Improving Group Decision Making for Techno-Economic Analysis

Submitted in partial fulfillment of the requirements for
the degree of
Doctoral of Philosophy
in
Engineering and Public Policy

Niles Guo
B.ASc. Systems Design Engineering, University of Waterloo
M.S. Public Policy and Management, Carnegie Mellon University

Carnegie Mellon University
Pittsburgh, PA

May, 2022

Acknowledgements

This research was possible through the support of Carnegie Mellon's Department on Engineering and Public Policy, Carnegie Mellon's Wilton E. Scott Institute for Energy Innovation (2017 Seed Grant for Energy Research), National Science Foundation Decision, Risk, and Management Sciences program (SES-2049333), U.S. Army Research Laboratory (ARL) (FP00003217), EQT Industry Grant, Carnegie Mellon University's Graduate Student Assembly CT Fund, and Carnegie Mellon University's Center for Behavior and Decision Research (CBDR) Small Grants Program.

I would like to express my most sincere gratitude towards my advisors Prof. Alex Davis (committee chair) and Prof. Meagan Mauter (committee co-chair). Alex, thank you for always being there for me, no matter how often I just showed up unannounced at your office door to discuss whatever gnarly issue I had at the moment, or when I needed someone to bounce ideas off, both for research and for life in general; I could not have made it through the past few years without you. Meagan, thank you for believing in me from the beginning and giving me the opportunity to start this journey. Without your guidance throughout my Ph.D. on how to be a researcher, I could not be where I am today; I am forever grateful for your guidance.

Next, I would like to thank the other committee members for their comments and input: Prof. Kate Whitefoot and Dr. Cristobal de la Maza. Your valuable input and insight during this process enriched my work and provided them with the necessary rigor.

In addition, I would like to thank Prof. Jay Whitacre. You and Meagan gave me the first opportunity to investigate how we can improve group decision making for complex systems and that became the seed that blossomed into this dissertation, thank you.

I also want to extend my gratitude to my co-authors for Chapter 4, Dr. Sangjin Jung, Dr. Patrick Funk, and Dr. Sandra DeVincent Wolf. I started the project not knowing the difference between additive and subtractive manufacturing, and your patience, guidance,

and willingness to answer all of my questions no matter how absurd were the only reasons that I stayed afloat during that time, thank you.

I was lucky enough to be part of two wonderful communities during the course of my Ph.D., and I cannot imagine how I was able complete this journey without the people in EPP and in the WE3Lab. The friendships I formed will last me a lifetime, and I will never forget the support you offered me. Without any particular order, to Amanda Quay, Daniel Gingerich, Conor Doherty, Alison Fritz, Yang Liu, Tim Bartholomew, Sneha Shanbhag, Alex Newby, Alex Dudchenko, and all the past and present members of WE3Lab (and all the honorary members who squatted in Scott Hall 5002, looking at you Rebecca Ciez), you made coming to campus each day fun and exciting. I will miss our Drink Talk Learns, soup swaps, early morning gym runs, and group lunch hours. To the EPPers, Ana Caceres, Luke Dennin, Logan Warberg, Prithvi Archarya, Priyank Lathwal, Turner Cotterman, Jessica Lovering, Sean Smillie, Gill Goobie (You are an honorary EPP), Aniruddh Mohan, Yanran Yang, Matt Bruchon, and many more, thank you for making the EPP community the welcoming space it is, and it will always have a special place in my heart. To everyone with whom I had the opportunity to work with in the Graduate Student Assembly, specifically Tamara Savage, Kristen Allen, Divyansh Kaushik, and Angelena Bohman, thank you for the opportunity to work on some really special initiatives both within EPP and the university at large, we made a difference! To all the wonderful staff who made EPP such a comforting place: Vicki Finney (sorry for always forgetting to register my research credits each semester until I see your email, Vicki!), Debbie Kuntz, Adam Loucks, and Kim Martin, thank you! To all the faculty members who I had the privilege of learning from at CMU, thank you!

I am also grateful for all of my non-EPP friends for keeping me sane and reminding what non Ph.D life could look like. To my (former) roommate Nate Breg, thank you not only for your friendship, but also the love and care to Felix (the cat) whenever I’m not home (there is still a reason you are one of his favourite humans). To the Heinz House (Nate, Emily Rosen, Payce Madden, Jafar Haider), thank you for always keeping me entertained through our wonderful daily distractions in our WhatsApp group. To Christian Bergland, thank you for always being available to chat, no matter the time. To Mickey

McGlasson and Aatir Siddique, thank you for always being down to have a kick around, no matter the weather. To Joanne Wong, Adrian Spânu, Albert Chen, Phecda Louie, Joyce Mak, Vivek Maharajh, Torehan Sharman, Gabe Fuhrman, Sally Huang, Ananyaa Ravi, Rebecca Dilla, and many others scattered around the world, thank you for your support and believing in me on this journey.

To my family, here in Pittsburgh, and in Xi'an, and back home in Toronto, without you I wouldn't be here today. Thank you to Nancy Brown and Michael Wright, you welcomed me into the family from the first day and I am forever grateful. To my grandparents (PoPo, LaoYe, NaiNai, and YeYe), thank you for instilling in me the curiosity of the world around us, and the work ethics I needed to complete this journey and I miss you all. To my aunts, uncles, and cousins, both in Calgary and in Toronto, thank you for believing in me and cheering me on from near and afar. To my dad Guo Pingbo, who was always proud of my achievements, always offered your support, and always gave me the freedom to make my own decisions. Thank you all for supporting me during this long journey and I can't wait to see you all in person soon!

Finally, and most importantly, I dedicate this dissertation to the two most important women in my life, my mom Xu Xin Ping and my wife Tenley Brownwright, both of whom finished this journey before me. To my mom, thank you for all of your sacrifice so I could have a better life, and all the years where you had to take on a role of both parents for me, I could not be here without you. To my wife, Tenley Brownwright, thank you for supporting me through this journey, and being there for me through all the ups and downs of the past few years. You are a constant inspiration for me. I love you!

Abstract

Group decision-making for techno-economic assessment often involves coordination between multiple stakeholders who have their own perspectives, motivations, objectives, and expertise. This increases the complexity of the decision-making process, requiring interventions that can help groups tradeoff multiple objectives and avoid suboptimal decisions. I focus on two critical problems faced by groups of system designers that have been incompletely explored in behavioral and decision science domain: generating efficient designs and reaching group consensus. I develop and test two classes of behavioral interventions to address these problems: real-time feedback and consensus-driven group recommender systems. Although these have the potential to help groups generate better designs and reach consensus, they lack quantitative evidence of their effectiveness and feasibility to improve group decision-making processes in techno-economic analyses.

The objective of this thesis is to provide quantitative evidence of the feasibility and effectiveness of those interventions using behavioral experiments in a laboratory setting. I achieved these objectives by recruiting a wide range of participants (students, laypeople in the general public, experts in an emerging technology), organizing them into groups, and evaluated the feasibility and effectiveness of the interventions on helping them to improve their decision-making in three different techno-economic areas. I first evaluated the effectiveness of providing real-time feedback to groups of students tasked with designing a complex wastewater management system with multiple objectives. Later, I extend the behavioral experiment framework to build and evaluate the feasibility and effectiveness of using a consensus-driven recommender system to help groups of laypeople to come to a consensus for climate change policy. Finally, I apply the model developed previously to help experts in metal additive manufacturing (MAM) domain to determine part and subassembly suitability for MAM in an U.S. Army context.

List of Figures

List of Tables

Table of Contents

Chapter 1 Introduction

Group decision-making for techno-economic assessment often involves coordination between multiple stakeholders who have their own perspectives, motivations, objectives and expertise.[1,2] This increases the complexity of the decision-making process, require qualitative evaluations around tradeoffs of multiple objectives, and can lead to unsatisfactory and suboptimal decisions.[3–7] Existing research in the behavioral and decision sciences has focused on two objectives: generating better designs, and making better decisions in order to reach group consensus.[8–10] There is a wealth of behavioral science literature on different mechanisms that would improve these objectives, including interventions that are effective for improving some aspects of group decision-making (provide computational tools, encourage interpersonal communications, incorporate individual preferences on group outcome to narrow down design options etc.).[11–26] On the other hand, there are two interventions, real-time feedback and consensus-driven group recommender system, with the potential to improve efficiency of designs and group consensus, that lack quantitative evidence of their effectiveness.[9,12,27,28]

Although empirical evidence is lacking, there is reason to believe that real-time feedback and consensus-driven group recommender systems can improve group decision-making. Prior research found that when provided with real-time feedback, groups can better understand the goals and interactions of the entire system rather than just each person's subsystem.[29,30] Feedback allows team members to gain better understanding of other members' roles and responsibilities,[31,32] and can improve the motivation of the group members, reducing social loafing and other undesirable behavior.[33–35] Finally, real-time feedback can allow group members to validate their assumptions about the overall system and how their actions influence it.[12] Therefore, by providing real-time feedback, groups can better understand how each piece of the overall system interacts and functions, what each member's roles and responsibilities are, and generate designs that are closer to the global optimum.

Consensus-driven group recommender systems work by estimating and aggregating individual preferences of members of the group.[9] All preference elicitation studies elicit individual preferences using some method (discrete choice experiments, contingent valuation, ranking etc.) then aggregate those preferences by assigning equal weighting to all individuals. The equal weights aggregation rule is not normatively required, nor consistent with observed experimental results that find individuals care about the outcomes of others, desiring equity and equality.[36,37] By combining individual preferences using an aggregation rule implicitly preferred by the groups, a consensus-based group recommender system can help to narrow down options that are acceptable to the group, and allow them to make decisions that will lead to greater group satisfaction.[9,10,38] As opposed to many existing group recommender systems, consensus-driven group recommender systems attempts to find solutions that have a high level of agreement amongst its users, rather than recommendations with a minimum level of agreement.[9,10,38,39] Consensus-driven group recommender systems that can incorporate a mechanism that aims to reach consensus allows individuals to express their preferences over the distribution of outcomes in a group and allow groups to make decisions that will have the highest likelihood of agreement.

1.1 Using real-time feedback to improve group decision-making in techno-economic assessment

The design of complex systems involves coordination between multiple experts who have their own perspectives, motivations, objectives, and expertise.[2,40] The heterogeneous objectives of these experts increase the complexity of design tasks, create information asymmetries, risking bounded rationality on the part of each participant, who ignores global complexities and instead rationally solves a focal sub-problem.[3,41,42] This approach generates locally, but not globally, optimal solutions to the systems.[4] In addition, the design space is often extremely large, making it difficult for groups to determine whether their design was globally optimal.[43] Prior work finds that when decision-makers work in teams, sub-optimal solutions may result from biases such as groupthink, egocentric biases, and social loafing.[13,24,44,45] While some work has suggested real-time feedback as a mechanism to reduce those biases, there is a question of whether it is a necessary mechanism for improving group design performance.[46–48] If real-time feedback is unnecessary, organizations can focus on other ways to improve team collaboration. While one prior study found little benefit of real-time system-level feedback for students designing a complex interconnected system (a space mission), the small sample limited the study's statistical power.[49] Other researchers have focused on either perceptual tasks (identifying visual cues), or simple decision-making tasks (selecting the right candidate for a hiring decision), as opposed to the design of complex interconnected systems.[12,24,29,30,33–35,46,50] Studies that evaluate the performance of design teams in large samples, as in the case of Gonzalez-Mule et al. (2016), also have lacked objective criteria for evaluating the team's solution (instead using subjective criteria based on the organization's leader).[29]

In addition, researchers also found that feedback is not a necessary component to improve group performance in a low-level perceptual decision-making task.[11,12] In fact, Bahrami *et al.* found social collaboration was the only required component for performance improvements, and feedback to the group accelerated the process, but was not sufficient on its own.[12] In groups where participants have different abilities, participants interacting with other participants learned information and credibility of each other's estimates through other signals.[27] Therefore, although prior research suggests a strong role of real-time feedback on performance, the results may not generalize to the design of complex systems, where correctness is evaluated as tradeoffs between different objectives.[28]

We hypothesize that groups with real-time feedback will generate solutions closer to the global optimum in a complex design task, and these groups will also generate better design solutions than participants working independently or through informal collaboration. Even when there is no single correct decision, real-time feedback can improve group decision making through several mechanisms. First, researchers have found that feedback clarifies the goals and interactions between subsystems for individuals within a team.[29,30] In a survey of 110 defense industry manufacturing firms in South Korea, feedback provided clarity on group goals and allowed individuals to understand their interactions with the overall system.[29] Teams that received feedback saw their members gain better understanding of their interactions with the overall system for problem solving tasks (landing an airplane for a flight simulator).[30] Second, feedback (when provided objectively and in a timely manner), allows team members to evaluate their assumptions about the overall system.[47,48] In a study of student groups making hiring decisions, teams that received feedback performed better than teams with no feedback, as feedback allowed team members to evaluate their preconceived ideas and assumptions.[47] Finally, real-time feedback can improve the group's collaboration process.[33–35,51] In two experiments that tasked students to generate ideas for improving the campus parking system, researchers found that feedback provided additional motivation for all participants to increase their effort.[33]

Researchers also found that feedback can also reduce social loafing by identifying individual contributions in a group.[13,52]

1.2 Using consensus-driven group recommender systems to generate group consensus in techno-economic assessment

Informal coalition formation and bargaining is the most common approach to develop consensus, where groups discuss and make concessions to form larger groups with the hope of finding solutions with widespread appeal.[53] For problems with a large number of stakeholders, one way to identify compromises is to use opinion polls.[54] However, those polls only estimate the distribution of public preferences for single issues (as opposed to the joint distribution across many issues), and do not permit quantification of either the strength of support for alternatives or whether supporters are willing to make tradeoffs. Preference elicitation studies can provide that information, for example by using discrete choice experiments to elicit stated preferences over sets of public policies.[37,55,56] In these studies, individuals are asked to make explicit choices between policy platforms, and statistical models like the multinomial logit or mixed logit are used to estimate individual preferences. Policies with the broadest appeal can be found by aggregating preferences across individuals using a voting rule, a social welfare function that gives each individual's vote equal weight.[57,58]

Use of a function that gives equal weight to all individuals is, however, neither normatively required, nor consistent with descriptive studies of social preferences. For both individuals expressing their preferences over outcomes for a group (the **individual social preference function**), as well as the group's aggregate preference (the **group social welfare function**), there are many alternatives to equal weights.[57] From the normative perspective, under the assumption that individual preferences are cardinal (invariant up to affine transformations) and fully comparable (CFC) across decision-makers (both utility levels and changes are comparable), there are many individual social preference functions and group welfare functions.[59,60] Although equal weights is one admissible function under CFC, choice based on the minimum utility (that is, choosing the alternative that maximizes the utility for the worst-off individual, or the least miserable), is another.[9,61,62] In fact, all individual social preference functions and group welfare functions have a common form under CFC. Consider $i \in \{1, 2, \dots, C\}$ decision-makers who have cardinal utility functions $u_i(.): \mathcal{X} \rightarrow R$ that map alternatives $x \in \mathcal{X}$ to real numbers. Call $w_1(.), w_2(.), \dots w_C(.)$ the social preference functions privately held by everyone that map the C individual utilities to a real number $w_i(.): \mathbb{R}^C \rightarrow \mathbb{R}$ If these social preference functions satisfy the Pareto and independence criteria, then they have the form:[60]

$$w_i\big(u_1(.), u_2(.), \dots, u_n(.)\big) = \bar{u}(.) + g_i\big(u_1(.) - \bar{u}(.), u_2(.) - \bar{u}(.), \dots, u_n(.) - \bar{u}(.)\big) \text{ (1-1)}$$

where $\bar{u}(.) = \frac{\sum_{i=1}^{C} u_i(.)}{C} and g_i(.): \mathbb{R}^C \rightarrow \mathbb{R}$ is a homogeneous degree-1 function. Similarly, any group social welfare function W must have the same form, minus the index i in g_i.

Descriptively, a consistent finding in the behavioral science and experimental economics literatures is that individuals do not just care about their own private outcomes, but also about the outcomes of others, meaning a simple equally weighted aggregation rule is not sufficient to capture the complexity of individual social preferences and group social welfare functions. For example, Bruhin *et al.* used a finite mixture model to uncover three types of preferences for inequality aversion among Swiss students.[19] Strong altruists (40% of the sample) were willing to pay 89 cents to reduce inequality by 1 dollar when they were ahead, and 19 cents when they were behind. Moderate altruists (50% of the sample) were willing to pay 15 cents and 7 cents, respectively. Behindness-averse individuals (10% of the sample) were willing to pay 78 cents to reduce inequality by one dollar, but only when they were behind. The Bruhin *et al.* study is one of several that use finite mixture models over predefined preference types (Iriberri & Rey-Biel, 2011, 2013; Conte & Moffatt, 2014; Conte & Levati, 2014; Bardsley & Moffatt, 2007), an approach that complements many studies that have used parametric models to fit individual social preferences (Andreoni and Miller, 2002; Bellemare et al., 2008; Fisman et al., 2007, 2015).[19–23,63–67] Other studies use non-parametric techniques (Kerschbamer, 2015, 2017), and evaluate the stability of social preferences, for example over time (Volk et al., 2012; Blanco et al., 2011) as well as in the field (Karlan, 2005; Benz & Meier, 2008; Fehr & Leibbrandt, 2011; Laury & Taylor, 2008).[68–75] There is also research examining contextual effects such as reputation and type of outcome on the form of social preferences. Findings from Bolton *et al.*

indicate that experimental results tend to be driven by unfairness averse preferences as opposed to reputational issues (decisions are unchanged when the experimenter cannot identify players).[76] Davis, Miller, and Bhatia find that a large majority of participants asked to split a painful experience had equality-seeking preferences (approximately 80% were characterized by equality-seeking, while 20% were selfish).[77]

These results indicate that a mechanism that aims to reach consensus should allow individuals to express their preferences over the distribution of outcomes in a group. However, prior work has either been descriptive, examining individual social preferences in different contexts, or has assumed a group social welfare function, as is done in prior work on group recommender systems and participatory budgeting.[9,38,78–83] This neglects an important relationship between the individual social preference functions w_i and the resulting group social welfare function W. It is possible to represent the group's social welfare function in terms of simple statistics over the individual utilities when each individual has a common (or similar) individual social preference function. Further, the same result could be obtained by a group of individuals with selfish individual social preference functions ($w_i = u_i$) but in aggregate uses an inequality averse group social welfare function.

We argue that finding policies that satisfy members of the public with different preferences can be seen as an optimization problem, that we call **social welfare optimization**, that uses information about individual preferences and group behavior to find the group social welfare function that implicitly best fits the group's values. We propose an empirical approach that models a group's social welfare function based on both individual preferences and group decisions, then validate the approach by observing whether the group reaches consensus on the recommendation.

1.3 Research objectives

This dissertation uses group behavioral experiments to investigate the effectiveness of these proposed behavioral interventions on group decision making for techno-economic assessments (Figure 1-1**Error! Reference source not found.**). This work evaluates how these interventions influence **the generation of designs and group consensus**. Specifically, By recruiting both laypeople and expert participants to make group decisions in three diverse technical domains (wastewater system design, federal decarbonization strategies in the United States, consolidation and part selection for metal additive manufacturing), I will a) determine the feasibility of real-time feedback and group recommender systems to improve the efficiency and consensus of group decisions, and b) evaluate the effectiveness of those interventions using laboratory experiments.

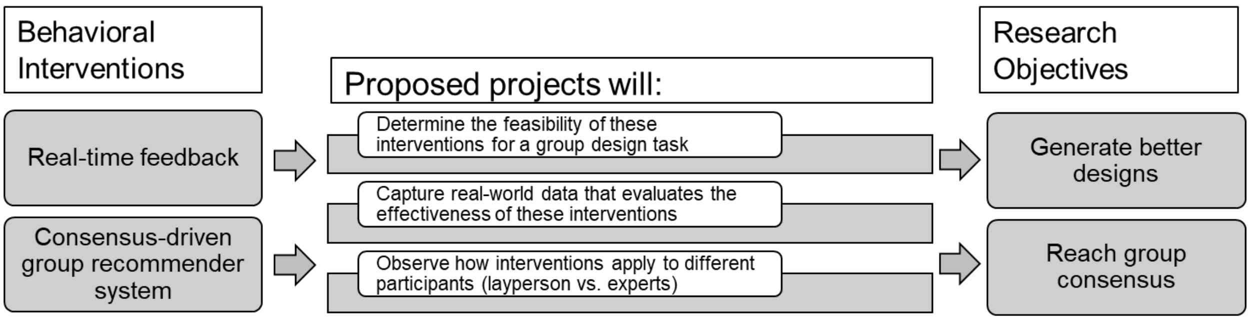


Figure 1-1 - Proposed work and their impact on research objectives

To accomplish these goals, I propose three research projects (Figure 1-2):

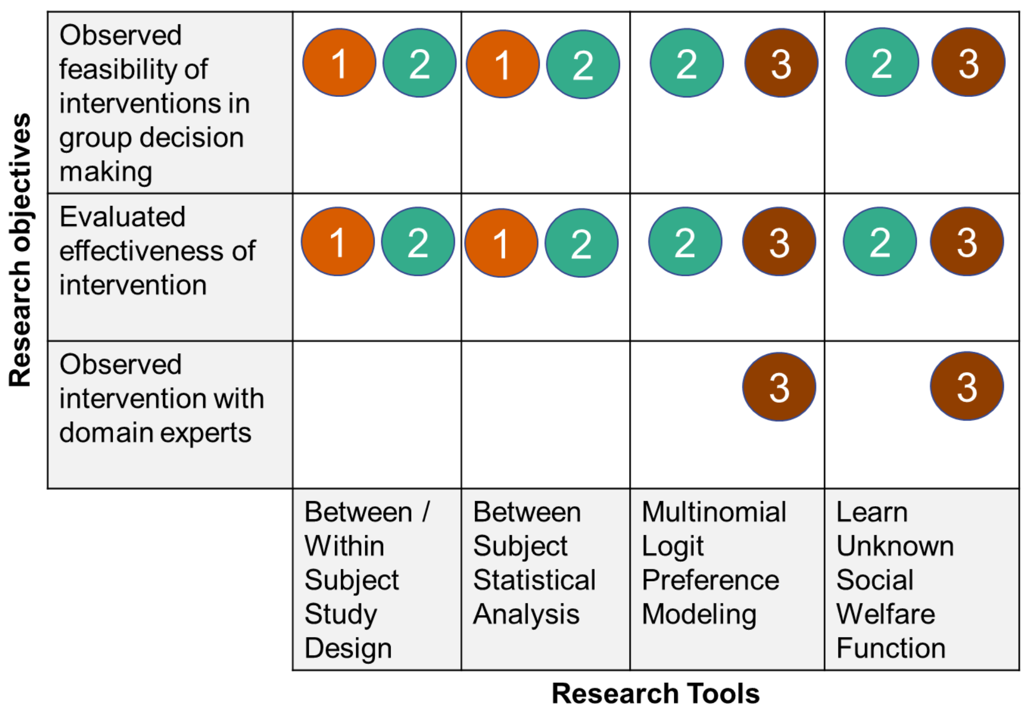


Figure 1-2 - Proposed research domains and tools associated with each proposed project

Chapter 2: Evaluating how real-time feedback improves multi-stakeholder design for complex environmental systems.

- This chapter evaluates the effectiveness of providing real-time feedback to groups (CMU students) tasked with designing a multi-objective wastewater treatment system for unconventional oil and gas exploration. Employing a **between/within subject experimental design**, I measured the differences in design performance as groups tackle the problem without any collaboration amongst its members, with informal communication between its members, and with real-time feedback provided to the group. By building the experiment on an existing multi-objective optimization model, we can evaluate how different designs generated by the groups differ from solutions on the Pareto frontier.

Chapter 3: A quantitative method for reaching consensus on federal climate change policy in the United States.

- This chapter evaluates the feasibility and effectiveness of reaching group consensus on climate and energy policy by **optimizing an unknown social welfare function** in a 3-stage experimental setting. Pittsburgh residents who planned to vote in the 2020 Democratic Presidential Primary were recruited in Spring 2020. I evaluated the feasibility of learning a group's social welfare function in real-time based on group members' individual choices. I then evaluated the effectiveness of the group recommender system by measuring the group's acceptance and their satisfaction with the recommendation.

Chapter 4: Automating Subsystem Consolidation Evaluations and Part Selection for Metal Additive Manufacturing

- This chapter measures the feasibility and effectiveness of the group recommender system for **experts in metal additive manufacturing**. Specifically, the chapter evaluates how a group recommender system can help experts with different technical backgrounds (material science, additive manufacturing, military logistics etc.) come to a consensus when faced with a decision to select appropriate parts for additive manufacturing for US Army priorities. By fostering deliberation between groups of experts and helping them come to consensus, the project 1) used expert-guided learning to filter out parts unsuitable for additive manufacturing, 2) expanded our existing techno-economic and expert decision models to incorporate Army priorities and procedures, and 3) prototyped a learning algorithm, seeded with expert knowledge, to automatically search Army databases for parts most suitable for additive manufacturing.

1.4 Dissertation structure

This dissertation consists of an introduction (Chapter 1), three chapters on three different research projects; one published (Chapter 2); one currently under review for publication (Chapter 3); one that will be submitted for publication in a peer-reviewed journal after the completion of the defense (Chapter 4), and an overall conclusion of this dissertation (Chapter 5). Chapter 1, the introduction, presents the motivation and background for the dissertation and the overall structure of this dissertation. Chapter 2, published in 2021 in Environmental Research and Communication[84], evaluated the impact of providing real-time performance feedback on groups tasked with a complex system design problem. Chapter 3, currently in review for publication in Nature Communications, investigates and evaluates the feasibility of generating climate change policy consensus using a consensus-driven group recommender system. In addition, this chapter investigates the first order policy cost and effectiveness of the recommended policies. Chapter 4, currently being prepared for submission, extends the method built in Chapter 3 and applied to eliciting expert judgment for part selection and part consolidation for metal additive manufacturing in the U.S. Army context. In addition, using the models described in the chapter, I developed an interactive platform demo that allows single experts enter potential candidates and receive a suitability score on their appropriateness for metal additive manufacturing and for subassembly consolidation. Finally, Chapter 5 presents a summary, future prospective, and conclusion of the dissertation.

Chapter 2 Real-Time Feedback Improves Multi-Stakeholder Design for Complex Environmental Systems[1]

[1] The contents of this chapter and its supplemental information (included as Appendix A) have been published as: Guo N, Davis A, Mauter M, Whitacre J. Real-time feedback improves multi-stakeholder design for complex environmental systems. *Environ Res Commun*. 2021;3(4):45006.

2.1 Abstract

We test whether providing quantitative real-time feedback relating design decisions to system objectives improve group solutions in an interdependent energy-water design task. While prior research suggests an important role of real-time feedback on task performance, few studies have examined the role of real-time feedback in the design of complex environmental systems. We tested a real-time feedback approach using a mixed within- and between-subject experiment (n = 88 Carnegie Mellon University students, divided into 22 groups of four). When compared to individual designs and informal collaborations, real-time performance feedback yielded solutions closer to the Pareto frontier and reduced both financial (by 26% and 21%) and environmental cost (by 34% and 12%). In addition, informal collaboration did not improve group decision-making when compared to individual designs. The results suggest that optimal solutions to meeting energy and water demand while minimizing cost and environmental impact can be obscured in informal collaborations, but that real-time feedback to system designers can help avoid waste of public resources.

2.2 Introduction

The design of complex systems involves coordination between multiple experts who have their own perspectives, motivations, objectives, and expertise.[40,85] The heterogeneous objectives of these experts increases the complexity of design tasks, risking bounded rationality on the part of each individual, who ignores global complexities and instead rationally solves a focal sub-problem.[3] This approach generates locally, but not globally, optimal solutions to the systems.[4] We propose an approach called Concurrent Assessment and Design of Systems (CADS) that combines a concurrent design process with real-time feedback to facilitate design decisions. Compared to existing engineering design frameworks, CADS helps decision-makers discretize systems into interconnected submodules, then adds real-time system performance feedback to relay design outcomes to decision-makers.[86–88] The goal of CADS is to help decision-makers converge on a globally optimal design through structured interaction and collaboration. In the present research, we experimentally examine the effect of one component of CADS, real-time performance feedback during group collaboration, to evaluate that component's importance as an environmental engineering design tool.

Compared to teams that work in isolation on their subproblem and then try to come to agreement on a global solution, real-time feedback provides each team member with the team's current total performance on a task.[89] This real-time feedback does not require another expert to evaluate and provide information to teams, as is typically done in after-the-fact performance evaluations.[89] Prior research finds that when decision-makers work in teams, biases such as groupthink, egocentrism, and social loafing can lead to suboptimal solutions.[13,44,45,90] While some findings have suggested that real-time feedback can reduce those biases, there is an open question about whether it is necessary for improving group design performance.[46–48] While one study found little benefit of real-time system-level feedback for students designing a complex interconnected system (a space mission), the small sample limited the study's statistical power.[49] Other researchers have focused on either perceptual tasks (identifying visual cues), or simple decision-making tasks (selecting the right candidate for a hiring decision), as opposed to the design of complex interconnected systems.[12,27,29–31,33–35,46,50,90] Studies that evaluate the performance of design teams in large samples, as in the case of Gonzalez-Mule et al. (2016), evaluated primarily the effect of ex-post feedback (self-reported performance reviews) rather than real-time feedback, and also have lacked objective criteria for evaluating the team's solution.[29] Although prior research suggests a strong role of real-time feedback on performance, the results may not generalize to the design of complex systems, where correctness is evaluated as efficient tradeoffs between different objectives.[28]

We hypothesize that groups receiving real-time feedback will generate better solutions in a complex design task than participants working independently or through informal collaboration. Real-time feedback can improve group decision-making through several mechanisms. First, researchers have found that feedback clarifies the goals and interactions between subsystems for individuals within a team.[29,30] In a survey of 110 defense industry manufacturing firms in South Korea, respondents indicated that feedback provided clarity on group goals and allowed individuals to understand their interactions with the overall system.[29] Teams that received feedback also gained a better understanding of the roles of other team members.[31] Second, feedback (when provided objectively and in a timely manner), allows team members to evaluate their assumptions about the system.[12,31,33,34,50] In a study of student groups making hiring decisions, teams that received feedback performed better than teams with no feedback, as feedback allowed team members to evaluate their preconceived ideas and assumptions.[47] Finally, real-time feedback can improve the group's collaboration process.[33–35,51,52] In two experiments that tasked students to generate ideas for improving the campus parking system, researchers found that feedback provided additional motivation for all participants to increase their effort.[33] Researchers also found that feedback can reduce social loafing by identifying individual contributions in a group.[13,52]

Building on this prior research, we created a group design task for complex interdependent energy-water systems to test two hypotheses about group performance with versus without real-time feedback:

H1: Group members in a complex system design task receiving real-time feedback will generate system-level solutions that are close to the Pareto optimal solution.

H2: Group members receiving real-time feedback will generate system-level solutions that are better than solutions generated independently or through informal collaborations.

To test our hypotheses, our working example is a wastewater treatment system for unconventional gas exploration operation in the Marcellus region (a region spanning New York, Pennsylvania, Maryland, Ohio, Virginia and West Virginia).[91] The problem is formally characterized using a multi-objective, mixed integer linear programming (MILP) model.[91] In this case study, potential designs had two competing system objectives, either to reduce project lifetime financial cost, or reduce the human health impacts from air emissions (see Figure 2-1). To design a system that meets those two objectives, analysts need to gather information from a range of experts on parameters such as wastewater flowback rate, wastewater composition in frack fluid, and wastewater treatment efficiency. Decision-makers also need to provide information on the set of system constraints, including fracking schedule, mass-balance, capacity, and finances. Finally, decision-makers need to make design and policy decisions about freshwater source use, wastewater reuse, and water transportation options.

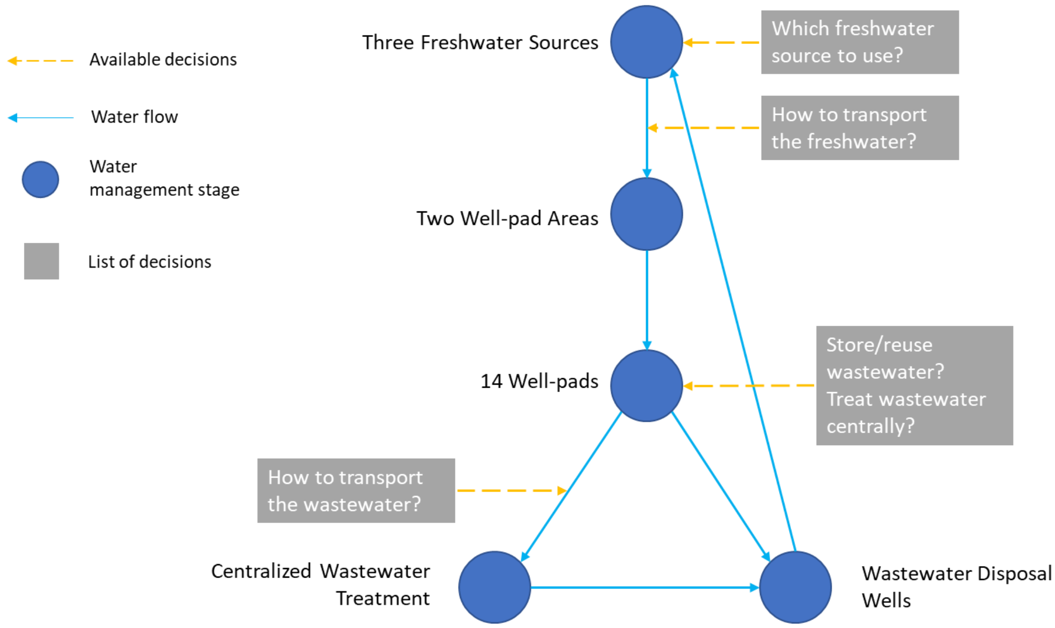


Figure 2-1 - Shale gas wastewater management system diagram, adapted from Bartholomew and Mauter. Experts needed to decide how to transport water between the different stages, along with other decisions such as whether to reuse wastewater, store wastewater, or to treat the water centrally. Experts also made policy decisions about where to draw freshwater while balancing different stakeholder preferences.

We used this case study to examine the effect of providing real-time feedback on system performance in a multi-participant, multi-objective case study. We compare real-time feedback against two alternatives: an independent design approach where each participant had to make decisions on their own, and an informal collaboration approach where participants worked together without real-time feedback.

2.3 Method

2.3.1 Research design

Participants were asked to design a wastewater management system for shale gas exploration in the Marcellus region.[91] Initially formulated as a MILP, changes were made to allow research participants to complete the tasks in a timely manner. These include:

- Pipelines can no longer be leased on a weekly basis.
- Water must be transported through trucking or constructing a new pipeline.

- Transportation option decisions are not made on a weekly, individual connection basis, but are applied across the entire system for the duration. This reduces the number of binary decision variables from 1750 to 13, making it tractable as a design exercise.
- Increased the number of freshwater sources from one to three. Participants needed to select the freshwater source for the system.
- Increased the number of pipeline options from one to three. Participants needed to choose a specific pipeline capacity if they chose to use it for water transportation.

Due to these changes, the problem transformed into these two objective functions:

$$Cost^{Financial} = \left(Cost^{WaterWithdrawal} + Cost^{FreshwaterTransportation} + Cost^{WastewaterTransportation} + Cost^{Storage} + Cost^{WastewaterTreatment}\right) \quad (2\text{-}1)$$

Where:

$Cost^{WaterWithdrawl}$ is the cost of freshwater withdrawal,

$Cost^{FreshwaterTransportation}$ is the cost of freshwater transportation,

$Cost^{WastewaterTransportation}$ is the cost of wastewater transportation,

$Cost^{Storage}$ is the cost of water storage,

$Cost^{WastewaterTreatment}$ is the cost of water treatment.

And:

$$Cost^{HHE} = \sum_e CE_e\left(M_e^{Storage} + M_e^{Pipeline} + M_e^{Trucking} + M_e^{CentralTreatment} + M_e^{Disposal}\right) \quad (2\text{-}2)$$

Where:

CE_e is the cost of emissions for each air pollutant e,

$M_e^{Storage}, M_e^{Pipeline}, M_e^{Trucking}, M_e^{CentralTreatment}, M_e^{Disposal}$ are the amount of air pollutant e generated for each activity.

Further details for each objective function can be found in Appendix A.1.

The model also consists of a series of constraints, ranging from mass balance constraints, to flow capacity constraints, to scheduling constraints. It is important to note that even with these constraints, there are ~2,200 different possible designs for the system, making it infeasible for participants to iterate through the entire solution space under a time constraint. The study employed a mixed within- and between-subjects design, with 88 participants drawn from the Carnegie Mellon University student population. Our experimental design was reviewed and approved by Carnegie Mellon University's Institutional Review Board prior to subject recruitment. There were no exclusion criteria. Students were recruited through posters around campus, and participants were compensated with a lottery of five $50 Amazon gift cards. Each participant in the group was randomly assigned an expert role and provided with briefing material on their expertise. There were four expert roles: Well-Pad Operator, Freshwater Expert, Wastewater Expert, and Environmental Regulator. The briefing material was unique to each expert role and included information on the parameters of their areas of expertise, their individual goals, their motivations, and the expected interactions with other members of the group. During the experiment, the experimenter was present in the room observing participant behavior and ensured that participants did not share their briefing material with other members of the group. Carnegie Mellon University students do not accurately represent the population of experts because they did not have years of experience working in the oil and gas industry, nor the technical expertise in designing a wastewater treatment system for shale gas exploration. To address their lack of expertise on the subject matter, the briefing material included a general overview of unconventional gas exploration, specific domain knowledge such as pipeline capacity and transportation distance, and historical system performance that provides a useful marker for the participants. We also added contextual information, including the motivation of their role as an expert. Additional details on the briefing material (including the specific briefing material for each role) can be found in Appendix A.2. Participants conducted the research tasks through a web-app (built through R Shiny) specific to their expert role. From the participant's perspective, these apps represented the submodules that provided all the information they needed to make decisions, as well as the performance of their submodules as a function of the group's decisions. Participants didn't have to enter their design decisions in a specific order (e.g., regulator first).

2.3.2 Research tasks

Pre-study Task: Each participant was briefed with the setup and the relevant information related to their role. To verify that participants sufficiently understood the setup prior to the tasks, three True or False validation questions were asked. If the participant did not answer the validation question correctly, the correct answer was provided, and the participant was given an opportunity to ask clarifying questions.

Task 2 (Collaboration Task): In this task, participants were asked to collaborate informally and generate a design solution. The results for both their individual design and the overall system were recorded. In addition to in-person dialogue, the participants could use markers and dry-erase whiteboards. When compared to the results from Task 1, this task measured the effects of informal discussion on both individual and group performance. The participants had 20 minutes to complete this task.

Task 3 (CADS Task): In this task, participants were asked to collaborate and generate a design solution with additional access to a dashboard with real-time performance feedback. The feedback included overall financial and environmental costs of the design they generated, along with a component-by-component breakdown of the cost. Finally, the feedback included the decisions each participant made for each design, allowing participants to understand how their collective decisions affected the overall system. The feedback did not provide suggestions for future designs, nor did it indicate whether the solution was on the Pareto Frontier, meaning participants needed to actively link the feedback they received with the decisions they made to make improvements to their design. Their results for both their individual design and the overall system were recorded. When compared with both Task 1 and 2, this task measured the effect of real-time performance feedback on both individual and group performance. The participants had 20 minutes to complete this task.

Each group consisted of four students who completed all three research tasks, and we compared the performance between each group and within each group, across all three tasks. To explore potential order effects, the order of Tasks 2 and 3 was randomized between-subjects. Task 1 always came first.

2.4 Results

H1: We hypothesized that participants receiving real-time feedback would generate system solutions that are close to optimal.

Because the task was a multi-objective optimization problem (both environmental and financial cost), there are multiple potential optimal solutions that generate a Pareto frontier. We classified distance to the optimal solution as the smallest Euclidean distance to the Pareto frontier across the multi-objective outcome space. Our hypothesis test constructed a 20% margin of non-inferiority surrounding each solution on the Pareto frontier. We assessed whether solutions generated by participants across the different tasks fell within that margin of non-inferiority.[92]

At the group level, participants receiving real-time feedback (CADS task) generated solutions that were closer to the Pareto frontier than the other non-CADS tasks. As shown in Figure 2-2, system outcomes created by groups with real-time feedback (Figure 2-2 panel B) overlapped much more closely with the Pareto non-inferiority margin than solutions generated through independent design (Figure 2-2 panel A) and through informal collaboration (Figure 2-2 panels C and D). The mean distance to the Pareto Frontier was \$2.4M for the CADS task (SD = \$2.2M, N = 22), compared to informal collaboration task of \$5.9M (SD = \$9.7M, N = 22), or independent task of \$14.5M (SD = \$14.7M, N = 22). The one-tailed test of whether the mean distances fell within the non-inferiority region is:

$$H_0:\ \mu_T\ \geq \mu_R + M_{NI}$$
$$H_a:\ \mu_T\ < \mu_R + M_{NI}$$

Where:

μ_T is the sample mean distance to the Pareto Frontier,

μ_R is the reference mean (a distance of 0 to the Pareto Frontier),

M_{NI} is the margin of non-inferiority (a distance = 20% of the closest solution on the Pareto Frontier).

Under this test, the CADS task has a t-statistic of -1.81 (df = 21, p = 0.04) compared to the informal collaboration task t-statistic of -0.05 (df = 21, p = 0.48) or independent task t-statistic of 0.55 (df = 21, p = 0.70). We could only reject the null hypothesis for outcomes generated through the CADS task, indicating that only the CADS task outcomes were statistically within the margin of non-inferiority of the Pareto efficient solutions. It is important to note that while we can only reject the null hypothesis for the outcomes generated through the CADS task, this is not evidence on its own that the CADS task performed statistically better than the other two tasks. H2 is used instead to evaluate the performance differences between the tasks.

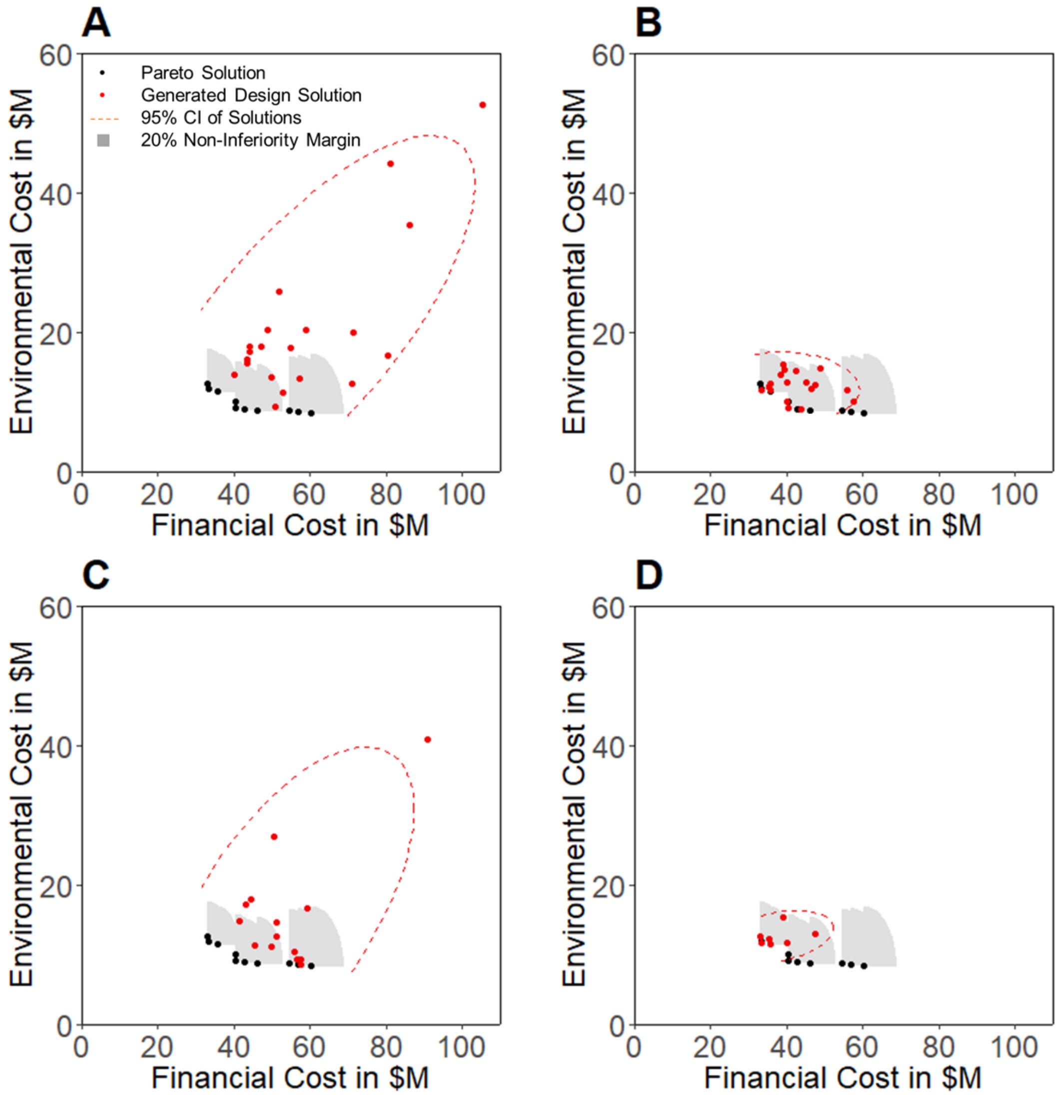


Figure 2-2 - Black dots represent Pareto outcomes, and gray area represents the 20% non-inferiority margin. (A) Design solutions created by participants from the individual design task. (B) Design solutions created by participants from the CADS design task. (C) Design solutions created by participants from informal collaboration design task before the CADS task. (D) Design solutions created by participants from informal collaboration design task after the CADS task. Red dotted lines denote 95% CI for the participant generated design solutions in the feasible region.

Participants receiving CADS feedback performed well regardless of the task order, whereas participants receiving informal feedback only performed well if that feedback came after the CADS task (Figure 2-2 panels D), but not before (Figure 2-2 panel C). This suggests an asymmetric transfer effect from the CADS task to informal collaboration, where participants were able to learn Pareto optimal solutions from the CADS task and transfer that to the informal collaboration task, but not vice versa. When the informal task came first the t-statistic for the informal collaboration task was 0.17 (df = 13, p = 0.57), showing no statistical evidence that groups generated designs inside the non-inferiority region when informally collaborating, whereas if the informal collaboration task was second, the t-statistic for the informal collaboration task is -2.30 (df = 7, p = 0.027) showing statistical evidence that these groups generated designs inside the non-inferiority region. On the other hand, if the CADS task was first, the t-statistic for the CADS task is -1.85 (df = 7, p = 0.053), whereas if the CADS task was second, the t-statistic for the CADS task is -1.81 (df = 13, p = 0.047) showing that the order of the CADS task did not appreciably change the result.

H2: Our second hypothesis was that participants receiving real-time feedback would generate system solutions that are better than solutions generated independently or through informal collaboration.

We tested H2 by constructing a linear regression model that evaluated the effect of providing participants with real-time feedback on their group performance, as measured by the system's financial and environmental impact costs. We found that participants receiving real-time feedback generated better solutions for both objectives than the independent design and informal collaboration design. Further, there was an asymmetric transfer effect, where participants who completed the real-time feedback task before the informal collaboration task performed better on the informal collaboration task than when the order of tasks was reversed. Figure 2-3 shows that the distribution of the financial cost (Figure 2-3 panel A) and environmental cost (Figure 2-3 panel B) are skewed with a long right tail. To account for the skew in the underlying distributions, we log transformed the dependent variables. The resulting distributions appear to be less skewed, with a shorter tail and fewer outliers. To address the concerns raised by Lo and Andrew (2015) about the log-normal assumption, we repeated the analysis with the model specification using a log link function, and there was very little difference between the log transformed and the log link result.[93] Additional details can be found in Appendix A.3. The model specification is:

$$log\ (ObjCost_i)\ = B_0 + \delta_1 \times CollabTask + \delta_2 \times CADSTask + \gamma \times 1(CADSFirst) + \tau_1 \times CollabTask \times 1(CADSFirst) + \tau_2 \times CADSTask \times 1(CADSFirst) + X_i\beta + \varepsilon_i \quad (2\text{-}3)$$

Participants receiving real-time feedback generated solutions that are better than the independent design and informal collaboration design for both objectives (shown in Figure 2-3). As shown in Table 2-1, participants who received the real-time feedback generated solutions that were on average ~\$17.2M lower in financial cost (26%[2], df = 58, t = -3.29, $p < 0.01$), and ~\$8.4M lower in environmental cost (34%, df = 58, t = -2.39, $p < 0.05$) when compared to the same group's performance in the independent design task, and ~\$11.6M (21%, df = 58, t = -3.21, $p < 0.05$) lower in financial cost and ~\$3.2M (12%, df = 58, t = -1.03, $p < 0.30$) lower in environmental cost when compared to the same group's performance in the informal collaboration task. The difference between the real-time feedback group and the informal collaboration task is not statistically significant for environmental cost, potentially because many groups chose to trade off environmental cost to increase their gains in the financial cost category between the informal collaboration stage and the feedback stage.

The cost reductions relative to independent design were statistically significant for the CADS task both when the CADS task was before informal collaboration and after. However, there was no improvement on the informal collaboration task relative to the independent design task when participants completed the informal collaboration task before the CADS task. This suggests there is significant learning and improvement from real-time feedback, and that improvement was transferred to the informal collaboration group when CADS came before informal collaboration, but not after. Without real-time feedback, there was little learning, suggesting the presence of asymmetric transfer effect. This can be directly observed from the regression results. We conducted a Z-test on the collaboration order effect ($\tau_{F1} = -0.29$) against the CADS order effect ($\tau_{F2} = -0.01$) for financial cost and found that the two coefficients are statistically different (Z = 2.24, $p < 0.05$). This showed the order effect for collaboration is significantly larger than the order effect for CADS, resulting in an additional decrease of approximately \$11.6M (~25%) on top of the effect of collaboration. However, this asymmetric transfer effect was not significant for environmental cost ($\tau_{E1} - \tau_{E1} = -0.06$, Z = 0.36, p = 0.36), showing that the order effect for collaboration was approximately the same for CADS.

---

[2] Percentages are calculated using the coefficients from the log-transformed models with methods from Halvorsen, R., & Palmquist, R. (1980).[218]

This could be the result of the design setup, where only one participant's primary objective in each group was environmental cost, and therefore most of the group emphasized financial cost.

We conducted an analysis of the residuals and found the residuals for the models are approximately normally distributed with conditional mean around 0. However, there is some heteroskedastic behaviour in the residuals that warranted the use of heteroskedastic and clustered robust standard errors that increased the p-values of the results. It is important to note that despite the use of heteroskedastic and clustered robust standard errors, the CADS treatment coefficient remained statistically significant for both objectives at $p < 0.05$ level. Additional details of this analysis without the clustered standard errors are reported in Appendix C.

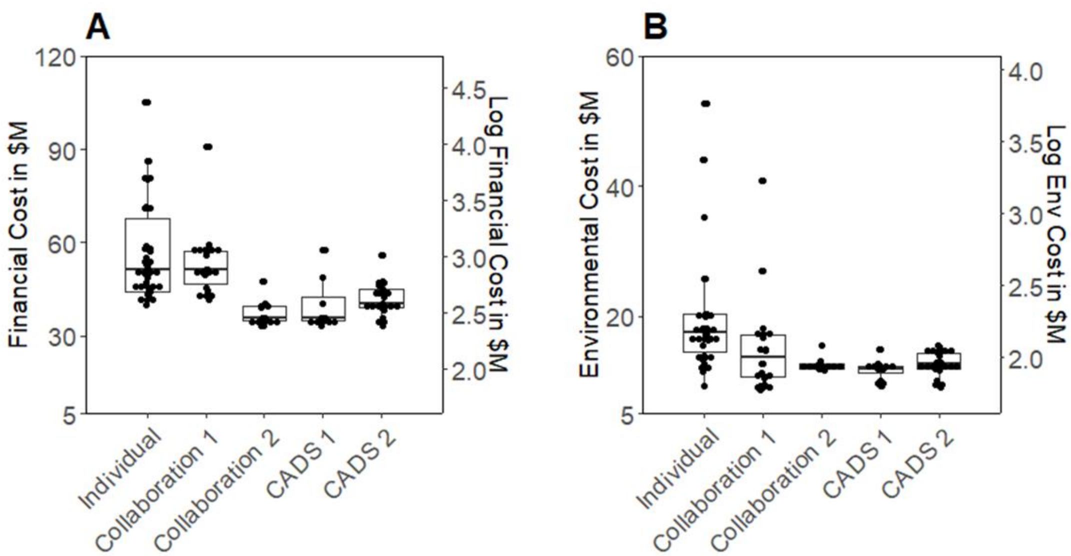


Figure 2-3 - Boxplots showing group performance for financial cost (A) and environmental cost (B), log transformed financial cost and log transformed environmental cost are presented as secondary y-axes. Collaboration corresponds to the Informal Collaboration research task, the number after Collaboration and indicates the order of the task (came first or came second). Boxes reflect the 25th and 75th percentile results, while the whisker extends to the value that is 1.5 times of the interquartile range (IQR).

Table 2-1 - Treatment effects of each task for both objectives, comparing against independent task with CADSFirst = 0. Basic refers to models with only the treatment variable, Covar refers to models with treatment variable and group-level covariates, Order refers to models with treatment variable and task order dummy variables, and Order + Covar refers to models with the specification in (3). Standard errors are heteroskedastic and auto-correlated robust. ***p < 0.001, **p < 0.01, *p <0.05

		Financial Cost Models				Environmental Cost Models			
	Coefficient	Basic	Covar	Order	Order + Covar	Basic	Covar	Order	Order + Covar
Independent (CADSFirst= 0)	B_0	4.02***	3.94***	4.04***	3.98***	2.92***	3.12***	2.94***	3.15***
Independent (CADSFirst = 1)	$B_0 + \gamma$			-0.06	-0.09			-0.06	-0.05
Collaboration (CADSFirst = 0)	$B_0 + \delta_1$	-0.18*	-0.18***	-0.07	-0.07	-0.30*	-0.30**	-0.28	-0.28*
Collaboration (CADSFirst =1)	$B_0 + \delta_1 + \gamma + \tau_1$			-0.42**	-0.45*			-0.42	-0.41*
CADS (CADSFirst = 0)	$B_0 + \delta_2$	-0.31***	-0.31***	-0.31**	-0.31**	-0.41***	-0.41***	-0.41**	-0.41*
CADS (CADSFirst = 1)	$B_0 + \delta_2 + \gamma + \tau_2$			-0.38**	-0.41**			-0.49**	-0.48*

Due to the multi-objective nature of the problem, each group can have different weights for each objective.[94] Certain groups might weigh financial cost more than environmental cost, and their design decisions would reflect that choice. To verify the robustness of the results from Table 2-1 with different objective weights, we applied relative weightings to the dependent variables, starting at 0% environmental cost, and varied it at a 5% interval until a weighting of 100% environmental cost. The model is as follows:

$$WeightedObjective_i = B_0 + \delta_1 \times CollabTask + \delta_2 \times CADSTask + \gamma \times 1(CADSFirst) + \tau_1 \times CollabTask \times 1(CADSFirst) + \tau_2 \times CADSTask \times 1(CADSFirst) + X_i\beta + \varepsilon_i \text{ (2-4)}$$

$$WeightedObjective = \omega_1 * log\ (FinancialCost)\ + \omega_2 * log\ \ (EnvironmentalCost) \text{ (2-5)}$$

Where:

$$\omega_1 + \omega_2 = 1 \text{ (2-6)}$$

The CADS task treatment coefficients showed a consistent progression from the financial objective model to the environmental objective model. In addition, the t-statistics show that, regardless of the objective weights, the CADS Task variable remained statistically significant (Figure 2-4). This shows that the model results are robust regardless of the relative weighting assigned to each objective.

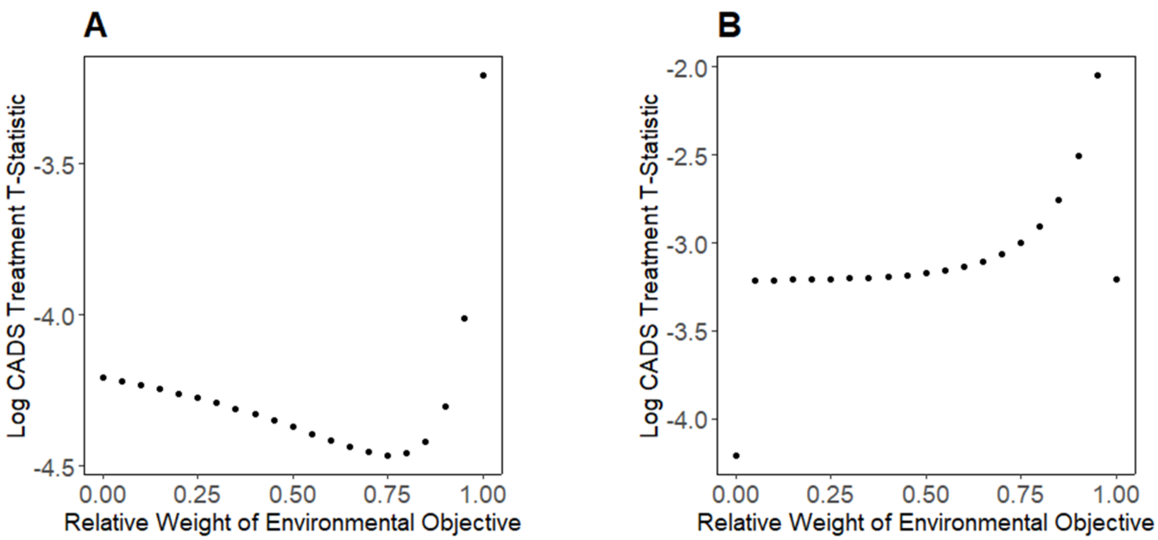


Figure 2-4 - (A) CADS task t-statistic compared to the independent task, plotted against environment objective weighting. (B) CADS task t-statistic compared to the informal collaboration task, plotted against environment objective weighting.

2.5 Discussion

In this study, we tested the effect of providing quantitative, real-time feedback on the relationship between design decisions and task objectives, then compared it against independent designs and informal collaborations in a multi-objective wastewater treatment system design task. Participants with real-time feedback generated solutions that are both closer to the Pareto frontier and with lower cost for both environmental and financial objectives than participants generating solutions independently or through informal collaboration. When participants attempted to generate solutions independently, they lacked insight about how their actions affected other members of the team and the overall system, leading to suboptimal outcomes. Even though the informal collaboration task allowed participants to communicate with each other and share information, the complexity of the system made it difficult for participants to understand the consequences of their actions. Some groups tried to understand the relationships between their roles and the overall system using the tools available to them in the room (e.g., whiteboards) during the informal collaboration task. However, we observed that no group was able to accurately map out the interdependent system relationship during the informal collaboration task. Finally, some groups decided to focus their attention on one individual module during the information collaboration task (akin to a depth-first search algorithm) and attempted to find the optimal solution with the lowest cost for that individual's role. However, that strategy did not always yield the best system solution, as sacrifices were made in other modules that made it globally inefficient.

When provided with real-time feedback, participants were able to better understand the impact of their decisions on the overall system through the display of objective metrics. We also observed that participants became more motivated as they grasped the connection between their decisions and the overall solution, which manifested in participants wanting to find better solutions. In addition, participants were able to use the real-time feedback to validate their assumptions for their decisions and recognize when those assumptions were false. Finally, real-time feedback improved the results of groups who had unmotivated members (observed through their lack of interaction with other members in their group), where the real-time feedback allowed groups to use that feedback as cues to point out where the unengaged members can improve the overall system.

There are several limitations to this study that are worth noting. First, the use of convenience sampling methods meant that participants are not representative of the intended population of industry experts.[95] We attempted to mitigate this limitation by providing briefing material to research participants that mimicked expert knowledge and behaviour (more detail in Appendix B), however there is still likely differences between how our participants responded to the research tasks compared to industry experts. Secondly, experimental design constraints such as time and cost meant that the design task is a facsimile of the real-world design challenge. There will be differences between the research task in this study and real design sessions, such as stronger personal motivations and familiarity with each participant due to prior experience.

Performance improvements could also have been due to activation of visual cues made available through the feedback mechanism, where participants performed better not because of the system-level feedback they received, but because of additional visual cues that activated their attention.[96,97] While this is a potential confounding factor, because we know that attention activation is also a function of time, and we observed no difference when the CADS task order was switched, some of this concern can be alleviated.[98,99]

This study finds a different effect of providing feedback during complex decision-making processes compared to simple perceptual tasks.[12] In our study, the effect of informal collaboration is not a statistically significant indicator of improved group performance. In contrast, the quantitative, real-time feedback mechanism provided to the groups had a large and statistically significant effect. This may be attributable to inherent differences between complex design tasks and perceptual tasks. In addition to advantages that real-time feedback provided to participants, *shared information bias* had a greater effect on team performance in complex design tasks than in perceptual tasks in prior work.[100,101] Without the ability to confirm their assumptions during the informal collaboration design task, participants lacked the ability to understand what information is essential, instead focusing on the information that all participants possessed. When provided with feedback, however, participants were able to confirm their assumptions and understand what information is needed from other participants to generate the optimal design. For complex design decisions, it appears that feedback is an important aspect of group success.

Although the use of students as participants threatens the external validity of the study, systems that real experts must deal with also have greater complexity than the one used in this study. Combined with the informational materials designed to get students up to speed on the task, the balance between participant expertise and task complexity may be similar in our study and the real world. Secondly, institutions such as the Jet Propulsion Lab (JPL), who in 1995 established the Advanced Projects Design Team (known as Team X) to design new space mission proposals, successfully used real-time feedback to improve their design process.[102] Experts were recruited for specific system modules in a design team (avionics, battery etc.), and they collaborated to generate mission designs through a series of concurrent design sessions.[103] When compared to previous space mission designs, Team X was able to design more missions per year, with lower average time for each design, and a lower average cost of design.[104] In addition, when simulated with experts against past mission parameters, Team X results were within 5% of actual mission costs.[105] The results of our study reinforce the idea that providing real-time feedback may have been an essential component of Team X's accuracy. Our results suggest that similar endeavors that aim to solve complex design problems, ranging from designing utility-scale grids to resilient public infrastructure, should include real-time feedback and collaboration among experts to avoid suboptimal outcomes.

2.6 Data Availability

Statistics were done using R 3.5.0 (R Core Team, 2018), the googledrive (1.0.0.9000; Bryan and McGowan, 2020), googlesheet (0.3.0; Bryan and Zhao, 2018), dplyr (0.8.5; Wickham et al., 2020), tidyr (1.0.2; Wickham and Henry, 2020), Rmisc (1.5; Hope, 2013), ggalt (0.4.0; Rudis et al., 2017), ggplot2 (3.3.2; Wickam, 2016), lme4 (1.1-21; Bates et al., 2015), car (3.0-7; Fox and Weisberg, 2019), texreg (1.36.23; Leifeld, 2013), ggforce (0.3.1; Pedersen, 2019), sandwich (2.5-1; Berger, Graham, Zeileis, 2017), and lmtest (0.9-37; Zeileis and Hothorn, 2002) packages.[106–117]

All data, code, and stimuli are available on GitHub: https://github.com/we3lab/MultiStakeholderDesign

Chapter 3 A quantitative method for reaching consensus on federal climate change policy in the United States[3]

[3] This chapter is being revised for publication in Nature Communications (February 2022) as: Guo, N., Mauter, M., Davis, A. "A quantitative method for reaching consensus on federal climate change policy in the United States."

3.1 Abstract

Methods that help identify policies that are acceptable to heterogeneous groups of individuals have the potential to improve the chance of group consensus. Existing methods, such as opinion polls and individual preference elicitation, identify policy platforms by estimating individual preferences for policies, then aggregate the preferences of those individuals to make recommendations to the group by giving equal weight to everyone. Such an approach is neither normatively required by axioms of preference aggregation, nor consistent with descriptive studies that find individuals care about the outcomes of others. Instead, we argue that finding policies that satisfy members of the group with different preferences can be seen as an optimization problem, that we call social welfare optimization, which uses information about individual preferences and group behavior to find a group social welfare function that implicitly best fits the group's values. We propose an empirical approach that models a group's social welfare function based on both individual preferences and group decisions, then validate the approach by observing whether the group reaches consensus on the recommendation. We apply this approach to the problem of setting climate policy among registered Democrats during the 2020 primary in a three-stage research design, where alternatives are multi-attribute potential policy platforms. We tested two approaches for estimating group social welfare functions, the mean-variance, which allows for inequality aversion, and weighted sum, which places different weights on each group member. Our first two pre-registered analyses found a suggestive, but not statistically significant, increase in agreement with the received recommendations from the mean-variance approach (n = 27) versus the weighted sum approach (n = 27) at both the group level (70% of groups reached consensus for mean-variance vs 52% for weighted sum) and individual level (85% vs. 69%, respectively). Our last two pre-registered analyses found that the mean variance approach resulted in greater satisfaction with the recommendation at both the group and individual levels (87% vs. 79%). Despite the lack of statistical significance for some of the results, there is some evidence that groups provided recommendations with the mean-variance model were more likely to reach a consensus and more likely to remain satisfied with the recommendation. Exploratory analyses revealed that many participants changed their policy preferences because of the group discussion. In addition, our results suggest support in the study population for a more aggressive federal 100% clean energy portfolio standard goal and more stringent limitations on fossil fuel exploration on public lands. Overall, our results suggest the

potential of improving the policy generation process and creating consensus for deep policy problems using a social welfare optimization approach.

3.2 Introduction

Although there is scientific consensus about the imminent danger of climate change, the United States federal government has so far been unsuccessful in passing national policies to reduce greenhouse gas (GHG) emissions that can limit global warming to 1.5C, despite attempts in both the legislative and executive branches.[118–123] The results of the 2020 U.S. federal election and the 2021 Georgia Senate run-off elections provide an opportunity for the U.S. federal government to enact meaningful climate change policies.[124,125] In response to President Biden's focus on climate change, policies such as a federal clean energy portfolio have been included in the U.S. House of Representatives' proposed budget during the summer of 2021.[126,127] There are still significant barriers for those policies to be enacted on a federal level. Conventional wisdom suggests that lack of progress has been primarily due to U.S. Republicans' skepticism about the reality of anthropogenic climate change despite most Democrats and Independents acceptance of the scientific consensus.[128–131] Yet, a recent nationally representative survey found that more than two thirds of Republicans agreed with statements that "global warming has probably been happening'" and "assuming global warming was happening, human activity was either equally or mostly responsible for global warming, in addition to natural causes."[128] Similarly, an analysis of the 109th and 110th U.S. congresses found an increasing consensus on the science of climate change, but disagreement about appropriate legislative action was responsible for impasse.[130] While such disagreements often take place between parties, intra-party disagreements can also prevent meaningful progress.[132] Despite having a super-majority in the U.S. Senate in 2009, President Obama was unable to pass a cap-and-trade bill due to a lack of consensus within the Democratic Party.[133] Similarly, the proposed Green New Deal in 2019 not only saw opposition by Republican lawmakers, but also Democratic members such as Rep. Tim Ryan (Ohio), Sen. Angus King (Maine) and others.[134] A similar fate threatens the Biden Administration's climate efforts (as seen in the recent opposition by Sen. Joe Manchin and Sen. Kyrsten Sinema).[135,136] Therefore, there is a need to identify policies that are acceptable within the Democratic party, where members are incentivized politically and ideologically to come to an agreement.

Informal coalition formation and bargaining is the most common approach to policy development, where groups discuss and make concessions to form larger groups with the hope of finding policy platforms with widespread appeal.[53] One way to identify compromise policies is to use opinion polls.[54] However, those polls only estimate the distribution of public preferences for single issues (as opposed to the joint distribution across many issues), and do not permit quantification of either the strength of support for policies or whether supporters are willing to make tradeoffs. Preference elicitation studies can provide that information, for example by using discrete choice experiments to elicit stated preferences over sets of public policies where individuals are asked to make explicit choices between platforms, and statistical models like the multinomial logit or mixed logit are used to estimate individual preferences.[37,55,56] In these studies, policies with the broadest appeal can be found by aggregating preferences across individuals using a voting rule, a social welfare function that gives each individual's vote equal weight.[57,58]

Use of a function that gives equal weight to all individuals is, however, neither normatively required, nor consistent with descriptive studies of social preferences. For both individuals expressing their preferences over outcomes for a group (**individual social preference function**), as well as the group's aggregate preference (**group social welfare function**), there are many alternatives to equal weights.[57] From the normative perspective, under the assumption that individual preferences are cardinal (invariant up to affine transformations) and fully comparable (CFC) across decision-makers (both utility levels and changes are comparable), there are many individual social preference functions and group welfare functions.[59–61] In fact, all individual social preference functions and group welfare functions have a common form under CFC (equal weighting being just one admissible function under CFC). Consider $i \in \{1, 2, \ldots, C\}$ decision-makers who have cardinal utility functions $u_i(.): \mathcal{X} \to R$ that map alternatives $x \in \mathcal{X}$ to real numbers. Call $w_1(.), w_2(.), \ldots w_C(.)$ the social preference functions privately held by everyone that map the C individual utilities to a real number $w_i(.): R^C \to R$. If these social preference functions satisfy the Pareto and independence criteria, then they have the form:[60]

$$w_i\big(u_1(.), u_2(.), \ldots, u_n(.)\big) = \bar{u}(.) + g_i\big(u_1(.) - \bar{u}(.), u_2(.) - \bar{u}(.), \ldots, u_n(.) - \bar{u}(.)\big) \text{ (3-1)}$$

where $\bar{u}(.) = \frac{\sum_{i=1}^{C} u_i(.)}{C}$ $and\ g_i(.): R^C \rightarrow R$ is a homogeneous degree-1 function.

Descriptively, a consistent finding in the behavioral science and experimental economics literatures is that individuals do not just care about their own private outcomes, but also about the outcomes of others, meaning a simple equally weighted aggregation rule is not sufficient to capture the complexity of individual social preferences and group social welfare functions.[19,21,67–70,74,76,77] These results indicate that a mechanism that aims to reach consensus should allow individuals to express their preferences over the distribution of outcomes in a group. However, prior work has either been descriptive, examining individual social preferences in different contexts, or has assumed a group social welfare function.[9,38,78–83] This neglects an important relationship between the individual social preference functions w_i and the resulting group social welfare function W. Work by Schmidt and Wichardt showed it is possible to represent the group's social welfare function in terms of simple statistics over the individual utilities when each individual has a common (or similar) individual social preference function.[137] We argue that finding policies that satisfy different preferences can be seen as an optimization problem, that we call **social welfare optimization**, that uses information about individual preferences and group behavior to find the group social welfare function that implicitly best fits the group's values. We propose an empirical approach that models a group's social welfare function based on both individual preferences and group decisions, then validate the approach by observing whether the group reaches consensus on the recommendation. Our approach focuses on estimating two unknown functions: 1) individual utilities for each of the climate policy alternatives ($\widehat{u_i}$), and 2) the group social welfare function that specifies how to combine individual utilities into a social decision rule ($\widehat{W}$).

We apply the approach to the problem of setting climate policy among registered Democrats for the 2020 primary in a three-stage research design, where alternatives m are multi-attribute potential policy platforms. We chose four different attributes (use of nuclear power in electric grid, 100% clean energy portfolio standard goal, carbon price, fossil fuel exploration on public lands), with three, three, six, and three attribute levels (respectively), resulting in 162 unique policy platform combinations. In Stage One, we model individual preferences; in Stage Two, we cluster participants into four clusters and generate groups with participants drawn from each cluster; finally in Stage Three, we observe the choice behavior of participants in groups of four (one from each cluster), use either a weighted-sum or mean-variance approach to recommend a policy platform the group, then measure the group's consensus on the recommended policy.

3.3 Methods

We apply the approach to the problem of setting climate policy among registered Democrats for the 2020 primary in a three-stage research design (seen in Figure 3-1), where alternatives m are multi-attribute potential policy platforms. The attributes of each alternative platform were chosen to be issues in climate policy that have gained significant attention within the Democratic Party, but where polls show disagreement within the party (to avoid trivial problems where consensus already exists).[138–140] The full set of attributes with their respective levels can be found in Appendix B.1. We limited the participation of the study to Democratic Party primary voters for two reasons: 1) the study was timed during the Democratic Primary of 2020, providing an opportunity to collect useful data about voter preferences, and 2) Democratic voters are more likely to have well-articulated preferences for climate policies, due to a combination of significant media coverage on different candidates' platforms during the primary, and the higher priority Democratic Party voters placed on climate change policies.[141–143]

We used a three-stage research design In Stage One, we model the individual preferences of each participant using a standard discrete choice modeling approach. We assumed individual choices followed a multinomial logit process with a linear utility function, which assumes an underlying cardinal utility function. We then estimate their utility function by assuming that the function for each individual i for alternative j is additive such that $u_i(m_j) = \sum_{k=1}^{K} \beta_{ki} x_{kj}$, where x_{kj} is the kth attribute of the jth alternative, M is the set of all policies, and choices are determined by the multinomial choice probability $p_{ij} = \frac{e^{u_i(m_j)}}{\sum_{x \in M} e^{u_i(m_x)}}$.[37,56,144]

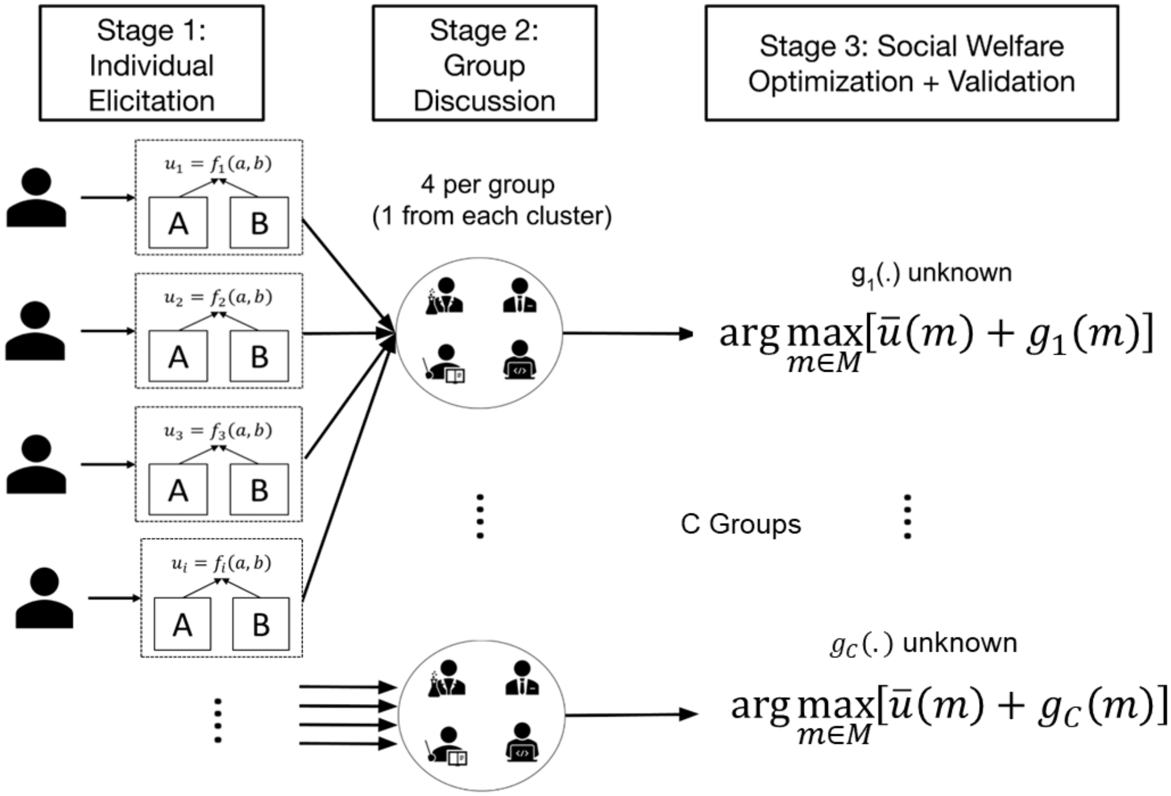


Figure 3-1 - The study used a three-stage research design. In Stage One, we elicited each participant's individual performance using their responses to a 44-question discrete choice task. In Stage Two, after we excluded participants who met our exclusion criteria, we clustered the remaining participants using their Stage One responses in four clusters using a balanced k-means clustering algorithm. Participants are then organized in groups of four, with one participant from each cluster. In Stage Three, groups are assigned to one of the two models (mean-variance or weighted sum) and are asked to first make choices among 15 pairs of policy platforms that they believed would be best for the group. Based on their responses, the model will make a policy recommendation to the group. Once the group decides on whether they will accept the recommendation, the group will validate the recommendation by comparing it against 10 other policies.

Participants completed an online survey where they were asked to choose between 44 pairs of policy platforms (with 2 attention check questions). Participants who did not finish the survey, or who met the exclusion criteria, were not included for the subsequent study stages. The example choice task shown in Figure B-1 (from Appendix B.2). Based on the 44 choices obtained from each participant, we used logistic regression to estimate the 8×1 parameter vector $\hat{\beta}$ of a linear utility function $u(m) = \widehat{\beta^T} m$, where m contains an intercept term (representing the status quo policy platform), plus 2 dummy variables for each of the attributes with three levels, and one variable representing the amount of the carbon tax (treating its six levels linearly). We selected the 44 pairs using a D-optimal design for the first alternative assuming $\beta = 0$, then rotated that design to create the second alternative.[145,146] From pilot testing we expected the task to take approximately 30 minutes, with each two-alternative choice taking approximately 30 seconds, and an additional 5-10 minutes to read the introductory material to the task.

In Stage Two, we used balanced k-means clustering to create equally sized clusters of participants who differed based on their estimated preferences in Stage One.[147] Using each participant's estimated parameters as a vector, the balanced K-means clustering algorithm randomly divided participants into four clusters, then calculated the mean distance of each participant’s estimated parameter from each cluster center. Each cluster then selected the participant that had the lowest mean distance to that cluster. Once all participants were assigned to a cluster, we identified a new center for each cluster, and recalculated the mean distance of each participant’s estimated parameters to the new cluster centers. The new clusters repeat the process of selecting participants that had the lowest mean distance to that cluster. The entire process repeated until cluster centers are no longer different between each iteration. At the end of the stage, all participants were clustered into four distinct clusters.

In Stage Three, participants were invited to sign up for the group task and asked to provide their scheduling availability to the experimenter. The experimenter then created groups by finding times where one participant from each cluster was available. Those four participants (one from each cluster) were assigned a time to meet for the group deliberation. The group was randomly assigned (blind to participants but not the experimenter) to have a recommendation come from one of the two models (weighted-sum or mean-variance). Groups were briefed at the beginning of the experiment that their goal was come to a decision about climate policy platforms. Due to the COVID-19 pandemic, participants were asked to participate online through a videoconferencing application. Participants were asked to not turn on their cameras on their devices, and communication between participants was exclusively through voice. The researcher was present for every group discussion, however, apart from briefing the participants at the beginning of the stage, did not participate in the discussion. There are three phases in Stage Three. In Phase One, participants were asked to choose between 15 pairs of policy platforms that they believed would be best for the group. They had two minutes to discuss each policy platform pair and were asked to make a choice before the end of the two minutes. They were informed that they did not have to agree before making their vote but were encouraged to discuss and share their perspectives with each other.

The 15 pairs of policy platforms selected for group deliberation were selected based on the randomly assigned treatment arm for the group. In the weighted-sum arm, the estimated social welfare function $\widehat{W_{ws}}(m)$ had the form:

$$\widehat{W_{ws}}(m) = \widehat{\alpha_1}u_1(m) + \widehat{\alpha_2}u_2(m) + \widehat{\alpha_3}u_3(m) + \widehat{\alpha_4}u_4(m) \text{ (3-2)}$$

where $u_i(m)$ are the estimated utility functions estimated from Stage One evaluated with respect to an attribute vector x. As participants made choices, a Bayesian logistic regression using stanglm with standard normal priors was used to estimate the 4×1 vector $\widehat{\alpha}$, where individual group-member choices were assumed to be independent and identically distributed Bernoulli random variables with mean $P(L \succ R) = \frac{e^{\widehat{W_{ws}}\left(m^L\right)-\widehat{W_{ws}}\left(m^R\right)}}{1+e^{\widehat{W_{ws}}(m^L)-\widehat{W_{ws}}(m^R)}}$.[148,149]

For the mean-variance arm, the estimated social welfare function $\widehat{W_{mv}}(m)$ was:

$$\widehat{W_{mv}}(m) = \bar{u}(m) + \hat{k} \text{ (3-3)}$$

where $\hat{k} \in \left[-1/\sqrt{4}, 1/\sqrt{4}\right]$ and $\sigma = \sqrt{\frac{\sum_{i=1}^{4}\left(u_i(m)-\bar{u}(m)\right)^2}{n-1}}$, where the constraints on $\hat{k}$ were selected to ensure that the social welfare function would not violate the Pareto condition (as indicated in Monte Carlo simulations). We estimated $\hat{k}$ using constrained logistic regression with the same assumptions as the weighted-sum social welfare function. Using this approach, the 15 pairs of policy platforms were selected, but in a different way for the weighted sum and mean-variance models based on simulations shown in Appendix B.6 indicating that these methods worked well for each method. For the weighted-sum model, we first selected 13 pairs using a D-optimal design for the first alternative, then rotated that design to create the second alternative.[145,146] In the last two pairs we used an upper confidence bound approach and selected the alternative that maximizes $\mathrm{E}(\widehat{W_{ws}}(m)) + \delta \mathrm{Var}(\widehat{W_{ws}}(m))$ with $\delta = 2\log\left(\frac{4n^2\pi^2}{6}\right)$ and n is 14 or 15.[150] The second alternative was then selected by rotating the first.[146] For the mean-variance model we used a D-optimal design to select the first alternative, and rotated that alternative to select the second. Once the group evaluated all 15 pairs of policy alternatives, phase two starts where we estimated the group social welfare function and present the policy recommendation to each group using the models described above.

Our approach does not require every member of the group to explicitly agree on the group's social welfare function, but instead that their behavior can be usefully approximated by such a function. The aim is to create a tool for making recommendations that aid group decision-making as part of an analytic-deliberative process, but not replace the group's final decision.[61] Although different treatment arms generated the recommended policy alternative through different models, there was no visual difference between the treatment arms for participants. Participants made their final decision for the group after being told that they were to evaluate whether the policy platform was the best for the group. Like phase 1, participants had two minutes to discuss, and they made their decision before the end of the two minutes.

The final phase was validation, where each group member independently made an additional 10 choices, but the left policy platform was always the recommended policy platform. In this section, they had only 30 seconds to make their choice for each pair of policy platforms. If the recommendation was truly the best alternative, then participants should always prefer the recommended alternative over the randomly selected alternative.

The study was reviewed and approved by Carnegie Mellon University Institutional Review Board (STUDY2019_00000401). The study was pre-registered and pre-planned using the Open Science Framework on March 4, 2020 (https://osf.io/f96bm/). Our primary hypotheses were that the mean-variance treatment arm would have a greater probability of consensus than the weighted sum arm, and that the average number of participants accepting the recommendation in the mean-variance treatment arm during the validation phase would be greater than the weighted sum arm. For the first hypothesis, consensus was defined as all participants in the group agreeing with the recommended alternative. This between-subjects test was conducted using a two-sample randomization test. For the second hypothesis, there were a total of 40 possible votes in favor of the recommended alternative in each group (10 pairs times four participants per group). Each group received a score from 0-40 and the average score across the two treatment arms was compared using a two-sample randomization test. Our secondary hypotheses were at the individual level. First, we used a mixed logit model to test whether participants in the two treatment arms differed in their probability of accepting the recommended alternative. Each individual's vote within a group was counted as 1 if they vote in favor of the recommended alternative and zero otherwise. We fit a mixed logit model, shown in log-odds form for group c:

$$\log\left(\frac{P(L \succ R)_c}{1-P(L \succ R)_c}\right) = \widehat{\gamma_c} + \hat{\theta} \times 1[\text{Treatment Arm} = \text{MV}]_c \quad (3\text{-}3)$$

where the group-level random intercepts (gamma) follow a multivariate normal distribution to allow for dependence of the choice between members of the same group. Our final pre-registered analysis repeated the mixed logit at the individual level during the validation stage. We used a mixed logit model of the probability of individual i voting in favor of the recommended alternative during the validation phase of the study. The model included both group-level intercepts and individual-level intercepts. We fit a mixed logit model, shown in log-odds form for participant i in group m:

$$\log\left(\frac{P(L\succ R)_{ic}}{1-P(L\succ R)_{ic}}\right) = \widehat{\omega_{\iota}} + \widehat{\tau_c} + \hat{\zeta} \times 1[\text{Treatment Arm} = \text{MV}]_{ic} \quad (3\text{-}4)$$

where $\widehat{\omega_{\iota}}$ and $\widehat{\tau_c}$ are modeled using a multivariate normal distribution to capture any group-level and participant-level dependence of the choices, and $\hat{\zeta}$ is the treatment effect for the mean-variance group.

3.4 Results

Our data collection process started in March 2020 and ended in July 2020. There were 684 participants from the Pittsburgh area who signed up for the study through Carnegie Mellon University's Center for Behavioral and Decision Research (CBDR) participant pool. Of those 684, 111 participants either did not finish the Stage One survey, or they indicated that they were not planning to vote in the Pennsylvania Democratic Party Primary (met the exclusion criteria). Out of the remaining 573 participants invited to participate in Stage 3, 357 participants did not participate or could not finish their study session, due to a combination of reasons, ranging from technical difficulties, to no longer being interested to the study, to not showing up to their schedule study session. 216 participants completed all three stages of the study, who were then divided into 54 groups of four participants each, with 27 groups in each treatment arm.

3.4.1 Stage one individual survey response and stage two participant clustering results

In Stage One, we estimated the individual policy preferences using a discrete choice approach. Participants completed a discrete choice survey between 44 pairs of climate and energy policy platforms for the US. In Stage Two, participants were clustered into four groups based on their fitted coefficients (representing the strength of their preference either for or against a particular attribute) from Stage One to maximize preference heterogeneity in the groups for Stage Three. Based on the estimated individual coefficients (summary statistics of the estimated coefficients can be found in Appendix B.3), the balanced k-means clustering algorithm organized participants into four clusters. We label these clusters (Figure 3-2) Indifferent, Fast Decarbonization, Against Fossil Fuel Exploration, and Against Status Quo (n = 54 for each cluster) based on the characteristics of the estimated coefficients.

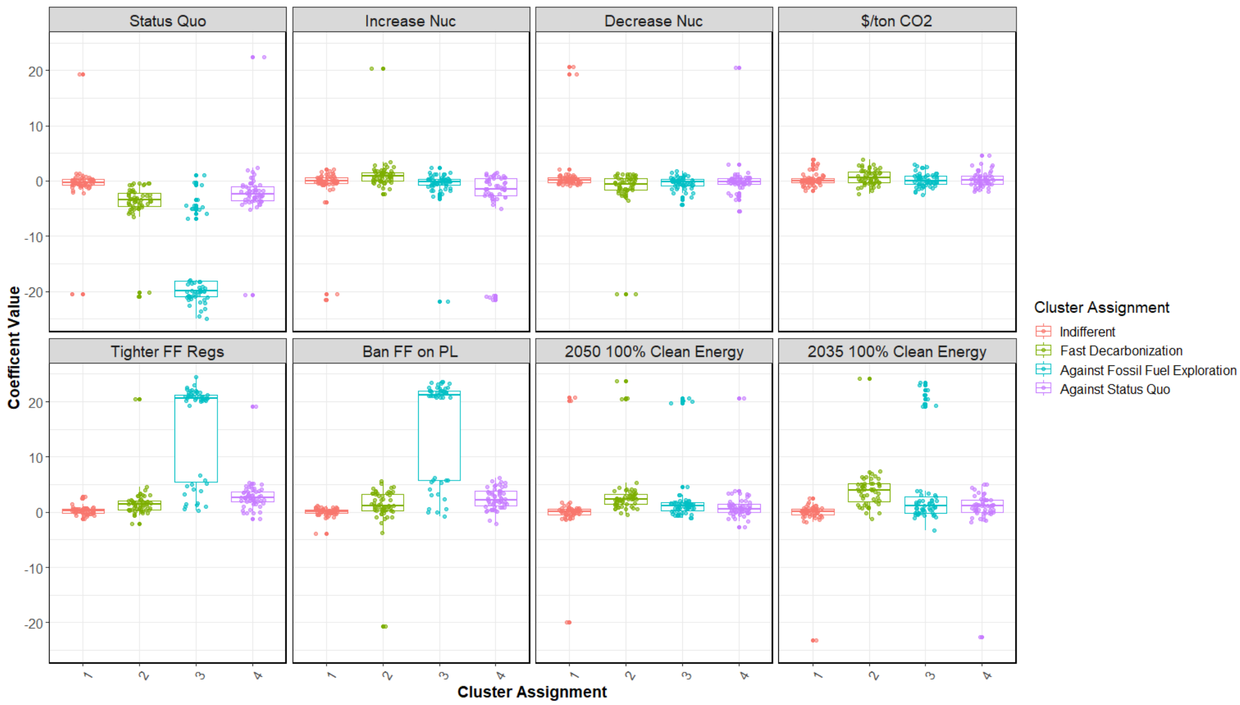


Figure 3-2 - Individual preferences of all participants who participated all three stages of the study, separated by four clusters. Each panel shows the coefficient estimate for each participant modeled through their individual survey results, grouped by Stage Two cluster assignments. For example, individuals in the Against Fossil Fuel Exploration cluster had very high estimated coefficient values for the tighter fossil fuel regulations and ban fossil fuel exploration on public lands attributes, suggesting their strong preference for policies that reduced fossil fuel exploration on public lands. Each dot represents one participant's estimated coefficient value for each attribute level. The box represents the interquartile range (IQR) while the whiskers represent min/max value $+/- 1.5 * IQR$ for coefficient estimates for each cluster in each panel.

To quantitatively characterize the differences between the clusters, we used Cohen's *d* to estimate the standardized mean difference in estimated coefficients across the clusters for each policy attribute (Table 3-1).[151] We see that every pair of clusters had a large difference on at least one attribute, with the attributes with most disagreement across clusters being tighter fracking regulations, ban of fossil fuel exploration on public lands, the status quo, and increasing nuclear power. Smaller differences emerged for decreasing nuclear power, the clean energy timeframes of 2050 and 2035, and carbon price. Only 9 of the 48 (19%) pairwise comparisons across clusters for all attributes had less than a small difference ($d < |0.2|$), suggesting that the clustering algorithm was able to create differentiated groups of participants with different policy preferences.

Table 3-1 - The table reports the Cohen's d value between each cluster pairs for the cluster's estimated coefficient for all attribute levels. Red cells represent clusters that have a large difference between them for that specific attribute, yellow cells represent medium difference, green cells represent small difference, and white cells represent no difference. For example, for the Tighter FF Regulations attribute, the Cohen's d value between the Indifferent cluster and the Fast Decarbonization cluster is -0.70, which indicates that for these two clusters, they had a medium difference for the Tighter FF Regulations attribute. The table shows there are significant differences between all cluster pairs for many attributes, showing that participants in each cluster generally have different policy preferences.

	Cluster Pairs					
Attributes	Indifferent – Fast Decarbonization	Indifferent – Against FF Exploration	Indifferent – Against Status Quo	Fast Decarbonization – Against FF Exploration	Fast Decarbonization – Against Status Quo	Against FF Exploration – Against Status Quo
Tighter FF Regulations	-0.70	-8.01	-1.26	-7.31	-0.56	6.74
Ban FF Exploration on Public Lands	-0.43	-7.40	-0.92	-6.98	-0.49	6.48
Status Quo	0.77	2.49	0.28	1.94	-0.51	-2.33
Increase Nuclear Power	-0.54	-0.03	0.71	0.51	1.24	0.73
Decrease Nuclear Power	0.57	0.33	0.39	-0.24	-0.18	0.06
2050 100% Clean Energy	-0.60	-0.40	-0.15	0.20	0.45	0.25
2035 100% Clean Energy	-0.51	-0.42	-0.04	0.09	0.47	0.38
$/ton CO2	-0.38	0.04	-0.07	0.42	0.31	-0.11

$d < |0.2|$ No Difference; $|0.2| \, d < |0.6|$ Small Difference; $|0.6| \, d < |0.8|$ Med. Difference; $d > |0.8|$ Large Difference

3.4.2 *Pre-registered analysis results*

Our pre-registered analyses refer to Stage Three of the study, where participants deliberated in groups of four and made choices over 15 pairs of climate and energy policies, spending two minutes deliberating about each pair. The weighted sum and mean-variance approaches were then fit to those choices to estimate the group social welfare function, and the scenario that maximized the social welfare function according to those choices was recommended to the group. Group members either accepted or rejected the recommendation (the group recommendation), then made 10 pairwise comparisons between the recommended policy and randomly selected alternatives (validation).

Group recommendation. The results from our first pre-registered analysis found that when compared to groups assigned to the weighted sum approach, groups assigned to the mean-variance approach were more likely to reach consensus on the final group recommendation, though the result is not statistically significant at our pre-registered alpha of 0.05 (Table 3-2). Comparing the two treatment arms, 19 groups out of 27 reached consensus (70%) for the mean-variance model, compared to 14 groups out of 27 (52%) for the weighted sum model. Applying the exact permutation test revealed that the difference between the two groups is not statistically significant, with a p-value of 0.26. A two sample, two-sided t-test of proportions found a similar result (t (54) = 1.35, p = 0.18).

On the individual level, our second pre-registered analysis reveals a similar pattern for the final group recommendation. Individuals in groups with the mean-variance model were more likely to agree with the recommended policy when compared to individuals in groups with the weighted sum approach (Table 3-2). We find that 92 out of 108 (85%) individuals agreed with the recommended policy in groups with the mean-variance model, compared to 74 out of 108 (69%) individuals who agreed with the recommended policy in groups with the weighted sum model. When using a mixed-logit model with random intercepts at the group level, the result is not statistically significant (Z = 1.77, p = 0.08).

Table 3-2 - Group and individual pre-registered analysis results for both acceptance of the model recommendation and the validation scores

	Group Result (Permutation Test)				**Individual Result (Mixed Logit)**			
	Recommendation		**Validation**		**Recommendation**		**Validation**	
	Mean-variance	Weighted Sum	Mean-variance	Weighted Sum	Mean-variance	Weighted Sum	Mean-variance	Weighted Sum
Number of Obs	27				108			
Accepted Rec. (%)	19 (70%)	14 (52%)	-	-	92 (85%)	74 (69%)	-	-
Mean Validation Score	-	-	34.6/40 (87%)	31.4/40 (79%)	-	-	8.71/10 (87%)	7.88/10 (79%)
Treatment Difference	19%		3.2		17%		0.83	
p-value	0.26		0.03		0.08		0.06	
T-Test Value	-1.35		-2.20		1.77		1.87	

Validation. The validation score for each group was the number of the 40 pairwise comparisons (10 pairwise comparisons for each person, 4 people) that were chosen in favor of the recommended alternative. When examining the validation scores for each group, our third pre-registered analysis found that groups assigned to the mean-variance approach generated higher validation scores when compared to groups assigned to the weighted sum approach (Table 3-2). Groups assigned to the mean-variance approach had an average validation score of 34.6, while groups assigned to the weighted sum approach has an average validation score of 31.4. This difference is statistically significant at the 0.05 level, with the exact permutation test p-value = 0.03. Similarly, a two-sided two-sample t-test had a t-value of -2.29 (p = 0.03). This suggests that groups with the mean-variance model were more likely to remain satisfied with their recommendation when provided with other alternatives, consistent with the previously discussed higher (though not statistically significant) rate of consensus for groups with the mean-variance model.

Our final pre-registered analysis found that when compared to those who were provided recommendations by the weighted sum model, individuals in groups with the mean-variance model were more likely to remain satisfied with the recommendation (Table 3-2). An individual in a group with the mean-variance model had an average validation score of 8.71 (out of 10), while an individual in a group with the weighted sum model had an average validation score of 7.88 (out of 10). Using a mixed-logit model with random intercepts at the individual and group level, the difference was however not statistically significant ($Z = 1.87$, $p = 0.06$). Even so, the result suggests that individuals who were provided recommendations using the mean-variance approach were more likely to be satisfied with the recommendation.

3.4.3 Qualitative observations of stage three discussions

We examined how participants changed their policy preferences by comparing their most preferred policy before group deliberation to the group recommendation made after group discussion (Figure 3-3). Figure 3-3 shows the number of participants whose most preferred policy prior to the group discussion was different from the group recommendation by each attribute, with shifts toward that attribute in blue, and away from that attribute in red. We see that all four clusters saw at least 16 participants change at least one of their attributes. This suggests that each group’s recommendation was not dominated by people from a particular cluster, as participants from all clusters modified at least some of their preferences and group discussions changed some participant preferences.

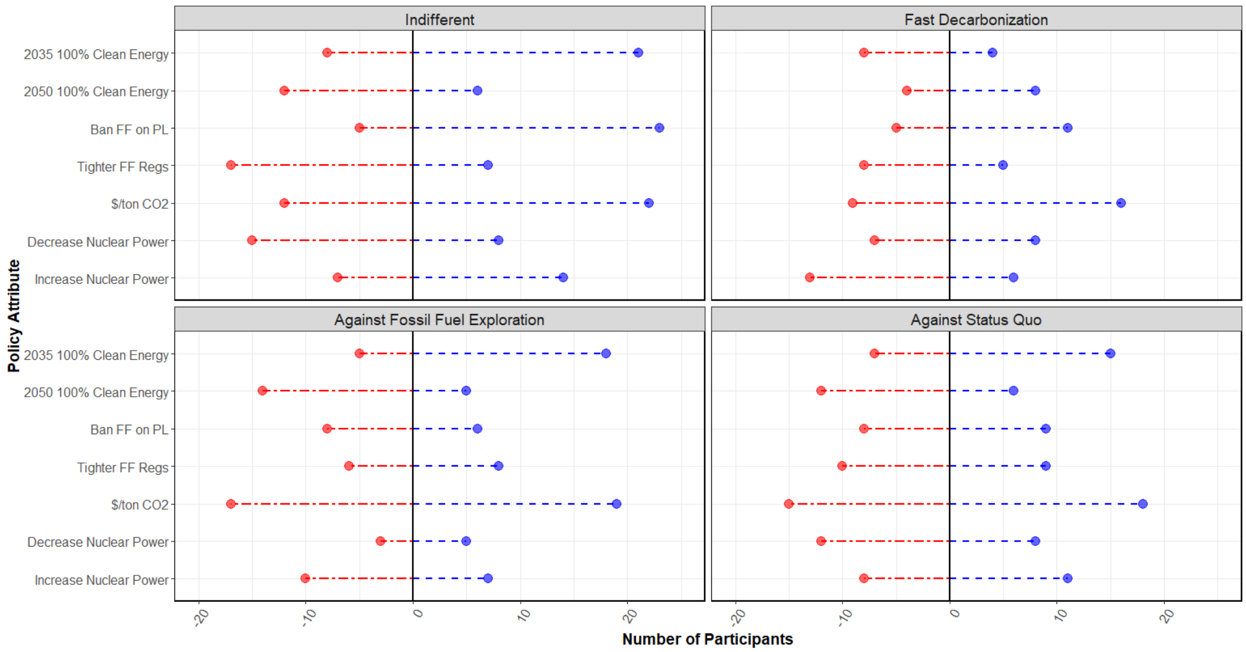


Figure 3-3 - Participants who changed their preferred policy attribute for each cluster. Each row represents the number of participants where their most preferred policy changed following the group discussion. Red line denotes the number of participants who changed their preferred policy away from that row, while the blue line denotes the number of participants who changed their preferred policy towards that row. For example, if a participant's most preferred policy prior to the group discussion was a 2100 100% Clean Energy Portfolio Standard goal, and following the group discussion, the recommendation was a 2035 100% Clean Energy Portfolio Standard goal, that participant would be counted in the blue section of 2035 100% Clean Energy row.

In addition, while the models generated a diverse set of recommended policies (Table 3-3 and Table 3-4), we can see some broad patterns of suggested policies by tallying the total number of policies recommended across the two groups for each of the policy attributes.

Table 3-3 - List of most common recommendations offered to the groups, categorized by their respective attribute levels.

Nuclear Power	100% Clean Energy Portfolio Standard	CO2 Price ($/ton)	Fossil Fuel Exploration on Public Lands	Count
Increase Nuclear Power	2035 100% Clean Energy Portfolio Standard Goal	150	Ban Fossil Fuel Exploration on Public Lands	6
Reduce Nuclear Power	2035 100% Clean Energy Portfolio Standard Goal	150	Tighter Regulations for Fossil Fuel Exploration	4
Maintain Nuclear Power	2035 100% Clean Energy Portfolio Standard Goal	0	Ban Fossil Fuel Exploration on Public Lands	3
Increase Nuclear Power	2050 100% Clean Energy Portfolio Standard Goal	150	Ban Fossil Fuel Exploration on Public Lands	3

Table 3-3 shows that only a few groups received the same policy recommendation, with the top policy only recommended to six out of 54 groups. In addition, examining the most common recommendations, we can see that there are some variations in the attribute levels (examined further in Table 3-4). The top two recommendations had the same attribute level for CO2 Price and 100% Clean Energy Portfolio Standard Goal, while the next two most common recommendations varied in both CO2 Price and 100% Clean Energy Portfolio Standard Goal. This shows that the recommender system generated different recommendations based on the participant individual preferences and the result of their group discussion.

Table 3-4 - Final recommended policy at the end of group discussion, tallied by the number of groups being offered a specific attribute for the final policy, separated by model, organized by whether the group reached consensus during the group discussion.

Attribute	**Level**	**Weighted Sum Model**		**Mean-Variance Model**	
Consensus?		*All*	*Only Consensus*	*All*	*Only Consensus*
Nuclear Power	Increase Nuclear Power	10 (37%)	**8 (58%)**	**10 (37%)**	6 (32%)
	Decrease Nuclear Power	6 (22%)	3 (21%)	8 (30%)	**8 (42%)**
	Maintain Nuclear Power	**11 (41%)**	3 (21%)	9 (33%)	5 (26%)
CO2 $ / ton	0$/ton	3 (11%)	0 (0%)	4 (14%)	2 (11%)
	$30/ton	6 (22%)	2 (14%)	1 (4%)	1 (5%)
	$60/ton	2 (8%)	1 (7%)	1 (4%)	0 (0%)
	$90/ton	1 (4%)	0 (0%)	1 (4%)	1 (5%)
	$120/ton	6 (22%)	4 (29%)	2 (8%)	1 (5%)
	$150/ton	**9 (33%)**	**7 (50%)**	**18 (66%)**	**14 (74%)**
Fossil Fuel Exploration on Public Lands	No Restrictions	1 (4%)	1 (7%)	0 (0%)	0 (0%)
	Tighter Regulations	6 (22%)	3 (21%)	11 (41%)	9 (47%)
	Ban Fossil Fuel Exploration on Public Lands	**20 (74%)**	**10 (72%)**	**16 (59%)**	**10 (53%)**
100% Clean Energy Portfolio Standard Goal	2100 Goal	3 (11%)	1 (7%)	2 (8%)	2 (11%)
	2050 Goal	5 (19%)	3 (21%)	4 (14%)	3 (15%)
	2035 Goal	**19 (70%)**	**10 (72%)**	**21 (78%)**	**14 (74%)**
Number of Groups		27	14	27	19

Aside from the nuclear power attribute, there was very little difference between the types of policies recommended depending on the group's acceptance of the recommendations. When examining the recommendations more closely, there were large differences across the levels of the with Nuclear Power (mean variance model had increase nuclear power as the most frequently recommended attribute level, while weighted sum model had maintain existing nuclear power as the most frequent recommended level) and CO2 Price attributes (mean variance model recommended $150 per ton CO2 price for 66.6% of groups, while weighted sum model recommended $150 per ton CO2 price for 42.2% of groups), while the Fossil Fuel Exploration on Federal Lands and the Clean Energy Standard Portfolio saw very little difference between the models. This pattern persisted even if we limit the data to groups who agreed with the recommendation (Columns labeled *All* versus *Only Consensus*).

When examining the types of policies that were recommended to the groups, and focusing on the policies that were accepted by the groups, we can see most groups were offered a 2035 100% Federal Clean Energy Portfolio Standard (74% for all groups, and 73% for groups who accepted the recommendation), and a small minority of groups (9% for all groups and 9% for groups who accepted the recommendation) were offered a 2100 100% Federal Clean Energy Portfolio Standard. In addition, most groups were offered banning fossil fuel exploration on federal lands (67% for all groups, and 61% for groups who accepted the recommendations), with only a small minority offered a policy with no restrictions for fossil fuel exploration on public lands (2% for all groups, and 3% for groups who accepted the recommendation). Finally, we see that many groups were offered a $150/ton CO2 price, suggesting the viability of a CO2 price policy (50% of all groups, 64% for groups that accepted the recommendation).

We see greater disagreement with the use of nuclear power for groups across the models. 37% of all groups were provided with a policy that increased the use of nuclear power, and 26% of all groups were provided with a policy that decreased the use of nuclear power. When focusing only on groups that agreed with the recommendation, we see 42% of groups agreeing with increasing nuclear power, and 33% of groups agreeing with decreasing nuclear power. This shows that there is large between-group disagreement on the use of nuclear power.

3.4.4 Effectiveness and efficiency of selected policies

Based on the attributes and policies that were most commonly recommended to participants, we performed a first-order analysis on their effectiveness (measured as cumulative CO2e reduction between 2020-2050 compared to a business-as-usual pathway), and efficiency (measured as total net cashflow of non-government sector between 2020-2050 compared to a business-as-usual pathway). While the measurement for effectiveness is self-explanatory, financial measures of efficiency is harder to define. Traditional economic measurements such as GDP or total spending should not be used to determine whether a policy package is “good” or “bad”, as the type of spending that drives GDP can be both good (capital investment in automation) and bad (increased medical services spending due to declining air quality). In addition, evaluating policy efficiency through changes to government revenue also paints an incomplete picture of the policy packages. Therefore, we chose non-government spending as the metric to measure policy efficiency, since it evaluates the direct industry impact of these policies, without making assumptions on how government revenue from these policies will be used.

Our policies were evaluated using the Energy Policy Simulator (EPS), created by Energy Innovation LLC (https://us.energypolicy.solutions/scenarios/home).[152] The simulation is an open-source computer model that allows users to control numerous policy levels that affect energy use and emissions and observe their result on a wide range of metrics, including greenhouse gas emissions reduction and financial outcomes. Some of the policy levers include 100% clean energy portfolio standard (with variable years of implementation), capital investment and fuel subsidy for specific power generation sources (nuclear for example), nation-wide carbon tax, fuel subsidy or tax and much more.[153] Unfortunately, the model does not allow the user model a decrease in fossil fuel exploration on federal lands. Instead, we integrated the model results from *Prest* on evaluating supply-side policy changes to oil and gas production on US federal lands.[154]

When examining each attribute separately and selecting the attribute level that was the most widely recommended, we see that they had very different levels of effectiveness and efficiency. As seen in Figure 3-4, a carbon price of $150/ton has the highest potential for cumulative CO2e reduction, followed by a 2035 goal for 100% clean energy portfolio. Despite its popularity, a ban on fossil fuel exploration on public lands have very little effect in overall CO2e emission reduction. This is due to the fact that fossil fuel exploration on public lands represent a small fraction of extraction activities, and a ban would be offset by a small increase in production from private lands.[154] This also means that a fossil fuel exploration ban on public lands will have very little price effects, as foreign production will increase to close the gap.[154]

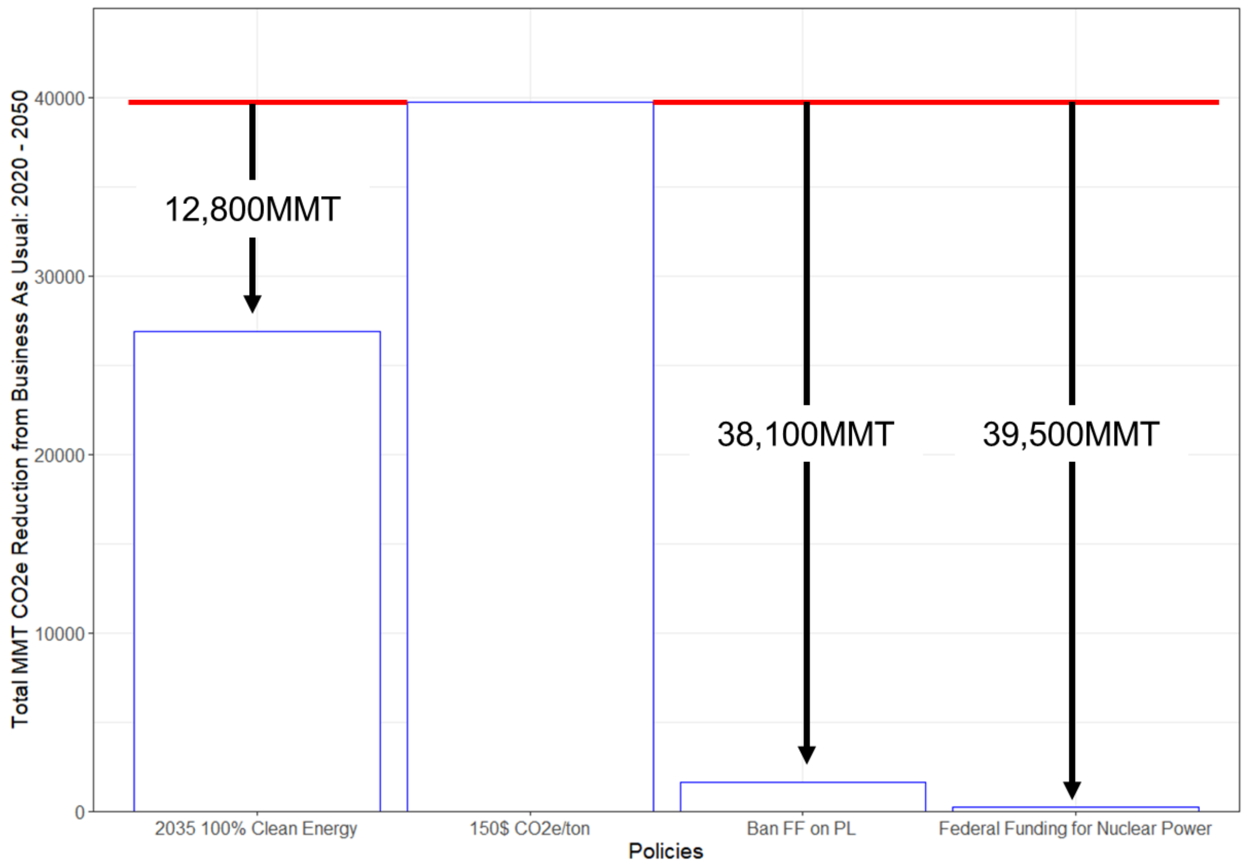


Figure 3-4 – Cumulative CO2e emission reduction (2020-2050) compared against a BAU case for the most popular attribute levels for all four attributes (2035 100% Clean Energy Portfolio Standard, $150 CO2 Price/ton, Ban Fossil Fuel Exploration on Public Lands, and Increase Federal Funding for Nuclear Power). Arrows represent the difference between the policy with the highest reduction in CO2e emissions (Carbon Pricing) against the other attributes.

When examining the efficiency of the different policy attributes (see Figure 3-5), we can see that a carbon price of $150/ton leads to the highest non-government spending out of the four policies. On the other hand, the 2035 100% Clean Energy Portfolio Standard generates a positive non-government spending, since this policy package allows industries to cut their fuel spending significantly over the long haul. Banning fossil fuel exploration on public lands also has some cumulative cost, due to companies losing revenue on future exploration opportunities.

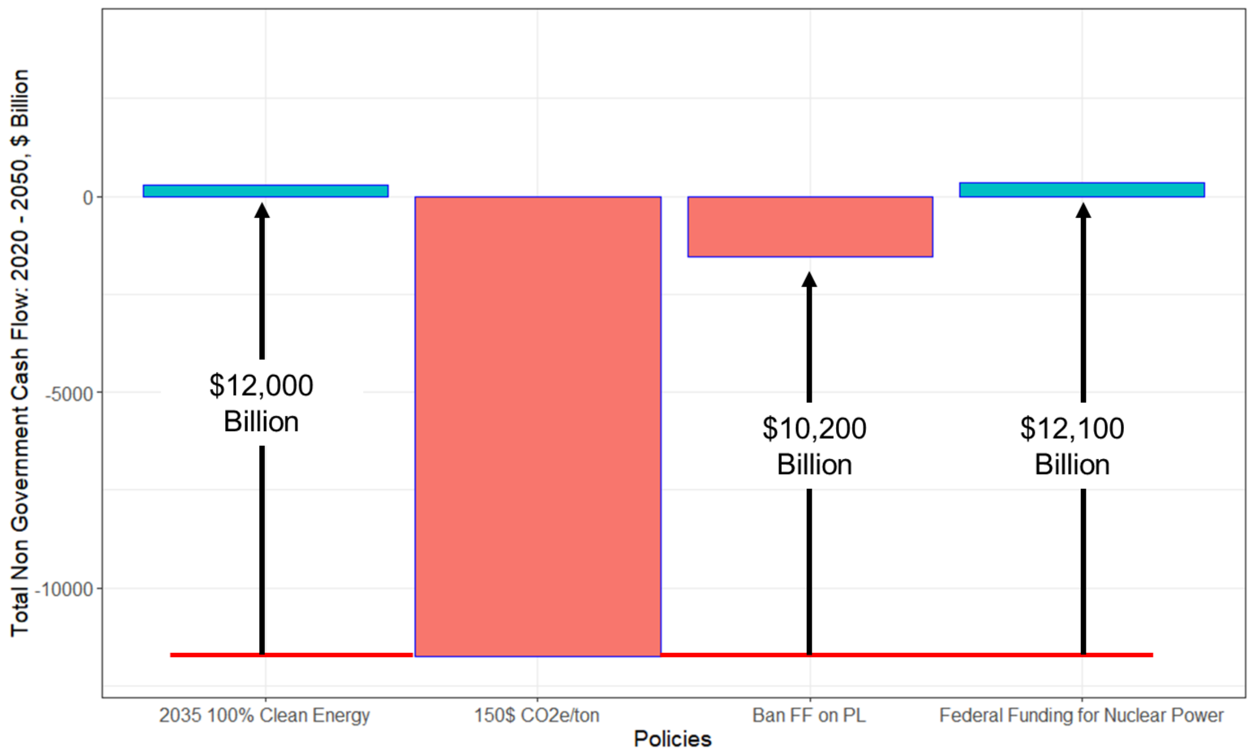


Figure 3-5 – Cost efficiency of different policy attributes measured in total non-government cash flow from 2020-2050. Red represents a negative cash flow, while blue represents a positive cash flow. Black arrows represent the difference between the costliest policy (CO2 price of 150$/ton) with the other three policies.

It is important to note that when examining a combination of these policy attributes, their effects are not additive. This is because there is significant overlap in public response to these policies. For example, behavioral changes due to a high carbon price would also be present in a world where there is a 100% Clean Energy Portfolio Standard.

When examining the top four most commonly recommended policy packages (see Table 3-3 for the top four policy packages), we see that the most commonly recommended policy package has both the highest cumulative reduction in CO2e emissions and the highest negative non-government cash flow (Figure 3-6 and Figure 3-7). In addition, while the policies themselves are different, the only attribute different that led to a significant difference in both CO2e emissions reduction and non-government cash flow is the carbon price.

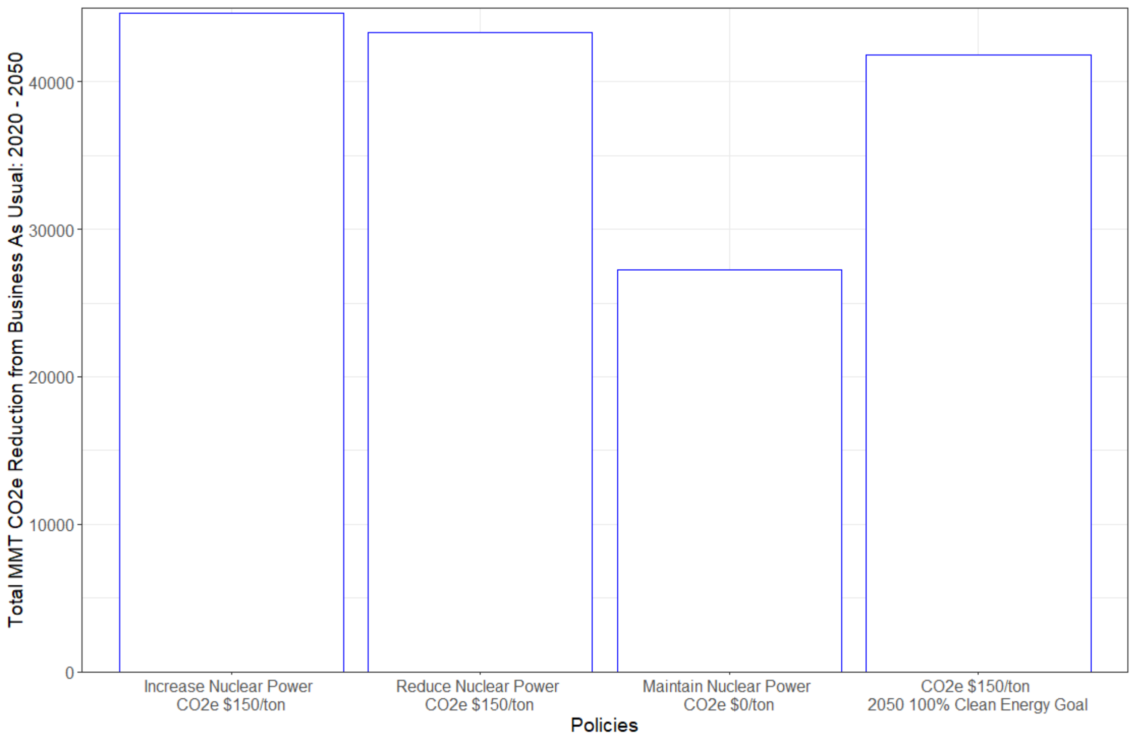


Figure 3-6 – Top four most common policies recommended to all groups. Left most column represents the most common policy recommended, while the right most column represents the fourth most commonly recommended policy.

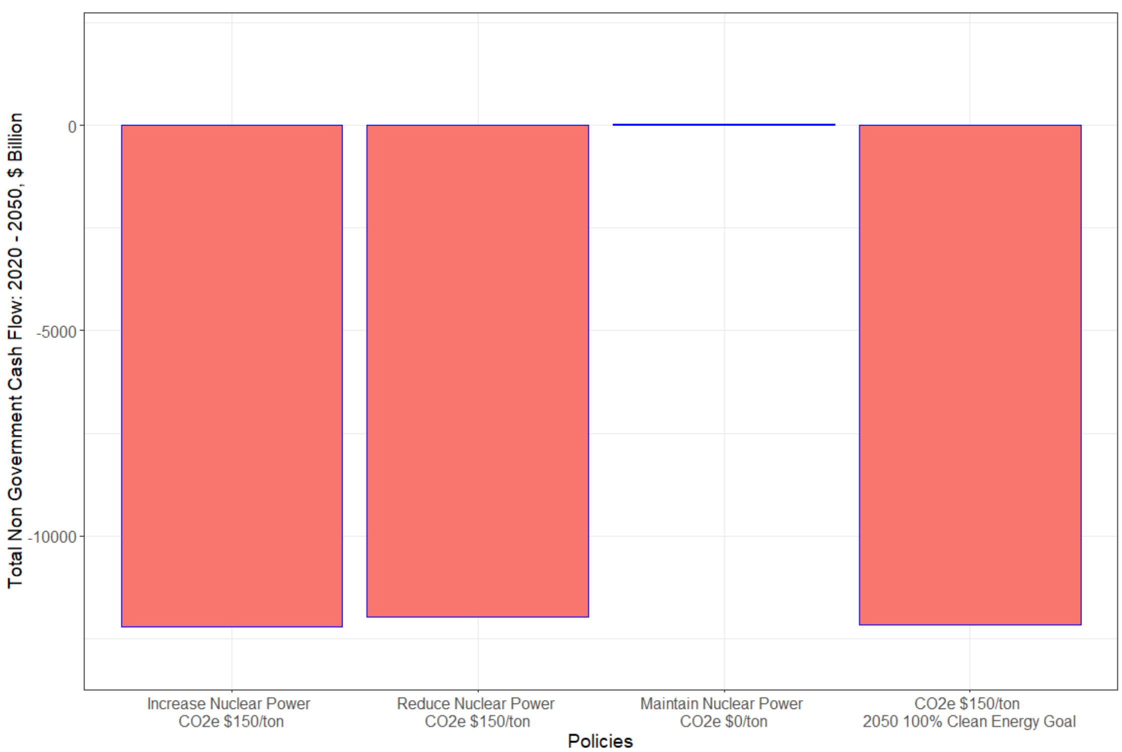


Figure 3-7 – Non-government cash flow from 2020-2050 for four most common recommended policies. Left most column represents the most commonly recommended policy, while the right most column represents the fourth most commonly recommended policy. Red represents a negative cash flow, while blue represents a positive cash flow.

In addition, we can examine the effectiveness and efficiency metrics of all the accepted group policies and compare those values to the most preferred policy metrics of the individuals in those groups. In Figure 3-8, every policy recommendation that was accepted by a group was plotted based on their efficiency and effectiveness values. The size of the circle represents the number of times those recommendations were accepted. We see that for most of the accepted policies, they had both high reduction of CO2 emissions and high private sector spending. This was expected as the framing of the study would naturally lead to participants to prefer policies that would lead to greater CO2 reduction.

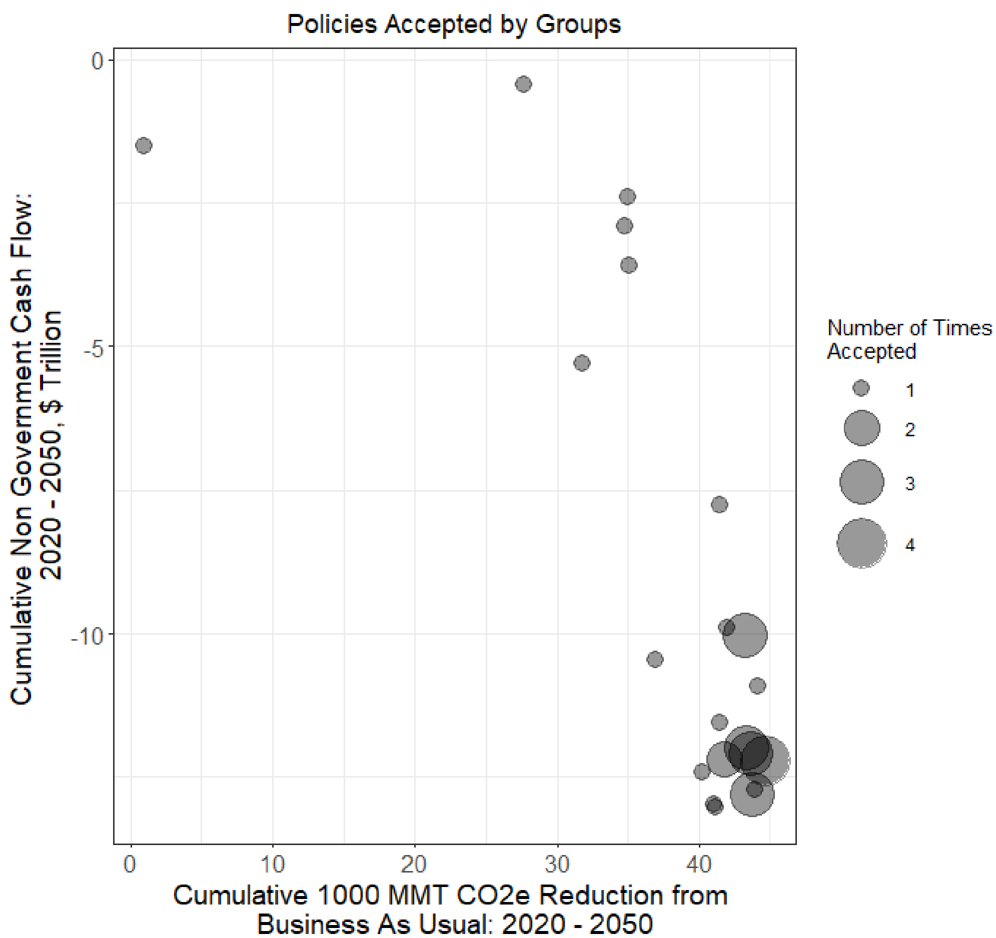


Figure 3-8 Policy effectiveness and efficiency of all accepted recommendations. The size of the circle represents the number of times a specific policy was accepted by the group.

In Figure 3-9, we first identified the participants in groups that accepted the recommendations, then plotted their most preferred policy's effectiveness and efficiency metrics prior to the group discussion. Similar to the group result, we see that most of these participants preferred policies that had high private sector spending and CO2 reduction. We can then take the group recommendation metrics from Figure 3-8 and subtract it from the individual metrics from Figure 3-9 to see if participants made any tradeoffs in order to get to consensus.

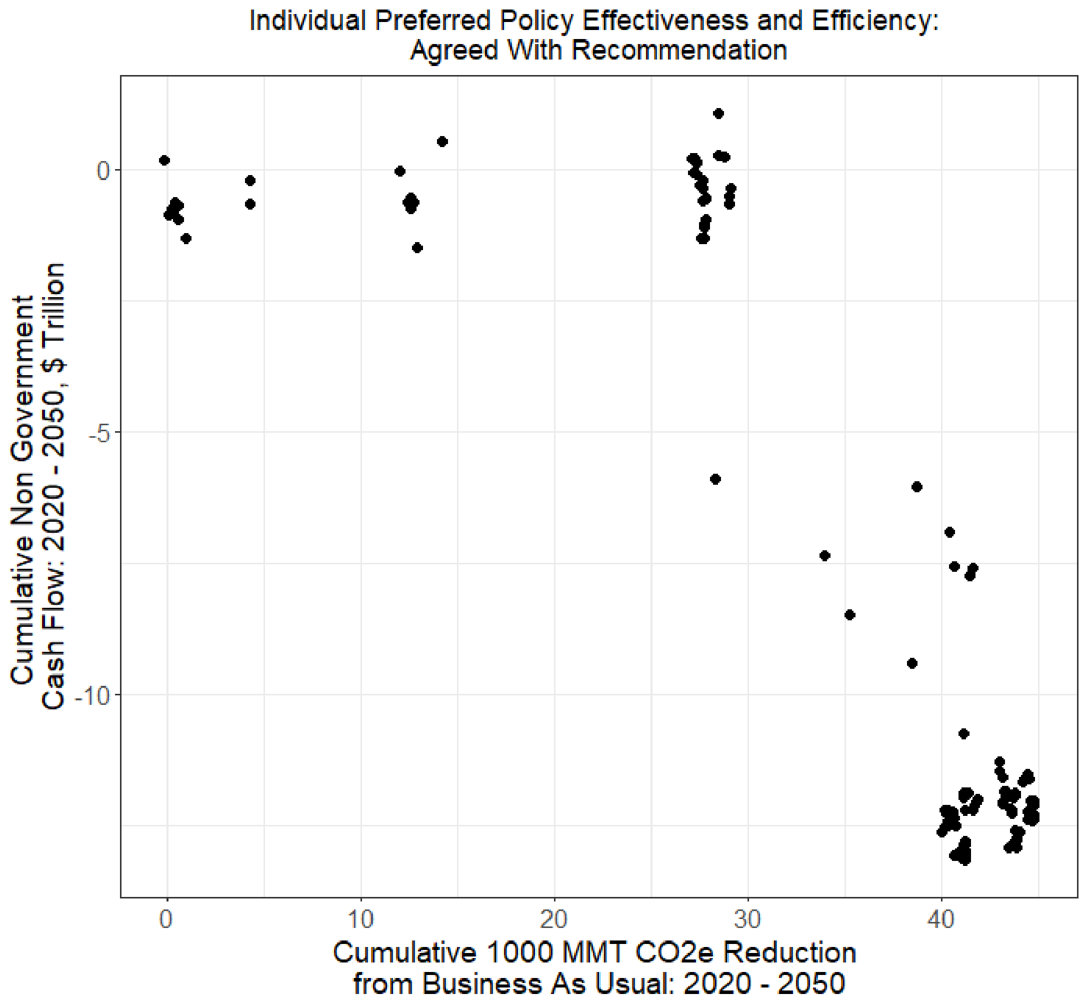


Figure 3-9 Most preferred policy efficiency and effectiveness metrics, of individuals in groups that accepted the recommendation. Each dot represents a specific individual.

In Figure 3-10, we can see the difference between the individual policy prior to the discussion and the accepted recommended policies. For most of these participants (71 out of 132, 54%), the group recommendations had higher emissions reduction and higher private sector spending. This suggests that rather than participants having to make tradeoffs during the group discussion, participants actually preferred greater emission reduction, and the group discussion actually helped them to identify policies that eventually lead to that outcome.

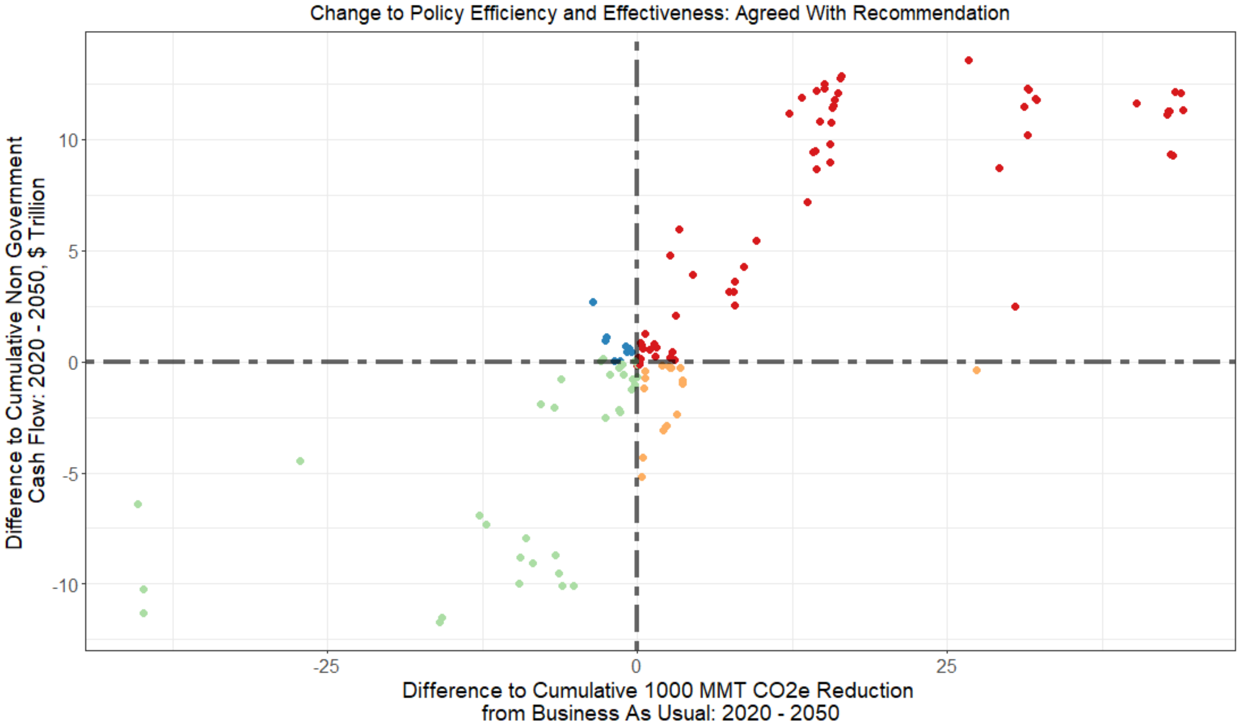


Figure 3-10 Change to policy efficiency and effectiveness of those who agreed with the recommendation. Each dot represents an individual who accepted the group recommendation. Individuals in the top right quadrant saw an increase in both effectiveness and efficiency when compared to their group recommendation.

3.5 Discussion

Our study examined the feasibility of dynamically learning a group's preferences for public policy in the form of an estimated social welfare function, then using those preferences to provide recommendations to a group whose members have heterogenous preferences for climate and energy policy. We tested two approaches for estimating social welfare functions, the mean-variance, which allows for inequality aversion, and weighted sum, which places different weights on each group member. Overall, we find suggestive but not statistically significant differences in the probability of consensus across the two approaches (70% vs 52%) and satisfaction with the recommendation (34.6/40 vs 31.4/40). While the exact permutation test showed that these results were not statistically significant at the alpha = 0.05 level, they do hint that the mean-variance model's ability to approximate the group's social welfare functions with the forms of min, max, and any function in-between allowed it to better capture the aggregated group preference and find recommendations with a high probability of consensus. In addition, the mixed-logit model found that found that while the results were not all statistically significant at $p < 0.05$ ($p = 0.09$ for group consensus, $p = 0.04$ for validation), the mean-variance model performed better than the weighted sum model to find policies that each participant agreed with (85.1% vs 68.5%) and that each participant was satisfied with (8.71/10 vs 7.88/10). Despite the lack of statistical significance for some of the results, likely due to the number of groups we were able to gather data from, there is evidence that groups provided recommendations with the mean-variance model were more likely to reach a consensus and more likely to remain satisfied with the recommendation.

Despite a wide range of personal preferences with the different climate change policies that were proposed during the 2020 Democratic Party Presidential Primary, our study found several policies that enjoy broad support for most participants in our sample. Most participants supported a 2035 100% Federal Clean Energy Portfolio Standard timeline, with almost 73% of groups accepting a policy recommendation with that attribute. When compared against their pre-deliberation preferences, 56 participants across all four clusters changed their preferences from either a 2050 or 2100 goal to a 2035 goal because of the group discussion. This result is higher than recent polling for federal action to reach a 100% Clean Energy Grid for US as a whole, and with Western Pennsylvania locally, though this could be a result of the partisan lean of the study population.[129,155] In addition, most participants supported a full fossil fuel exploration ban on public lands, with almost 61% of groups accepting a policy recommendation with that attribute. When compared against their pre-survey preferences, 41 participants across all four clusters changed their preferences to a ban because of the group discussion. This is much higher than a contemporary poll on fossil fuel exploration on public lands for American voters, though once again this could be a result of the partisan lean of the study population.[156] These results suggest that for our participants (a convenience sample of registered Democrats who intend to vote in Southwestern PA), people with different policy preferences can come to an agreement with the aid of group discussion and a group recommender system. The policies that saw high levels of support included a 2035 100% clean energy portfolio goal and a ban on fossil fuel exploration on public lands. In addition, these results are also significant due to the geographic nature of the study, where the vast majority of participants live in Southwestern PA, an area that have significant ties with the fossil fuel industry, and voters often reward political candidates that are seen as moderates.[157] In addition, the first-order techno-economic analysis of these preferred policies showed that there is a greater affinity to policies with greater CO_2 emissions reduction, which would be accompanied by greater private sector spending.

There are several shortcomings with our study. First, due to time constraints, we were not able to evaluate the social welfare optimization approaches against a control group, either being asked to come up with a group policy themselves, or with a policy recommended at random. While this makes it difficult to examine whether providing any recommendation (even a random one) might lead to consensus, the difference between the mean-variance and weighted sum approaches suggests a benefit of using a non-linear approximator to social welfare functions. Second, due to the onset of the COVID-19 pandemic, we could not conduct the group discussion session in person, and participants instead relied on using tele-conferencing software. Because participants only communicated through voice rather than through both voice and video, non-verbal communications cues were not available to participants. Third, despite our best efforts, there is always a risk that participants did not fully understand the nuances of the different levels within each attribute, especially in abstract attributes such as the carbon price. That could lead to participants selecting a policy that is not reflective of their actual preferences. We observed during the discussion stage that some participants had trouble with the carbon price attribute, both in terms of who will pay it (consumers vs. businesses), and how it relates to their own lives. This might explain the relatively low importance that participants placed on that attribute in their individual surveys, and how that attribute had the least difference across all four clusters. Finally, the exclusion criterion of the study led to only Pittsburgh residents who were planning to vote in the presidential primary campaign to participate in the study. Therefore, we could not evaluate the effectiveness of the models to find consensus when polarized political groups are trying to find common ground. For large societal issues such as climate change, it is impossible to ignore the role of polarization on the likelihood of finding policies that have bi-partisan support, especially in an environment of record-high level of political polarization, often preventing constructive discussion and search for mutually acceptable solutions.[158,159] However, we felt tackling the issue of political polarization is beyond the scope of this project, and wanted to focus more narrowly around the policy deliberation problem itself, where participants are motivated to come to a consensus.

Our study finds that there is some promise in a modeling approach to identify policies that can help to generate group consensus. While our study focused on climate change, this approach can also be used in other areas where there is a heterogenous set of stakeholder preferences such as healthcare and immigration. However, it is important to note that these recommendations are just suggestions, and should not be used in a prescriptive manner, but rather to aid group decision-making instead of replacing it. The value in the approach is in the narrowing of the entire policy space (often incredibly large) and identify policies that have the foundations for a potential consensus, thus reducing the time and effort to identify policies that have the potential for agreement.

3.6 Data Availability

Statistics were done using R 3.5.0 (R Core Team, 2018), the dplyr, tidyr, googleAuthR, googledrive, googlesheets4, shinyjs, lubridate, rstanarm, mlogit, survival, DoE.base, DoE.wrapper, Metrics, mgcv, support.CEs, mlmRev, lme4, rlist, Matrix, matrixStats, pROC, purrr, magrittr, gargle, httr, readxl, fmsb, gridExtra, ggpubr, shiny, shinydashboard, ggplot2, and Hmisc, packages.[106,107,110,111,113,114,148,160–187]

All data, code, and stimuli are available on GitHub: https://github.com/powerguo/PolicyRecommendation

Chapter 4 Automating Subsystem Consolidation Evaluations and Part Selection for Metal Additive Manufacturing[4]

[4] This chapter is currently being prepared for publication (February 2022). This chapter should be referred to as: Guo, N., Jung, S., Funk, P., Davis, A. "Automating Subsystem Consolidation Evaluations and Part Selection for Metal Additive Manufacturing".

4.1 Abstract

Metal additive manufacturing (MAM) could revolutionize military operations by reducing production costs and fuel use while increasing readiness and capabilities. Not only existing conventional manufacturing parts can be replaced with parts using MAM, but subassemblies can also be consolidated using MAM, further reducing cost and weight. However, not all parts can be printed using MAM for military operations. Therefore, there is still a need to scrutinize parts and assemblies that are appropriate for MAM. Extending the prior work done by *Funk et al.* and *Guo et al.*, we elicit and aggregate expert judgment from different domains to understand what attributes are important for MAM and for part consolidation. Specifically, we 1) use semi-structured interviews to filter for attributes that are specific to the Army context, 2) expand our existing techno-economic and expert decision models to incorporate Army priorities and tech authority procedure, and 3) prototype an interface, seeded with expert knowledge and our existing models, to provide an Army database that experts can access and evaluate parts most for their suitability for MAM and parts consolidation. Our results showed that for both questions, experts valued both technical and logistic attributes when making decisions. For part selections, experts considered technical attributes such as the presence of overhangs along with logistical attributes such as OEM purchase price when selecting parts for MAM. For subassembly consolidation, attributes such as the presence of overhangs and the number of parts in the consolidation were used by experts to make their decisions. In addition, our result shows that existing heuristic rules that either focus on ease of printing or economic results do not fully capture the expert decision rules when making their judgment. Instead, when aggregated, experts use a few attributes from different domains to make their decisions. Despite individual experts employed different heuristic rules, aggregated results found clear and consistent decision-rules that are important. This further reinforces prior findings that aggregate expert knowledge have the potential to increase the speed of technology adoption in emerging technologies.

4.2 Introduction

Metal additive manufacturing (MAM) could revolutionize military operations by reducing production costs and fuel use while increasing readiness and capabilities.[188–191] Not only existing conventional manufacturing parts can be replaced with parts using MAM, but subassemblies can also be consolidated using MAM, further reducing cost and weight.[192,193] While the US military has deployed AM labs into the field by the Army, Navy and Department of Defense (DoD) contractors since 2012, there is still significant opportunities for MAM technology to be distributed more widely.[189,194–197] Effective use of MAM will allow for on-demand printing of products at the point of requirement, eliminating the need for large warehousing and supply chain management requirements.[194,195] This would not only reduce the cost of part purchases and replacement, but also decrease equipment downtime, hasten maintenance speed, and eventually leading to greater system readiness.[189,195,197]

However, not all parts can be printed using MAM for military operations. In addition to technical challenges with complex geometry or material, the process is still more suited for single part production than for mass production.[190,197] Therefore, there is still a need to scrutinize parts and assemblies that are appropriate for MAM. Yet, selecting which parts to produce and consolidate using MAM is challenging as there are many candidate parts, few experts, and no one expert has all the required knowledge.[188,196] This process is especially challenging in an emerging technology field such as MAM, where there is high level of uncertainty and tacit knowledge is predominant.[188,198]

While experts can be used to make these assessments in emerging technology, errors and inconsistencies in their judgment can reduce their accuracy and usefulness. Prior research showed that at the scientific frontier, experts utilize different heuristic rules to substitute complex processes for simpler ones in order to make their decisions, similar to lay-people.[199,200] This implies that inconsistencies that plague people for human-decision making also applies to experts.[201,202] In addition, experts with relevant technological knowledge of MAM do not necessarily overlap significantly with those who have relevant MAM experience in the Army context.[197] Therefore, there is a need to improve how we gather and aggregate expert judgment in these applications. Prior research has shown that we can leverage statistical decision making and combine it with expert insight to improve prediction accuracy, and we can use this combination to develop expert-guided algorithms to identify which components and systems should be produced or consolidated with MAM.[203,204] Extending the prior work done by *Funk et al.* and *Guo et al.*, we can elicit and aggregate expert judgment from different domains to understand what attributes are important for MAM and for part consolidation.[18,205] Specifically, we 1) use semi-structured interviews to filter for attributes that are specific to the Army context, 2) expand our existing techno-economic and expert decision models to incorporate Army priorities and tech authority procedure, and 3) prototype an interface, seeded with expert knowledge and our existing models, to provide an Army database that experts can access and evaluate parts most for their suitability for MAM and parts consolidation.

4.3 Methods

To understand the set of attributes that experts consider to be important when judging on part suitability for MAM and for consolidation, we used both existing literature and 1-on-1 semi-structured interviews with selected experts with military experience. While existing work in the literature have primarily focused on technical feasibility of AM (part orientation, part complexity, existence of overhangs, support structure etc.), experts with domain subject knowledge could also have logistic or mission-oriented attributes that they would consider for both part selection and part consolidation.[18,190,206–210] Our 1-on-1 semi-structured interviews were then designed to elicit non-technical attributes from experts, while at the same time confirming the importance of the set of technical attributes in the literature.

We interviewed seven AM experts, five within academia and two from the US military. Expert backgrounds ranged from material science to mechanical engineering, to aerospace industry. The interviews were semi-structured, with seven questions focused on parts consolidation and seven questions focused on parts selection (questions can be found in Appendix C.1). The questions started by asking for the experts' prior experience with the topic and depending on their responses we adjusted the other questions that attempted to understand how they made decisions for both parts consolidation and part selection. For most of the experts, they addressed both parts selection and consolidation in their answers at the same time. While their answers helped to identify the attributes for parts consolidation and parts selection, and helped to verify our assumptions, their perspective was primarily focused on academia. We generated two different sets of attributes based on the semi-structured interviews, one for part selection and one for part consolidation.

4.3.1 MAM part selection

Combining attributes identified through existing literature and the semi-structured interviews, we identified eight attributes that would be important for experts to judge whether a part is suitable for MAM (Table 4-1).

Table 4-1 - Expert judgment attributes for MAM part suitability selection. Attributes and their associated levels were identified through existing literature and 1-on-1 semi-structured interviews.

Attributes	*Description*	*Levels*
Critical Part	This attribute is critical to the safe operation of either the sub-assembly or the assembly.	Yes/No
Overhang	Additive production works by building parts up by individual layers. An overhang occurs during the printing process if the consolidated part's upper layers extend outward over the layers below. This attribute specifies whether the candidate will have an overhang under the optimal build orientation.	Yes/No
Common Material	Common alloys for processes such as laser powder beds have well-defined printing parameters for MAM. If a part can be constructed from those common alloys, it would be more suitable for MAM than parts that must be constructed from new alloys that require extensive tuning of machine parameters.	Yes/No
Mechanical Cyclic Load	The attribute specifies whether the part will experience cyclic load (mechanical loading and unloading, or regular load reversals), during normal operation.	Yes/No
Purchase Price	The estimated purchase price of this part from the OEM (**not using additive manufacturing**) to the nearest dollar.	Expected purchase price
Mission Criticality	A failure in this part would prevent the assembly from completing the mission.	Yes/No
Failure Rate	The estimated expected failure rate of the part in the field. There are three levels, low (2 or less failures in the field), and high (more than 2 failures in the field)	Low/High
Supply Shortage	This attribute estimates whether the part will experience supply shortage in the field. A supply shortage is when a part fails in the field and there is no replacement readily available.	Yes/No

It is important to note that these attributes span across technical, economic, and logistic considerations. We would then expect our expert decision model to differ from results that focus only on one of these considerations (cost minimization for example).

Using the attributes identified above, we used the multi-stage research design from *Guo et al.* and extended to this problem.[205] The key difference is that in stage one, we modeled the individual judgment of each expert using a mixed-logit model $u(m) = \widehat{\beta^T} m + individual|m$, where β^T is 8 ×1 parameter vector and m contains the 7 dummy variables for each attribute except for price, and one continuous variable for price. We included random slopes for each participant on each attribute to capture how individual experts might differ on their relative weighting of the attributes. The fixed effect estimates would represent an aggregated expert judgment on the importance of each attribute, while the random effects would represent the estimate of each individual expert's judgment on the importance of each attribute.

The recruited experts completed an online survey where they were asked to choose the most appropriate part for MAM between 30 pairs of parts (with 2 attention check questions). The survey candidates and their associated levels are shown in Appendix C.2, with an example choice task shown in Figure C-1 **(**from Appendix C.2**)**. 15 Candidate parts were selected based on the US Army's Improved Ribbon Bridge (IRB) that consists of the M16 Ramp Bay and M17 Interior Bay.[211] These parts have a wide range of attributes, from small and cheap washers, to complex and expensive subassemblies such as a direct control linear valve. The full list of the candidate parts can be found in Appendix A. Based on the 30 choices obtained from each participant, we used a mixed logit model with random slopes on the individual level to estimate the relative importance of each attribute for both each expert and the aggregated group. The 30 pairs of comparisons were first generated using a modified coordinate exchange algorithm (CEA) that can identify optimized designs.[212,213] However, since the candidate sets do not cover the full set of attribute levels, alternatives with unavailable attribute levels that were identified through the CEA were replaced with the most similar available alternatives. From pilot testing we expected the task to take approximately 30 minutes, with each two-alternative choice taking approximately 30 seconds, and an additional 5-10 minutes to read the introductory material to the task.

In stage 2, we used balanced k-means clustering to create equally sized clusters of participants who differed based on their estimated preferences in stage 1.[147] Using the model's random effects estimated in stage 1 as a vector, the balanced K-means clustering algorithm randomly divided participants into four clusters, then calculated the mean distance of each participant's estimated parameter from each cluster center. Each cluster then selected the participant that had the lowest mean distance to that cluster. Once all participants were assigned to a cluster, we identified a new center for each cluster, and recalculated the mean distance of each participant's estimated parameters to the new cluster centers. The new clusters repeat the process of selecting participants that had the lowest mean distance to that cluster. The entire process repeated until cluster centers are no longer different between each iteration. At the end of the stage, all participants were clustered into four distinct clusters.

In stage 3, participants were invited to sign up for the group discussion task and asked to provide their scheduling availability to the experimenter. The experimenter then created groups by finding times where 1 participant from each cluster was available. Participants (one from each cluster) were assigned a time to meet for the group discussion. Given the limited number of participants, groups consisted of either three or four participants. However, all participants for each group comes from a different cluster in identified in stage 2. Groups were briefed at the beginning of the experiment that their goal was come to a decision what the most suitable part for MAM. Due to the COVID-19 pandemic, participants were asked to participate online through a videoconferencing application. The researcher was present for every group discussion, however, apart from briefing the participants at the beginning of the stage, did not participate in the discussion. There are three phases in stage 3. In phase 1, participants were asked to choose between 15 pairs of potential parts for MAM as a group. They had two minutes to discuss each comparison and were asked to make a choice before the end of the two minutes. They were informed that they did not have to agree before making their vote but were encouraged to discuss and share their perspectives with each other. The 15 pairs of parts selected for group deliberation were selected by finding pairs of parts where the difference in the mean utility values of all participants in the group are maximized.

Once the group evaluated all 15 pairs of alternatives, phase two starts where we estimated the group social welfare function and present the recommendation to each group using *Guo et al.*[205] Participants made their final decision for the group after being told that they were to evaluate whether the part was the most suited for MAM given the group's preferences. Like phase 1, participants had two minutes to discuss, and they made their decision before the end of the two minutes.

The final phase was validation, where each group member independently made an additional 10 choices, but the left option was always the recommended part. In this section, they had only 30 seconds to make their choice for each pair of comparison. If the recommendation was truly the best alternative, then participants should always prefer the recommended alternative over the randomly selected alternative.

The study was reviewed and approved by Carnegie Mellon University Institutional Review Board (STUDY2021_00000156).

4.3.2 MAM part consolidation

Similar to our approach to part selection, we combined attributes identified through existing literature and the semi-structured interviews and identified six attributes that would be important for experts to judge whether different subsystems of parts are suitable for consolidation. (Table 4-2).

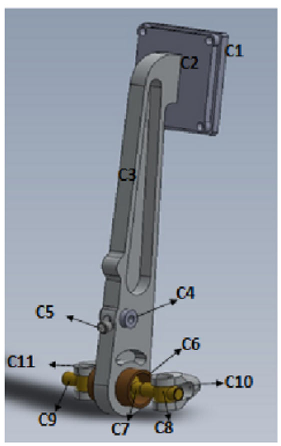
Table 4-2 - Expert judgment attributes for subsystem part consolidation. Attributes and their associated levels were identified through existing literature and 1-on-1 semi-structured interviews.

Attributes	*Description*	*Levels*
Material	The material used to print the consolidated part. In this study, we will consider three types of materials: steel, aluminum, or titanium.	Steel (Stainless Steel 316L) / Aluminum (Al6061) / Titanium (Ti6Al4V)
Critical Parts in the Consolidation	A critical part is one that is required for the assembled system to function. This attribute counts the number of critical parts included in the consolidation.	Number of critical parts
Overhangs in the Consolidation	Additive production works by building components up by individual layers. An overhang occurs during the printing process if the consolidated component's upper layers extend outward over the layers below. This attribute specifies whether the consolidation candidate will have an overhang under the optimal build orientation.	Yes/No
Support Structure Volume	The total volume required for support structures to print the part under its optimal orientation.	Support structure volume in cubic CM
Cyclic Load	The attribute specifies whether the consolidated component will experience cyclic load (mechanical loading and unloading, or regular load reversals, during normal operation) in the consolidation.	Yes/No
Total Production Cost	The expected total new production cost of the entire brake pedal assembly including the consolidation candidate. This cost is estimated with an optimization model that includes manufacturing, material, labor, and post processing costs.	Expected total production cost per brake pedal (nearest dollar)

It is important to note that the Support Structure Volume and Total Production Cost attributes are generated using a topological optimization model as described in *Nie et al.*, and therefore not elicited from experts themselves.[206]

Using the attributes identified above, we modeled expert judgment by recruiting experts to complete an online survey of 40 discrete choice tasks. Similar to the Part Selection task, we modeled the individual judgment of each expert using a mixed-logit model $u(m) = \widehat{\beta}^T m + individual|m$, with random slopes for each participant on each attribute to capture how individual experts might differ on their relative weighting of the attributes. The fixed effect estimates would represent an aggregated expert judgment on the importance of each attribute, while the random effects would represent the estimate of each individual expert's judgment on the importance of each attribute.

The recruited experts completed an online survey where they were asked to choose the most appropriate subassembly for consolidation between 40 pairs of subassemblies. The survey candidates and are shown in Appendix C.3, with an example choice task shown in Figure C-3 in Appendix C.3. Candidate subassemblies are selected from a brake pedal assembly that has 11 components as seen in Figure 4-1 below. After filtering out unfeasible consolidations, there are 288 (96 possible consolidation orientations multiplied by three different materials) possible subassembly consolidation candidates in the brake pedal assembly.



	Components
C1	pedal face
C2	pedal base spacer
C3	brake pedal arm
C4	brake pedal axle
C5	axle hold screw
C6	balance bar sleeve
C7	balance bar
C8	balance bar clevis joint Right
C9	balance bar clevis joint Left
C10	balance bar clevis Right
C11	balance bar clevis Left

Figure 4-1 - Candidate subassembly consolidations are selected from a combination of components in this brake pedal assembly.

Based on the 40 choices obtained from each participant, we used a mixed logit model with random slopes on the individual level to estimate the relative importance of each attribute for both each expert and the aggregated group. Similar to the part selection model, the 40 pairs of comparisons were first generated using a modified coordinate exchange algorithm (CEA) that can identify optimized designs. Alternatives with unavailable attribute levels that were identified through the CEA were replaced with the most similar available alternatives. From pilot testing we expected the task to take approximately 40 minutes, with each two-alternative choice taking approximately 30 seconds, and an additional 5-10 minutes to read the introductory material to the task.

4.4 Results

4.4.1 MAM part selection

The data collection processed started in June 2021 and ended in July 2021. We reached out to 77 experts through the CMU Army Research Lab (ARL) community. 19 experts responded and completed the stage one survey. These 19 individuals had a wide range of expertise and experience as seen in Table 4-3**Error! Reference source not found.**. These 19 experts had experience ranging from industry, to academia, to military, with most indicating they had a high degree of self-assessed expertise with MAM.

Table 4-3 - Experience and self-assessed expertise of experts who participated in the part selection survey. Note that some participants had multiple years of experience in more than one area (industry and academia for example), and therefore the tally of participant experience in different sectors is higher than 19. * We suspect that the participant who indicated they had no self-assessed expertise with MAM was a mistake, as they indicated they had 10+ years of industry and military experience with MAM.

Metal Additive Manufacturing Experience				
Sector	2-4 Years	5-7 Years	8-10 Years	10+ Years
Industry	3	4	4	4
Academia	8	2	0	0
Military	1	1	0	1

Self Assessed Expertise		
None	Moderate	High
1*	3	15

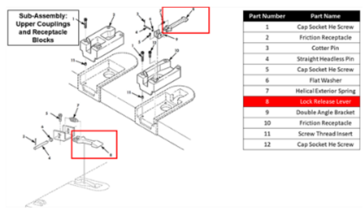
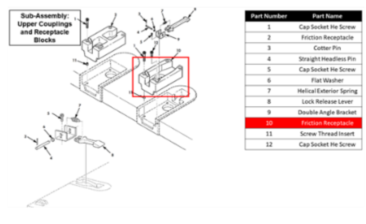
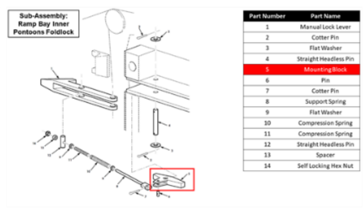
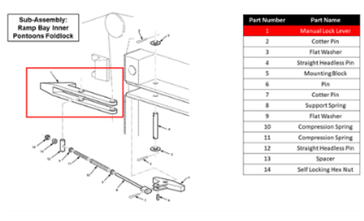
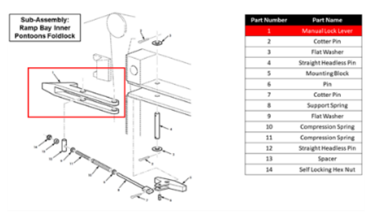
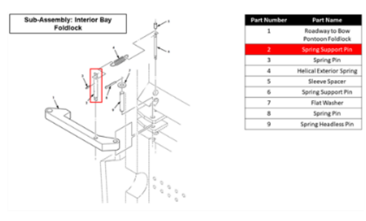
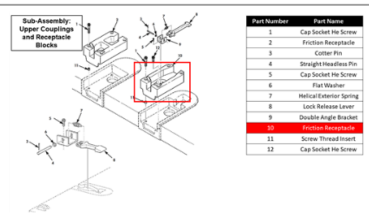
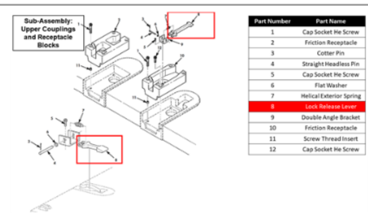
In stage one, using the mixed-logit model, the estimated relative importance of each attribute can be seen in Table 4-4. The result suggests that the three most important attributes for selecting the right part for MAM are whether overhang exists, whether the part is critical, and the price of the part if purchased directly from the OEM. In addition, our result is consistent with existing AM literature, where the presence of overhang, cyclic load, and criticality would decrease a part's suitability for MAM, while if a part can be constructed from a common allow with known printing parameters, it would increase its suitability for MAM.[210] Finally, our result suggests that experts do not only use technical attributes to judge a part's suitability for MAM, and considerations around failure rate, supply shortage, and OEM purchase price are still noticeable in an aggregated expert model.

Table 4-4 - Estimated relative importance of part selection attributes for MAM suitability. Standard errors represent fixed effects standard errors, standard deviation represents random effects standard deviation. Accuracy was calculated using in-sample AUC value. * $p < 0.1$, **$p < 0.05$, *** $p < 0.01$

Coefficients	***Estimates***	***Standard Errors***	***Standard Deviation***
Common Material	1.508	1.019	2.920
Overhang Yes	-1.634**	0.722	0.965
Cyclic Load Yes	-0.464	0.515	1.356
Critical Yes	-1.953**	0.852	2.094
Mission Critical Yes	0.504	0.773	2.069
Low Failure Rate	-0.595	0.579	1.254
Supply Shortage	0.807	0.784	2.234
Log (Purchase Price)	1.051***	0.263	0.638
Accuracy (in-sample AUC)	***0.950***		

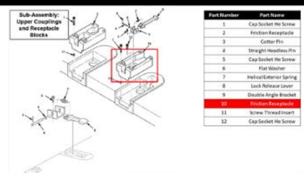
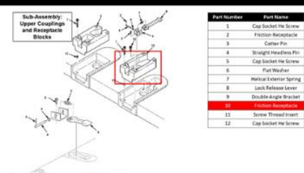
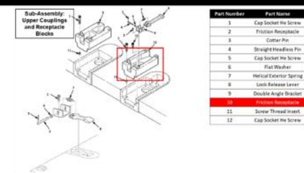
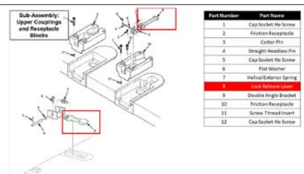
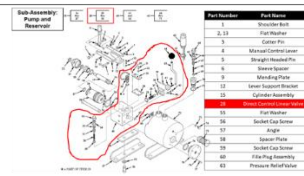
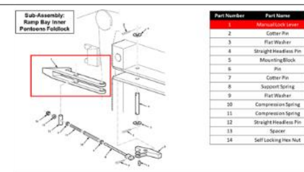
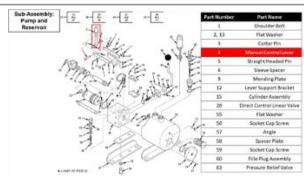
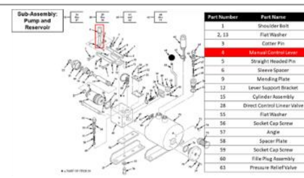
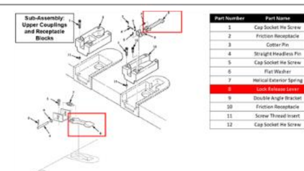
Using the estimated fixed effects attribute parameters, we identified the MAM suitability of every candidate part in our study. The full ranking of part suitability for our candidates can be found in Appendix C.4. We then compared our ranking with rankings generated through two other known logistic heuristics for MAM suitability (price, high failure rate and low supply). We can see that the top ranked candidates from the aggregated expert model are different than candidates identified through existing known heuristic rules, though candidates identified from the purchase price heuristic have the same top three candidates but in different order (Table 4-5).

Table 4-5 - Top ranked MAM suitable candidates using the aggregated expert model, purchase price heuristics, and high failure rate + low supply heuristic. Note, only two parts in the candidate list have high failure rate and low supply.

Ranking	Aggregated Expert Model	Purchase Price	High Failure + Low Supply
1			
2			
3			

In stage three, nine out of 19 experts participated in the group discussion, forming three groups of three participants each. In this group of experts, seven had industry experience, five had academia experience, and two had military experience. Out of the three groups, two groups reached consensus and agreed with the recommendation provided through the estimated social welfare function, while in the third group one participant did not agree with the recommendation (Table 4-6). When examining the validation score of the three groups, we see that in the two groups that agreed with the recommendation, participants continued to choose the recommended part over random alternatives, suggesting that the recommended part was the most suitable part for MAM for the group. While the limited number of groups (due to limited number of experts who participated in the group discussion) prevented any in-depth statistical analysis of the recommendation algorithm, the fact that two out of three groups had high satisfaction with the recommendation suggests some utility in the model approach.

Table 4-6 - Group results of expert discussion on appropriate part for MAM. While the model generated a ranked list of all candidate parts based on the group responses, only the top ranked choice is presented to the group.

Ranking	Group 1	Group 2	Group 3
1			
2			
3			
Recommendation	2/3	3/3	3/3
Validation	19/30	29/30	30/30

4.4.2 Part consolidation

The data collection process for assessing part consolidation appropriateness ran parallel with the part selection problem. Out of the 77 experts we reached out to, 12 experts responded to the part consolidation survey. These 12 experts had prior experiences from industry, academia, and military, with 9 of them indicating they had high level of self-assessed expertise (Table 4-7).

Table 4-7 - Expert demographic of participants who responded to the part consolidation survey. Note that some participants had multiple years of experience in more than one area (industry and academia for example), and therefore the tally of participant experience in different sectors is higher than 12.

Metal Additive Manufacturing Experience				
Sector	2-4 Years	5-7 Years	8-10 Years	10+ Years
Industry	1	4	0	3
Academia	5	1	0	0
Military	0	1	0	2

Self Assessed Expertise		
None	Moderate	High
0	3	9

Using the mixed-logit model, we were able to estimate the relative importance of different attributes for expert judgment of part consolidation. Due to the fact that the Critical Part attribute range from 0 – 4 and its distribution is not uniform, there are different ways to model that attribute in the mixed-logit model. Options include treating number of critical parts as factorized variable with four levels (0 to 3), factorized variable with three levels (0 and 1 as one level, 2, and 3), factorized variable with two levels (0 and 1 as low, 2 and 3 high), and continuous variable on the number of critical parts. Our models also examined how including the total number of parts being consolidated as a continuous covariate would impact the final result. All eight model results can be found in Table 4-8.

Our result shows that models with the added covariate of number of consolidated parts perform better than models without the covariate. In-sample AUC values for all four models performed better after adding the number of consolidated parts. In addition, number of consolidated parts is the most statistically significant covariate out of all estimated parameters, suggesting its importance to the model.

Out of the four different ways to model number of critical parts, we see that modeling the variable as high/low provided the best model result. While the in-sample AUC of that approach was the lowest (0.888), this is expected given there is less variables in the model. The number of significant covariates is higher than the other approaches. Therefore, we believe the best model (highlighted in blue in Table 4-8) to aggregate expert judgment on part consolidation is to treat the number of critical parts as a two-level factorized variable between high/low and including the number of parts in the consolidation.

Our results shows that the number of consolidated parts in the consolidation candidate, whether there were two or more critical parts in the consolidation, and the existence of overhangs are significant predictors of expert judgment on consolidation appropriateness. Out of these parameters, only the number of parts in the consolidation is statistically significant at $p< 0.05$ level. On the other hand, production cost of the consolidation was not a significant predictor of expert judgment on consolidation suitability.

Table 4-8 - Model estimates and performance. First four models did not include number of parts being consolidated as a covariate, while the last four models included the number of parts as a covariate. Model performance is measured using in-sample AUC. Values in brackets denote standard deviation of random effects. *p < 0.1, **p < 0.05, ***p < 0.01

	Without Number of Parts				With Number of Parts			
Coefficients	Factored Criticality	Factored Criticality, 0+1 Combined	Factored Criticality, High/Low	Continuous Criticality	Factored Criticality	Factored Criticality, 0+1 Combined	Factored Criticality, High/Low	Continuous Criticality
Overhang Yes	-0.676** (0.20)	-0.942*** (0.41)	-0.886*** (0.38)	-0.884** (0.68)	-0.369 (0.96)	-0.522 (0.90)	-0.528* (0.92)	-0.143 (0.60)
Cyclic Load Yes	-0.087 (1.16)	-0.021 (1.18)	0.031 (1.19)	-0.003 (1.12)	-0.710 (1.37)	-0.620 (1.38)	-0.473 (1.30)	-0.711 (1.50)
1 Critical Part	1.164* (0.82)	--	--	-0.045 (1.32)	0.886 (1.08)	--	--	-0.520 (1.07)
2 Critical Parts	0.692 (1.98)	-0.071 (2.31)	-0.110 (2.51)	--	-0.452 (1.61)	-1.089 (1.74)	-0.972 (1.92)	--
3 Critical Parts	0.900 (2.42)	-0.028 (2.67)	--	--	-0.778 (1.85)	-1.460 (1.84)	--	--
Production Cost	-0.004 (0.02)	-0.003 (0.02)	-0.005 (0.02)	-0.005 (0.02)	-0.003 (0.02)	-0.004 (0.02)	-0.003 (0.02)	-0.001 (0.00)
Support Volume	-0.000 (0.02)	0.009 (0.03)	0.009 (0.03)	0.006 (0.02)	-0.009 (0.02)	-0.002 (0.03)	0.000 (0.03)	-0.004 (0.02)
Number of Parts	--	--	--	--	0.438* (0.57)	0.465* (0.57)	0.363** (0.54)	0.382* (0.58)
Model AUC	**0.890**	**0.884**	**0.888**	**0.869**	**0.911**	**0.905**	**0.888**	**0.892**
Correlation w/ Cost	**-0.754**	**-0.839**	**-0.750**	**-0.898**	**-0.500**	**-0.523**	**-0.551**	**-0.209**

We also compared the results of our models with the topological optimization model. Using the estimated fixed effects parameters, we constructed the full ranking of all 288 candidates in the candidate set. We then calculated the correlation of that ranking against the ranking of candidates generated through the topological optimization model. We see that the correlation between the ranking generated through the expert aggregated model is around -0.5 when compared to the topological optimization model rankings, suggesting that the candidates deemed to be appropriate for consolidation by experts are different than the candidates that have the lowest production cost.

4.5 Discussion

We created expert aggregated models to evaluate how experts from different backgrounds evaluated part suitability for MAM and subassembly suitability for consolidation. Our results showed that for both questions, experts valued both technical and logistic attributes when making decisions. For part selections, experts considered technical attributes such as the presence of overhangs along with logistical attributes such as OEM purchase price when selecting parts for MAM. For subassembly consolidation, attributes such as the presence of overhangs and the number of parts in the consolidation were used by experts to make their decisions. In addition, our result shows that existing heuristic rules that either focus on ease of printing or economic results do not fully capture the expert decision rules when making their judgment. Instead, when aggregated, experts use a few attributes from different domains to make their decisions.

Our results also showed that there was large disagreement amongst the experts themselves on the importance of each attribute for both questions. Only three attributes across both models had a standard deviation (measure of variance of the random effects) smaller than the parameter estimate (presence of overhang and OEM purchase price for part selection model; production cost for the consolidation model). This suggests that on an individual level, experts had very different relative importance placed on different attributes. However, confirming the result from *Funk et al.*, we saw that when aggregated, we can identify a consistent decision-rule from the collection of experts.[18] This is supported through the group discussion results for part selection, where despite having different individual decision rules, experts were able to agree quickly on a consensus of what the best candidate is for MAM.

It is important to note several shortcomings of our results. The first is the relatively low number of experts that were able to complete the surveys for both questions, and the low number of experts who participated in group discussion for part selection. Out of 77 experts we contacted, only 19 responded to the part selection survey (with 9 experts participating in the group discussion), and 12 responded to the consolidation survey. This reduced our results' statistical power and limited our ability to conclude more definitely on the different attributes that drives expert decision making. In addition, given MAM is an emerging technology, certain experts might have more accurate assessment on either part MAM suitability or subassembly consolidation suitability. Ideally then, we would weigh those experts' evaluations more during the aggregation process. While there are a variety of different methods for weighting experts, most of them rely on measuring their responses against a known calibration variable.[214,215] However, given the emerging nature of the MAM field, it is difficult to validate and calibrate collected expert responses. While we attempted to use self-assessed expertise to infer expert evaluation accuracy, the small sample size of experts who responded to the surveys made it impractical to use for this purpose. Finally, since both models are built on specific assembly (pontoon bridge) and specific subassembly (brake pedal assembly), extending these models would require collecting data from experts on additional parts and assemblies. Different contexts with different business case or regulatory environment (aerospace, nuclear etc.) could also change how experts view the importance of these attributes. Therefore, it is important to take these considerations into account before extending these model results.

Finally, to help Army stakeholders navigate the implications of our models, we created interfaces for both models where single experts can input a single candidate (either part or subassembly) and receive a suitability score based on the aggregated models. The interfaces were built using R Shiny and it is hosted through the shinyapps.io server. The demos evaluate potential candidates using the fix effects parameters estimated by the aggregated expert models. User can enter attributes of a potential candidate through a series of drop-down menus and textboxes (see Figure 4-2). The suitability score would then be provided to the user based on their inputs. In addition, the interfaces will show candidate parts that have similar suitability for either MAM or consolidation based on the user defined characteristics. These interfaces can easily be extended to other parts and subassemblies by substituting different attributes or parameter values in the underlying model. The two interfaces can be found at https://nilesxug.shinyapps.io/PartSelectionDemo/ for part selection, and https://nilesxug.shinyapps.io/ConsolidationDemo/ for subassembly consolidation.

We extended prior work on aggregating expert judgment for selecting appropriate parts for MAM and consolidation for the U.S. Army context and found that there is value incorporating attributes from different expert domains.[18,205] Despite individual experts employed different heuristic rules, aggregated results found clear and consistent decision-rules that are important. This further reinforces prior findings that aggregate expert knowledge have the potential to increase the speed of technology adoption in emerging technologies.

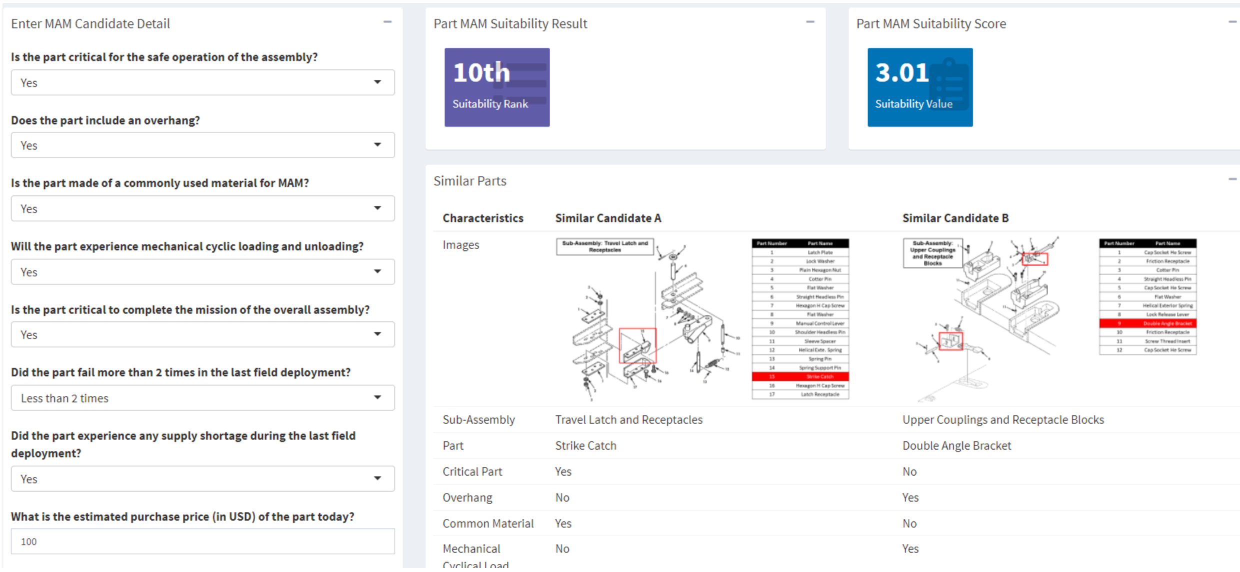


Figure 4-2 – Screenshot of the MAM part selection interface to evaluate potential candidates using the developed expert aggregate model. Users are expected to enter the candidate details using the controls on the left-hand side of the screen, while the right hand side of the screen will display 1) ranking of the candidate against known set of candidates, 2) a numerical suitability score, and 3) other parts that have similar suitability scores to compare against the candidate.

4.6 Data availability

Statistics were done using R 3.5.0 (R Core Team, 2018), the dplyr, tidyr, googleAuthR, googledrive, googlesheets4, shinyjs, lubridate, mlogit, survival, DoE.base, DoE.wrapper, Metrics, mgcv, support.CEs, mlmRev, lme4, rlist, Matrix, matrixStats, pROC, purrr, magrittr, gargle, httr, readxl, fmsb, gridExtra, ggpubr, shiny, shinydashboard, ggplot2, idefix, and Hmisc, packages.[106,107,164–168,170–174,110,175,177–185,111,212,113,114,160–163]

All data, code, and stimuli are available on GitHub: https://github.com/powerguo/MAMDemo.

Chapter 5 Conclusions and Contributions

5.1 Conclusions

Group decision-making for techno-economic assessment often involves coordination between multiple stakeholders who have their own perspectives, motivations, objectives and expertise.[1,2] This increases the complexity of the decision-making process, require qualitative evaluations around tradeoffs of multiple objectives, and can lead to unsatisfactory and suboptimal decisions.[3–7] Although empirical evidence is lacking, there is reason to believe that real-time feedback and consensus-driven group recommender systems can improve group decision-making. Therefore, my dissertation fills the gap in the existing literature and used group behavioral experiments to investigate the effectiveness of these proposed behavioral interventions on group decision making for techno-economic assessments.

I achieved these objectives by recruiting a wide range of participants (students, laypeople in the general public, experts in an emerging technology), organizing them into groups, and evaluated the feasibility and effectiveness of the interventions on helping them to improve their decision-making in three different techno-economic areas.

Chapter 2 extended a prior MILP optimization model for wastewater management and evaluated how groups of students, if given different roles and objectives, would respond to real-time feedback during a design task with two competing objectives. Utilizing a between/within subject design, I was able to use the objective results to measure the additive effectiveness of providing real-time feedback to groups against the groups working independently and groups working collaboratively, but without real-time feedback. My results suggest that real-time feedback provided a statistically significant design improvement for groups against both working independently and with only collaboration. When provided with real-time feedback, participants were able to better understand the impact of their decisions on the overall system through the display of objective metrics. We also observed that participants became more motivated as they grasped the connection between their decisions and the overall solution, which manifested in participants wanting to find better solutions. In addition, participants were able to use the real-time feedback to validate their assumptions for their decisions and recognize when those assumptions were false. Finally, real-time feedback improved the results of groups who had unmotivated members (observed through their lack of interaction with other members in their group), where the real-time feedback allowed groups to use that feedback as cues to point out where the unengaged members can improve the overall system.

In addition to advantages that real-time feedback provided to participants, *shared information bias* had a greater effect on team performance in complex design tasks than in perceptual tasks in prior work.[100,101] Without the ability to confirm their assumptions during the informal collaboration design task, participants lacked the ability to understand what information is essential, instead focusing on the information that all participants possessed. When provided with feedback, however, participants were able to confirm their assumptions and understand what information is needed from other participants to generate the optimal design. For complex design decisions, our result shows feedback is an important aspect of group success.

For Chapter 3, I pivoted to evaluate the feasibility and effectiveness of providing a consensus-driven group recommender system for groups to reach consensus for climate change policies inspired by proposed policies from the 2020 Democratic Party Presidential Primaries. Modifying the group behavioral experiment from Chapter 2, I was able to recruit people from the general population in the Pittsburgh area and examined how their preferences changed after providing them with a recommendation from the recommender system. After eliciting their individual preferences using a standard discrete choice approach, I clustered the participants based on their policy preferences, and then recruited them in groups of four to discuss their climate change policy preferences. Providing them with recommendations from two different underlying approaches, weighted-sum versus mean-variance, I measured the groups' acceptance of the recommendations, and found that groups who were given the mean-variance approach were more likely to accept the recommendation provided by the recommender system, and they were more likely to stay with the recommendation when provided with other options. The results demonstrate that consensus driven group recommender systems have the potential to help teams reach consensus on difficult policy domains by providing participants the opportunity to discuss their preferences with others in real-time, allowing for participants to change their underlying preferences, and filtering out policy options in a large policy space that would have low probability of acceptance to all members. In addition, the observational data showed that participants in the study (Democratic Party primary voters in the Pittsburgh area) had strong policy preferences for an aggressive 100% clean energy portfolio goal and for a ban on fossil fuel exploration on public lands, despite the large differences in those policies' CO2e emissions reduction effectiveness and cost to the private sector.

For Chapter 4, I extended the behavior experiment design from Chapter 3, and applied the mean-variance driven group recommender approach to the problem of selecting appropriate parts for metal additive manufacturing (MAM) from experts with domain knowledge. Using the same discrete choice platform developed in Chapter 3, after recruiting MAM experts from industry, academia, and the U.S. Army, I captured expert judgment of both selecting appropriate parts for MAM and for selecting appropriate subassemblies for part consolidation. In addition, I examined the effect of providing the consensus driven group recommender system to three groups of experts tasked with selecting appropriate parts for MAM. The results show that while individual experts may have very different judgment on appropriateness for MAM, aggregating expert judgment together using a mixed-logit approach provides a clear and consistent decision-rule. This confirms the prior work of *Funk et al.*, which shows that pooling expert judgment using statistical methods can create valuable insights of expert judgment for domains at the technology frontier even if individual experts have different judgments. In addition, the results show that due to the emerging nature of MAM, different domains will have different attributes that are important for its adoption. While some technological attributes such as the presence of overhangs do drive expert decision making for MAM, logistic attributes such as OEM purchase price and number of consolidated parts are also valued by experts. Furthermore, some attributes such as presence of cyclic loading and support structure volume that prior literature identified as important drivers for MAM adoption were not significant drivers of expert judgment. Finally, the group discussion results show real-world evidence of the feasibility and usefulness of the consensus-driven group recommender system, as it helped two out of three groups reach consensus on selecting the most appropriate part for MAM.

5.2 Research Contributions

This dissertation aimed to evaluate the feasibility and effectiveness of providing real-time performance feedback and consensus-driven group recommender systems to groups for techno-economic assessment. The following publications, including open-sourced datasets, are products of the research and methods used in this dissertation.

- Chapter 2: Real-Time Feedback Improves Multi-Stakeholder Design for Complex Environmental Systems
  - Journal: Guo, N., Davis, A., Mauter, M. & Whitacre, J. Real-time feedback improves multi-stakeholder design for complex environmental systems. *Environ. Res. Commun. 3*, 45006 (2021) doi:10.1088/2515-7620/abf466
  - Code, stimulus, and dataset: N., Davis, A., Mauter, M. & Whitacre, J. Multistakeholder Design. (2021) https://github.com/we3lab/MultiStakeholderDesign
- Chapter 3: A quantitative method for reaching consensus on federal climate change policy in the United States
  - Journal: Guo, N., Mauter, M. S. & Davis, A. L. A quantitative method for reaching consensus on federal climate change policy in the United States. (2022). *In Submission.*
  - Pre-registered experiment plan: Guo, N., Davis, A. L. Optimizing an Unknown Social Welfare Function to Aid Public Decision-Making. *Open Science Foundation*. (2020) doi: 10.17605/OSF.IO/F96BM
  - Code, stimulus, and dataset: Guo, N., Mauter, M. S. & Davis, A. L. Policy Recommendations. (2022). https://github.com/powerguo/PolicyRecommendation
- Chapter 4: Automating Subsystem Consolidation Evaluations and Part Selection for Metal Additive Manufacturing
  - Journal: Guo, N., Jung, S., Funk, P., Davis, A. L. Automating Subsystem Consolidation Evaluations and Part Selection for Metal Additive Manufacturing. (2022). *In Preparation.*
  - Code, stimulus, and dataset: Guo, N., Davis, A. L., MAM Demo. (2022). https://github.com/powerguo/MAMDemo

5.3 Future Prospective and Recommendations

5.3.1 Evaluate the impact of real-time feedback with experts

In chapter 2, I evaluated the impact of providing real-time feedback to groups of students recruited from Carnegie Mellon University. To ensure that students had sufficient information to act as experts, I provided them with briefing material that would provide both technical information and personal motivation about the role they were randomly assigned to. However, this cannot replace observations with real experts in their domain. Briefing material, no matter how well prepared, is no replacement for personal experience with the problem at hand. Even if provided with the information, a student may not identify the implications of their actions and the relationships between different modules that an expert would. This also necessitated the simplification of the system design task in order to ensure participants were able to complete the task in a reasonable amount of time, without feeling discouraged or frustrated with any lack of progress.

Recruiting real experts with domain knowledge would allow the researchers to repeat the analysis with more complex systems (either in terms of decision variables, or objectives, or system parameters), which can further quantify the benefit of real-time feedback. In addition, this would allow researchers to observe how experts interact with the information, and how they might react if the information is inconsistent with their expectations. Would experts quickly adjust their priors, or would they distrust information provided to them through a black-box system? Finally, repeating this analysis with experts will allow researchers to observe if experts have any behavioral changes with the presence of the feedback. Since experts will have real-time access to system performance, there is a question of where they will refocus attention.

5.3.2 Evaluate trust for group recommender system results

In Chapter 3 and 4, I used a consensus-driven group recommender system to identify recommendations that would are acceptable to groups of individuals. However, for the system to work as intended and generate accurate recommendations, participants need to be motivated to come to a consensus and trust that the recommender generates recommendations that adequately captures their own preferences. If participants are motivated by negative partisanship, tribalism, and distrust of others, then it is unlikely that the recommender system in its currently form can generate an accurate recommendation that can satisfy everyone.

However, many of our policy problems such as healthcare, gun control, pandemic response, and immigration are plagued motivated reasoning and low trust in both government and expertise.[216] For example, recent research showed that most variables such as GDP, pandemic preparedness, population density, and cancer/COPD prevalence did not adequately explain cross-country variations in COVID-19 infection rates.[217] Instead, the two largest predictors of COVID-19 infections other than age were trust in government and interpersonal trust.[217] Therefore, when attempting to find policy solutions in these domains that are acceptable to individuals, there is work to be done to create community engagement strategies to foster a greater sense of interpersonal trust.

While the recommender system requires trust to work as intended, the approach presented in Chapter 3 and 4 luckily does represent an opportunity to foster greater interpersonal trust when attempting to identify potential solutions. By creating an environment where participants are asked to share their perspectives and reasoning, participants have the opportunity to build trust through the process. Through informal observation of results of Chapter 3, policy options no longer seem impersonal when they are presented by someone else in a small group. Therefore, I believe there is potential for the group discussion process to increase trust and compromise between participants.

5.3.3 Evaluate the effect of group size on the quality of recommendation

In Chapters 3 and 4, I used a group size of four participants to examine the effects of the group recommender system on the probability of consensus. While that group size was necessary to have a controlled and manageable experimental setting, in many real-world applications (in policy areas such as public health, immigration etc.), the number of stakeholders could vary greatly. Since the dynamic of any group discussion changes as the number of participants increases, there is a potential for group recommender system to have different effectiveness due to a change in the group size. Therefore, there are additional opportunities to investigate the effectiveness of the consensus driven recommender system with a variety of different group sizes.

For example, rather than using the cluster size of 4 with a balanced cluster size in Chapter 3, potential investigation could vary the number of clusters to identity other categories of individual preferences and evaluate the performance of the recommender system on different cluster sizes with different Davies-Bouldin index values. This would allow people to grouped more accurately with others who share their preferences, and the identification of groups that have greater differences (for example, preferences for fast-decarbonization and increased nuclear power).

In addition, as the number of clusters increase, the number of participants in a group would increase. It is reasonable to expect the dynamic of the group discussion to change. With more participants, it would be harder for every member of the group to voice their desires and objectives during discussion, especially in an ad-hoc setting without a pre-defined moderator. This would then affect the ability of the consensus-driven model to find the recommendation that would have a high probability of success. Therefore, there is a potential for further examination on the effect of group size on the probability of consensus.

Chapter 6 References

1. Kling, C. L., Arritt, R. W., Calhoun, G. & Keiser, D. A. Integrated Assessment Models of the Food, Energy, and Water Nexus: A Review and an Outline of Research Needs. *https://doi.org/10.1146/annurev-resource-100516-033533* **9**, 143–163 (2017).
2. Ascough, J. C., Maier, H. R., Ravalico, J. K. & Strudley, M. W. Future research challenges for incorporation of uncertainty in environmental and ecological decision-making. *Ecol. Modell.* **219**, 383–399 (2008).
3. Simon, H. Theories of bounded rationality. *Decis. Organ.* **1**, 161–176 (1972).
4. Booker, J. M. & McNamara, L. A. Solving black box computation problems using expert knowledge theory and methods. *Reliab. Eng. Syst. Saf.* **85**, 331–340 (2004).
5. Kahneman, D., Slovic, S. P., Slovic, P. & Tversky, A. *Judgment under uncertainty: Heuristics and biases*. (Cambridge university press, 1982).
6. Thurston, D. L. Real and perceived limitations to decision based design. in *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference* **19739**, 91–98 (American Society of Mechanical Engineers, 1999).
7. Schuler, D. & Namioka, A. *Participatory design: Principles and practices*. (CRC Press, 1993).
8. DWARAKANATH, S. & WALLACE, K. M. Decision-making in Engineering Design: Observations from Design Experiments. *https://doi.org/10.1080/09544829508907913* **6**, 191–206 (2007).
9. Dara, S., Chowdary, C. R. & Kumar, C. A survey on group recommender systems. *J. Intell. Inf. Syst.* 1–25 (2019).
10. Castro, J., Quesada, F. J., Palomares, I. & Martínez, L. A consensus-driven group recommender system. *Int. J. Intell. Syst.* **30**, 887–906 (2015).
11. Bahrami, B. *et al.* Optimally Interacting Minds. *Science (80-.).* **329**, 1081–1085 (2010).
12. Bahrami, B. *et al.* Together, slowly but surely: The role of social interaction and feedback on the build-up of benefit in collective decision-making. *J. Exp. Psychol. Hum. Percept. Perform.* **38**, 3–8 (2012).
13. Latané, B., Williams, K. & Harkins, S. Many hands make light the work: The causes and

consequences of social loafing. *J. Pers. Soc. Psychol.* **37**, 822–832 (1979).

14. Huber, G. P. Issues in the design of group decision support sytems. *MIS Q. Manag. Inf. Syst.* **8**, 195–204 (1984).
15. Wong, Y. K. Computer-aided group judgment: A study of group decision support systems effectiveness. (1989).
16. DeSanctis, G. COMPUTER GRAPHICS AS DECISION AIDS: DIRECTIONS FOR RESEARCH*. *Decis. Sci.* **15**, 463–487 (1984).
17. Ashton, R. H. Combining the judgments of experts: How many and which ones? *Organ. Behav. Hum. Decis. Process.* **38**, 405–414 (1986).
18. Funk, P., Davis, A., Vaishnav, P., Dewitt, B. & Fuchs, E. Individual inconsistency and aggregate rationality: Overcoming inconsistencies in expert judgment at the technical frontier. *Technol. Forecast. Soc. Change* **155**, 119984 (2020).
19. Bruhin, A., Fehr, E. & Schunk, D. The many faces of human sociality: Uncovering the distribution and stability of social preferences. *J. Eur. Econ. Assoc.* **17**, 1025–1069 (2018).
20. Conte, A. & Moffatt, P. G. The econometric modelling of social preferences. *Theory Decis.* **76**, 119–145 (2014).
21. Conte, A. & Levati, M. Use of data on planned contributions and stated beliefs in the measurement of social preferences. *Theory Decis.* **76**, 201–223 (2014).
22. Bellemare, C., Kröger, S. & Van Soest, A. Measuring Inequity Aversion in a Heterogeneous Population Using Experimental Decisions and Subjective Probabilities. *Econometrica* **76**, 815–839 (2008).
23. Fisman, R., Kariv, S. & Markovits, D. Individual Preferences for Giving. *Am. Econ. Rev.* **97**, 1858–1876 (2007).
24. Bahrami, B. *et al.* What failure in collective decision-making tells us about metacognition. *Philos. Trans. R. Soc. B Biol. Sci.* **367**, 1350–1365 (2012).
25. Koriat, A. When Are Two Heads Better than One and Why? *Science (80-.).* **336**, 360–362 (2012).
26. Mellers, B. *et al.* Psychological Strategies for Winning a Geopolitical Forecasting Tournament. *Psychol. Sci.* **25**, 1106–1115 (2014).
27. Bang, D. *et al.* Does interaction matter? Testing whether a confidence heuristic can replace interaction in collective decision-making. *Conscious. Cogn.* (2014).

doi:10.1016/j.concog.2014.02.002

28. Hogarth, R. M., Gibbs, B. J., McKenzie, C. R. M. & Marquis, M. A. Learning From Feedback: Exactingness and Incentives. *J. Exp. Psychol. Learn. Mem. Cogn.* (1991). doi:10.1037/0278-7393.17.4.734
29. Gonzalez-Mulé, E., Courtright, S. H., DeGeest, D., Seong, J.-Y. & Hong, D.-S. Channeled Autonomy: The Joint Effects of Autonomy and Feedback on Team Performance Through Organizational Goal Clarity. *J. Manage.* **42**, 2018–2033 (2016).
30. Gabelica, C., Van den Bossche, P., De Maeyer, S., Segers, M. & Gijselaers, W. The effect of team feedback and guided reflexivity on team performance change. *Learn. Instr.* **34**, 86–96 (2014).
31. Luthans. *Organizational Behavior*. (McGraw-Hill/Irwin).
32. Ciucci, F., Honda, T. & Yang, M. C. An information-passing strategy for achieving Pareto optimality in the design of complex systems. *Res. Eng. Des.* **23**, 71–83 (2012).
33. Jung, J. H., Schneider, C. & Valacich, J. Enhancing the Motivational Affordance of Information Systems: The Effects of Real-Time Performance Feedback and Goal Setting in Group Collaboration Environments. *Manage. Sci.* **56**, 724–742 (2010).
34. Severo, M. C., Walentowska, W., Moors, A. & Pourtois, G. Goals matter: Amplification of the motivational significance of the feedback when goal impact is increased. *Brain Cogn.* **128**, 56–72 (2018).
35. Ederer, F. Feedback and Motivation in Dynamic Tournaments. *J. Econ. Manag. Strateg.* **19**, 733–769 (2010).
36. Davis, A. L., Krishnamurti, T., Fischhoff, B. & Bruine de Bruin, W. Setting a standard for electricity pilot studies. *Energy Policy* **62**, 401–409 (2013).
37. Sergi, B., Davis, A. & Azevedo, I. The effect of providing climate and health information on support for alternative electricity portfolios. *Environ. Res. Lett.* **13**, 24026 (2018).
38. Zhang, H., Dong, Y., Chiclana, F. & Yu, S. Consensus efficiency in group decision making: A comprehensive comparative study and its optimal design. *Eur. J. Oper. Res.* **275**, 580–598 (2019).
39. O'connor, M., Cosley, D., Konstan, J. A. & Riedl, J. PolyLens: A recommender system for groups of users. in *ECSCW 2001* 199–218 (Springer, 2001).
40. Kling, C. L., Arritt, R. W., Calhoun, G. & Keiser, D. A. Integrated Assessment Models of

the Food, Energy, and Water Nexus: A Review and an Outline of Research Needs. *Annu. Rev. Resour. Econ.* **9**, 143–163 (2017).

41. AKERLOF, G. A. THE MARKET FOR “LEMONS”: QUALITY UNCERTAINTY AND THE MARKET MECHANISM. *Uncertain. Econ.* 235–251 (1978). doi:10.1016/B978-0-12-214850-7.50022-X
42. Hayek, F. A. The use of knowledge in society. *Am. Econ. Rev.* **35**, 519–530 (1945).
43. Hazelrigg, G. A. A Framework for Decision-Based Engineering Design. *J. Mech. Des.* **120**, 653–658 (1998).
44. Keysar, B., Lin, S. & Barr, D. J. Limits on theory of mind use in adults. *Cognition* **89**, 25–41 (2003).
45. Stasser, G. & Titus, W. Pooling of unshared information in group decision making: Biased information sampling during discussion. *Small Groups Key Readings* **48**, 227–239 (2006).
46. Ludwig, T. D., Geller, E. S. & Clarke, S. W. The Additive Impact of Group and Individual Publicly Displayed Feedback: Examining Individual Response Patterns and Response Generalization in a Safe-Driving Occupational Intervention. *Behav. Modif.* **34**, 338–366 (2010).
47. Bose, U. Design and evaluation of a group support system supported process to resolve cognitive conflicts. *Comput. Human Behav.* **49**, 303–312 (2015).
48. Sigrist, R., Rauter, G., Marchal-Crespo, L., Riener, R. & Wolf, P. Sonification and haptic feedback in addition to visual feedback enhances complex motor task learning. *Exp. Brain Res.* **233**, 909–925 (2015).
49. Austin-Breneman, J., Honda, T. & Yang, M. C. A Study of Student Design Team Behaviors in Complex System Design. *J. Mech. Des.* **134**, 124504 (2012).
50. Cohen, I., Brinkman, W.-P. & Neerincx, M. A. Effects of different real-time feedback types on human performance in high-demanding work conditions. *Int. J. Hum. Comput. Stud.* **91**, 1–12 (2016).
51. Nijstad, B. A., Stroebe, W. & Lodewijkx, H. F. M. Persistence of Brainstorming Groups: How Do People Know When to Stop? *J. Exp. Soc. Psychol.* **35**, 165–185 (1999).
52. Parks, C. D. & Sanna, L. J. *Group Performance and Interaction*. (Westview Press, 1999).
53. Kingdon, J. W. & Stano, E. *Agendas, alternatives, and public policies*. **45**, (Little, Brown Boston, 1984).

54. Galesic, M. *et al.* Asking about social circles improves election predictions. *Nat. Hum. Behav.* **2**, 187–193 (2018).
55. De La Maza, C., Davis, A., Gonzalez, C. & Azevedo, I. A graph-based model to discover preference structure from choice data. in
56. Sergi, B., Azevedo, I., Xia, T., Davis, A. & Xu, J. Support for Emissions Reductions Based on Immediate and Long-term Pollution Exposure in China. *Ecol. Econ.* **158**, 26–33 (2019).
57. Arrow, K. J. Little's critique of welfare economics. *Am. Econ. Rev.* **41**, 923–934 (1951).
58. Little, I. M. D. *A critique of welfare economics*. (Oxford University Press, 2002).
59. Sen, A. Informational bases of alternative welfare approaches: aggregation and income distribution. *J. Public Econ.* **3**, 387–403 (1974).
60. Roberts, K. W. S. Interpersonal comparability and social choice theory. *Rev. Econ. Stud.* 421–439 (1980).
61. Dewitt, B., Davis, A., Fischhoff, B. & Hanmer, J. An approach to reconciling competing ethical principles in aggregating heterogeneous health preferences. *Med. Decis. Mak.* **37**, 647–656 (2017).
62. Gorla, J., Lathia, N., Robertson, S. & Wang, J. Probabilistic group recommendation via information matching. *WWW 2013 - Proc. 22nd Int. Conf. World Wide Web* 495–504 (2013). doi:10.1145/2488388.2488432
63. Iriberri, N. & Rey-Biel, P. The role of role uncertainty in modified dictator games. *Exp. Econ.* **14**, 160–180 (2011).
64. Iriberri, N. & Rey-Biel, P. Elicited beliefs and social information in modified dictator games: What do dictators believe other dictators do? *Quant. Econom.* **4**, 515–547 (2013).
65. Bardsley, N. & Moffatt, P. G. The Experimetrics of Public Goods: Inferring Motivations from Contributions. *Theory Decis.* **62**, 161–193 (2007).
66. Andreoni, J. & Miller, J. Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica* **70**, 737–753 (2002).
67. Fisman, R., Jakiela, P., Kariv, S. & Markovits, D. The distributional preferences of an elite. *Science (80-.).* **349**, aab0096 (2015).
68. Kerschbamer, R. The geometry of distributional preferences and a non-parametric identification approach: The Equality Equivalence Test. *Eur. Econ. Rev.* **76**, 85–103

(2015).

69. Kerschbamer, R., Sutter, M. & Dulleck, U. How Social Preferences Shape Incentives in (Experimental) Markets for Credence Goods. *Econ. J.* **127**, 393–416 (2017).

70. Volk, S., Thöni, C. & Ruigrok, W. Temporal stability and psychological foundations of cooperation preferences. *J. Econ. Behav. Organ.* **81**, 664–676 (2012).

71. Blanco, M., Engelmann, D. & Normann, H. T. A within-subject analysis of other-regarding preferences. *Games Econ. Behav.* **72**, 321–338 (2011).

72. Karlan, D. S. Using Experimental Economics to Measure Social Capital and Predict Financial Decisions. *Am. Econ. Rev.* **95**, 1688–1699 (2005).

73. Benz, M. & Meier, S. Do people behave in experiments as in the field?—evidence from donations. *Exp. Econ.* **11**, 268–281 (2008).

74. Fehr, E. & Leibbrandt, A. A field study on cooperativeness and impatience in the Tragedy of the Commons. *J. Public Econ.* **95**, 1144–1155 (2011).

75. Laury, S. & Taylor, L. Altruism spillovers: Are behaviors in context-free experiments predictive of altruism toward a naturally occurring public good. *J. Econ. Behav. & Organ.* **65**, 9–29 (2008).

76. Bolton, G. E., Brandts, J., Katok, E., Ockenfels, A. & Zwick, R. Testing theories of other-regarding behavior: A sequence of four laboratory studies. *Handb. Exp. Econ. Results* **1**, 488–499 (2008).

77. Davis, A. L., Miller, J. H. & Bhatia, S. Are preferences for allocating harm rational? *Decision* **5**, 287 (2018).

78. Zhang, B., Dong, Y. & Herrera-Viedma, E. Group decision making with heterogeneous preference structures: An automatic mechanism to support consensus reaching. *Gr. Decis. Negot.* **28**, 585–617 (2019).

79. Goel, A., Krishnaswamy, A. K., Sakshuwong, S. & Aitamurto, T. Knapsack voting for participatory budgeting. *ACM Trans. Econ. Comput.* **7**, 1–27 (2019).

80. Benade, G., Nath, S., Procaccia, A. D. & Shah, N. Preference elicitation for participatory budgeting. in *Thirty-First AAAI Conference on Artificial Intelligence* (2017).

81. Boutilier, C. *et al.* Optimal social choice functions: A utilitarian view. *Artif. Intell.* **227**, 190–213 (2015).

82. Anshelevich, E. & Postl, J. Randomized social choice functions under metric preferences.

J. Artif. Intell. Res. **58**, 797–827 (2017).

83. Anshelevich, E., Bhardwaj, O., Elkind, E., Postl, J. & Skowron, P. Approximating optimal social choice under metric preferences. *Artif. Intell.* **264**, 27–51 (2018).
84. Guo, N., Davis, A., Mauter, M. & Whitacre, J. Real-time feedback improves multi-stakeholder design for complex environmental systems. *Environ. Res. Commun.* **3**, 45006 (2021).
85. Ascough, J. C., Maier, H. R., Ravalico, J. K. & Strudley, M. W. Future research challenges for incorporation of uncertainty in environmental and ecological decision-making. *Ecol. Modell.* **219**, 383–399 (2008).
86. Verhagen, W. J. C., Stjepandić, J. & Wognum, N. *Concurrent Engineering in the 21st Century*. *Concurrent Engineering in the 21st Century: Foundations, Developments and Challenges* (Springer International Publishing, 2015). doi:10.1007/978-3-319-13776-6
87. Chachere, J., Levitt, R. E., John Kunz & Addo-Tenkorang, R. Concurrent Engineering (CE): A Review Literature Report. *Cent. Integr. Facil. Eng.* **II**, 1–9 (2003).
88. Rock, G. & Bil, C. *Concurrent Engineering Approaches for Sustainable Product Development in a Multi-Disciplinary Environment*. (Springer London, 2013). doi:10.1007/978-1-4471-4426-7
89. Navajas, J., Niella, T., Garbulsky, G., Bahrami, B. & Sigman, M. Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds. *Nat. Hum. Behav.* **2**, 126–132 (2018).
90. Bahrami, B. *et al.* What failure in collective decision-making tells us about metacognition. *Philos. Trans. R. Soc. B Biol. Sci.* **367**, 1350 (2012).
91. Bartholomew, T. V. & Mauter, M. S. Multiobjective Optimization Model for Minimizing Cost and Environmental Impact in Shale Gas Water and Wastewater Management. *ACS Sustain. Chem. Eng.* **4**, 3728–3735 (2016).
92. Laster, L. L. & Johnson, M. F. Non-inferiority trials: The ‘at least as good as’ criterion. *Stat. Med.* **22**, 187–200 (2003).
93. Lo, S. & Andrews, S. To transform or not to transform: using generalized linear mixed models to analyse reaction time data. *Front. Psychol.* (2015). doi:10.3389/fpsyg.2015.01171
94. Pohekar, S. D. & Ramachandran, M. Application of multi-criteria decision making to

sustainable energy planning - A review. *Renew. Sustain. Energy Rev.* **8**, 365–381 (2004).

95. Hultsch, D. F., MacDonald, S. W. S., Hunter, M. A., Maitland, S. B. & Dixon, R. A. Sampling and generalisability in developmental research: Comparison of random and convenience samples of older adults. *Int. J. Behav. Dev.* **26**, 345–359 (2002).
96. LANG, A. Involuntary Attention and Physiological Arousal Evoked by Structural Features and Emotional Content in TV Commercials. *Communic. Res.* **17**, 275–299 (1990).
97. Egeth, H. & Kahneman, D. *Attention and Effort*. *The American Journal of Psychology* **88**, (1975).
98. VaezMousavi, S. M., Barry, R. J. & Clarke, A. R. Individual differences in task-related activation and performance. *Physiol. Behav.* **98**, 326–330 (2009).
99. Boksem, M. A. S., Meijman, T. F. & Lorist, M. M. Effects of mental fatigue on attention: An ERP study. *Cogn. Brain Res.* **25**, 107–116 (2005).
100. Kerr, N. L. & Tindale, R. S. Group Performance and Decision Making. *Annu. Rev. Psychol.* **55**, 623–655 (2004).
101. Baker, D. F. Enhancing Group Decision Making: An Exercise to Reduce Shared Information Bias. *J. Manag. Educ.* **34**, 249–279 (2010).
102. Mark, G. Extreme collaboration. *Commun. ACM* **45**, 89–93 (2002).
103. Oberto, R. E. *et al.* The NASA Exploration Design Team: blueprint for a new design paradigm. in *2005 IEEE Aerospace Conference* 4398–4405 (IEEE, 2005). doi:10.1109/AERO.2005.1559745
104. Smith, J. L. Concurrent Engineering in the Jet Propulsion Laboratory Project Design Center. in (1998). doi:10.4271/981869
105. Parkin, K. L. G., Sercel, J. C., Liu, M. J. & Thunnissen, D. P. Icemaker/sup TM/: An excel-based environment for collaborative design. *IEEE Aerosp. Conf. Proc.* **8**, 3669–3679 (2003).
106. D'Agostino McGowan, L. & Bryan, J. googledrive: An Interface to Google Drive. (2020).
107. Bryan, J. googlesheets4: Access Google Sheets using the Sheets API V4. (2020).
108. Rudis, B., Bolker, B. & Schulz, J. ggalt: Extra Coordinate Systems, 'Geoms', Statistical Transformations, Scales and Fonts for 'ggplot2'. (2017).
109. Leifeld, P. {texreg}: Conversion of Statistical Model Output in {R} to {\LaTeX} and

{HTML} Tables. *J. Stat. Softw.* **55**, 1–24 (2013).

110. Wickham, H., François, R., Henry, L. & Müller, K. dplyr: A Grammar of Data Manipulation. (2021).
111. Wickham, H. & Henry, L. tidyr: Tidy Messy Data. (2020).
112. Hope, R. M. Rmisc: Rmisc: Ryan Miscellaneous. (2013).
113. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag New York, 2016).
114. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using {lme4}. *J. Stat. Softw.* **67**, 1–48 (2015).
115. Fox, J. & Weisberg, S. *An {R} Companion to Applied Regression*. (Sage, 2019).
116. Berger, S., Graham, N. & Zeileis, A. *Various Versatile Variances: An Object-Oriented Implementation of Clustered Covariances in {R}*. (2017).
117. Zeileis, A. & Hothorn, T. Diagnostic Checking in Regression Relationships. *R News* **2**, 7–10 (2002).
118. Yeganeh, A. J., McCoy, A. P. & Schenk, T. Determinants of climate change policy adoption: A meta-analysis. *Urban Clim.* **31**, 100547 (2020).
119. Bedsworth, L. W. & Hanak, E. Climate policy at the local level: Insights from California. *Glob. Environ. Chang.* **23**, 664–677 (2013).
120. Betsill, M. M. & Bulkeley, H. Cities and the multilevel governance of global climate change. *Glob. Gov. A Rev. Multilater. Int. Organ.* **12**, 141–160 (2006).
121. Turner, J. M. *The Republican reversal : conservatives and the environment from Nixon to Trump*. *The Republican reversal : conservatives and the environment from Nixon to Trump* (Harvard University Press).
122. Goulder, L. H. & Stavins, R. N. Challenges from State-Federal Interactions in US Climate Change Policy. *Am. Econ. Rev.* **101**, 253–257 (2011).
123. 2021, I. *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. (Cambridge University Press, 2021).
124. 2020 Presidential Election Results: Joe Biden Wins - The New York Times. Available at: https://www.nytimes.com/interactive/2020/11/03/us/elections/results-president.html. (Accessed: 31st August 2021)

125. Full Georgia U.S. Senate runoff results | The Washington Post. Available at: https://www.washingtonpost.com/elections/election-results/georgia-senate-runoffs-2021/. (Accessed: 31st August 2021)

126. Cochrane, E. House Passes $3.5 Trillion Budget Plan for Vast Expansion of Safety Net - The New York Times. *New York Times* (2021). Available at: https://www.nytimes.com/2021/08/24/us/politics/house-budget-social-safety-net.html. (Accessed: 31st August 2021)

127. Leber, R. Infrastructure plan: Congress advances clean electricity standard for climate - Vox. *Vox* (2021). Available at: https://www.vox.com/22579218/clean-energy-standard-electricity-infrastructure-democrats. (Accessed: 31st August 2021)

128. Van Boven, L., Ehret, P. J. & Sherman, D. K. Psychological barriers to bipartisan public support for climate policy. *Perspect. Psychol. Sci.* **13**, 492–507 (2018).

129. Howe, P. D., Mildenberger, M., Marlon, J. R. & Leiserowitz, A. Geographic variation in opinions on climate change at state and local scales in the USA. *Nat. Clim. Chang.* **5**, 596–603 (2015).

130. Fisher, D. R., Waggle, J. & Leifeld, P. Where does political polarization come from? Locating polarization within the US climate change debate. *Am. Behav. Sci.* **57**, 70–92 (2013).

131. Ehret, P. J., Van Boven, L. & Sherman, D. K. Partisan barriers to bipartisanship: Understanding climate policy polarization. *Soc. Psychol. Personal. Sci.* **9**, 308–318 (2018).

132. Hess, D. J. Sustainability transitions: A political coalition perspective. *Res. Policy* **43**, 278–283 (2014).

133. Brewer, P. R. Polarisation in the USA: climate change, party politics, and public opinion in the Obama era. *Eur. Polit. Sci.* **11**, 7–17 (2012).

134. Meyer, R. Democrats Are Shockingly Unprepared to Fight Climate Change - The Atlantic. *The Atlantic* (2017). Available at: https://www.theatlantic.com/science/archive/2017/11/there-is-no-democratic-plan-to-fight-climate-change/543981/. (Accessed: 31st August 2021)

135. Davenport, C. Joe Manchin Will Craft U.S. Climate Plan - The New York Times. *New York Times* (2021). Available at: https://www.nytimes.com/2021/09/19/climate/manchin-

climate-biden.html. (Accessed: 20th September 2021)

136. Meyer, R. Democrats May Be on the Verge of Climate Disaster. *The Atlantic* (2021). Available at: https://www.theatlantic.com/science/archive/2021/09/democrats-may-be-verge-climate-disaster/620148/.

137. Schmidt, U. & Wichardt, P. C. Inequity aversion, welfare measurement and the Gini index. *Soc. Choice Welfare* **52**, 585–588 (2019).

138. Reinhart, R. 40 Years After Three Mile Island, Americans Split on Nuclear Power. *Gallup* (2019). Available at: https://news.gallup.com/poll/248048/years-three-mile-island-americans-split-nuclear-power.aspx. (Accessed: 7th September 2021)

139. Kaufman, A. Voters Back Ban On Fracking, New Poll Finds | HuffPost. *Huffington Post* (2019). Available at: https://www.huffpost.com/entry/fracking-ban-polling_n_5d701602e4b09bbc9ef95ef9. (Accessed: 7th September 2021)

140. Is the Public Willing to Pay to Help Fix Climate Change? Is the Public Willing to Pay to Help Fix Climate Change? FINDINGS FROM A NOVEMBER 2018 SURVEY OF ADULTS AGE 18 AND OLDER. (2018).

141. Fischhoff, B. Value elicitation: Is there anything in there? *Am. Psychol.* **46**, 835–847 (1991).

142. Collinson, S. 2020 election: Highlights from CNN's climate town hall | CNN Politics. *CNN* (2019). Available at: https://www.cnn.com/2019/09/05/politics/climate-town-hall-highlights/index.html. (Accessed: 7th September 2021)

143. Tyson, A. How important is climate change to voters in the 2020 election? | Pew Research Center. *Pew Research Center* (2020). Available at: https://www.pewresearch.org/fact-tank/2020/10/06/how-important-is-climate-change-to-voters-in-the-2020-election/. (Accessed: 7th September 2021)

144. Luce, R. D. & Suppes, P. *Preference, Utility, and Subjective Probability*. (Wiley, 1965).

145. Atkinson, A. C. & Donev, A. N. The construction of exact D-optimum experimental designs with application to blocking response surface designs. *Biometrika* **76**, 515–526 (1989).

146. Street, D. J., Burgess, L. & Louviere, J. J. Quick and easy choice sets: constructing optimal and nearly optimal stated choice experiments. *Int. J. Res. Mark.* **22**, 459–470 (2005).

147. Malinen, M. I. & Fränti, P. Balanced k-means for clustering. in *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)* 32–41 (2014).
148. Goodrich, B., Gabry, J., Ali, I. & Brilleman, S. rstanarm: {Bayesian} applied regression modeling via {Stan}. (2020).
149. Brilleman, S. L., Crowther, M. J., Moreno-Betancur, M., Buros Novik, J. & Wolfe, R. Joint longitudinal and time-to-event models via {Stan}. (2018).
150. Srinivas, N., Krause, A., Kakade, S. M. & Seeger, M. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv Prepr. arXiv0912.3995* (2009).
151. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*. *Statistical Power Analysis for the Behavioral Sciences* (Academic press, 1977). doi:10.4324/9780203771587
152. Energy Innovation LLC. Energy Policy Simulator 3.3.1. (2022).
153. Energy Innovation LLC. Documentation Introduction. (2022). Available at: https://us.energypolicy.solutions/docs/introduction.html.
154. Prest, B. Supply-Side Reforms to Oil and Gas Production on Federal Lands: Modeling the Implications for Climate Emissions, Revenues, and Production Shifts. (2020).
155. Berke, C., Walter, L. & Mildenberger, M. Americans Support Federal Action to Reach a 100% Clean Energy Grid – Third Way. *Third Way* (2021). Available at: https://www.thirdway.org/memo/americans-support-federal-action-to-reach-a-100-clean-energy-grid. (Accessed: 17th September 2021)
156. Jenkins, L. M. As Interior Issues Arctic Oil Leasing Plan, Over 2 in 5 Voters Oppose Opening Public Lands to Drilling. *Morning Consult* (2020). Available at: https://morningconsult.com/2020/08/24/oil-gas-drilling-public-lands-arctic-refuge-poll/. (Accessed: 17th September 2021)
157. Gomez, H. J. The next big existential fight for Democrats? Pennsylvania. *NBC News* (2021). Available at: https://www.nbcnews.com/politics/elections/next-big-existential-fight-democrats-pennsylvania-n1276745. (Accessed: 17th September 2021)
158. Baldassarri, D. & Gelman, A. Partisans without Constraint: Political Polarization and Trends in American Public Opinion. *Am. J. Sociol.* **114**, 408–446 (2008).

159. Heltzel, G. & Laurin, K. Polarization in America: two possible futures. *Curr. Opin. Behav. Sci.* **34**, 179–184 (2020).

160. Grolemund, G. & Wickham, H. Dates and Times Made Easy with {lubridate}. *J. Stat. Softw.* **40**, 1–25 (2011).

161. Croissant, Y. mlogit: Multinomial Logit Models. (2020).

162. Therneau, T. M. A Package for Survival Analysis in R. (2020).

163. Grömping, U. {R} Package {DoE.base} for Factorial Experiments. *J. Stat. Softw.* **85**, 1–41 (2018).

164. Groemping, U. DoE.wrapper: Wrapper Package for Design of Experiments Functionality. (2020).

165. Hamner, B. & Frasco, M. Metrics: Evaluation Metrics for Machine Learning. (2018).

166. Bates, D., Maechler, M. & Bolker, B. mlmRev: Examples from Multilevel Modelling Software Review. (2019).

167. Ren, K. rlist: A Toolbox for Non-Tabular Data Manipulation. (2016).

168. Bengtsson, H. matrixStats: Functions that Apply to Rows and Columns of Matrices (and to Vectors). (2020).

169. Henry, L. & Wickham, H. purrr: Functional Programming Tools. (2020).

170. Bache, S. M. & Wickham, H. magrittr: A Forward-Pipe Operator for R. (2020).

171. Bryan, J., Citro, C. & Wickham, H. gargle: Utilities for Working with Google APIs. (2020).

172. Wickham, H. httr: Tools for Working with URLs and HTTP. (2019).

173. Wickham, H. & Bryan, J. readxl: Read Excel Files. (2019).

174. Nakazawa, M. fmsb: Functions for Medical Statistics Book with some Demographic Data. (2019).

175. Auguie, B. gridExtra: Miscellaneous Functions for ‘Grid’ Graphics. (2017).

176. Kassambara, A. ggpubr: ‘ggplot2’ Based Publication Ready Plots. (2020).

177. Chang, W., Cheng, J., Allaire, J. J., Xie, Y. & McPherson, J. shiny: Web Application Framework for R. (2020).

178. Chang, W. & Borges Ribeiro, B. shinydashboard: Create Dashboards with ‘Shiny’. (2018).

179. Harrell Jr, F. E., with contributions from Charles Dupont & many others. Hmisc: Harrell

Miscellaneous. (2020).

180. Wood, S. N. *Generalized Additive Models: An Introduction with R*. (Chapman and Hall/CRC, 2017).

181. Wood, S. N., N., Pya & S"afken, B. Smoothing parameter and model selection for general smooth models (with discussion). *J. Am. Stat. Assoc.* **111**, 1548–1575 (2016).

182. Wood, S. N. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. R. Stat. Soc.* **73**, 3–36 (2011).

183. Wood, S. N. Thin-plate regression splines. *J. R. Stat. Soc.* **65**, 95–114 (2003).

184. Wood, S. N. Stable and efficient multiple smoothing parameter estimation for generalized additive models. *J. Am. Stat. Assoc.* **99**, 673–686 (2004).

185. Edmondson, M. googleAuthR: Authenticate and Create Google APIs. (2019).

186. Attali, D. shinyjs: Easily Improve the User Experience of Your Shiny Apps in Seconds. (2020).

187. Stan Development Team. {RStan}: the {R} interface to {Stan}. (2020).

188. Russell, J. D. & Fielding, J. C. *America Makes: The National Additive Manufacturing Innovation Institute (NAMII) Status Report and Future Opportunities*. (2014).

189. Schrand, A. M. Additive Manufacturing From Form to Function. *Strateg. Stud. Q. Emerg. Technol. Spec. Ed.* **10**, 74–90 (2016).

190. Moore, T. A., McConnell, B. M. & Wilson, J. R. Simulation-based evaluation on integrating additive manufacturing capability in a deployed military environment. *Proc. - Winter Simul. Conf.* **2018-Decem**, 3721–3729 (2019).

191. Meisel, N. A., Williams, C. B., Ellis, K. P. & Taylor, D. Decision support for additive manufacturing deployment in remote or austere environments. *J. Manuf. Technol. Manag.* **27**, 898–914 (2016).

192. Yang, S., Min, W., Ghibaudo, J. & Zhao, Y. F. Understanding the sustainability potential of part consolidation design supported by additive manufacturing. *J. Clean. Prod.* **232**, 722–738 (2019).

193. Schmelzle, J. *et al.* (Re)Designing for Part Consolidation: Understanding the Challenges of Metal Additive Manufacturing. *J. Mech. Des.* **137**, (2015).

194. Drushal, J. R. *Additive Manufacturing: Implications to the Army Organic Industrial Base in 2030*. (2013).

195. Zimmerman, B. A. & Allen III, E. E. *Analysis of the Potential Impact of Additive Manufacturing on Army Logistics*. (2013).

196. Huang, S. H., Liu, P., Mokasdar, A. & Hou, L. Additive manufacturing and its societal impact: A literature review. *Int. J. Adv. Manuf. Technol.* **67**, 1191–1203 (2013).

197. Boer, J. den, Lambrechts, W. & Krikke, H. Additive manufacturing in military and humanitarian missions: Advantages and challenges in the spare parts supply chain. *J. Clean. Prod.* **257**, 120301 (2020).

198. Bohn, R. E. From Art to Science in Manufacturing: The Evolution of Technological Knowledge. *Found. Trends® Technol. Inf. Oper. Manag.* **1**, 1–82 (2005).

199. Dunbar, K. How scientists think: On-line creativity and conceptual change in science. in *Creative thought: An investigation of conceptual structures and processes.* 461–493 (American Psychological Association, 1997). doi:10.1037/10227-017

200. McComb, C., Cagan, J. & Kotovsky, K. Optimizing design teams based on problem properties: Computational team simulations and an applied empirical test. *J. Mech. Des. Trans. ASME* **139**, (2017).

201. Moore, D. A. & Healy, P. J. The trouble with overconfidence. *Psychological Review* **115**, 502–517 (2008).

202. Tversky, A. Intransitivity of preferences. *Psychol. Rev.* **76**, 31–48 (1969).

203. Abdollahi, S., Davis, A., Miller, J. H. & Feinberg, A. W. Expert-guided optimization for 3D printing of soft and liquid materials. *PLoS One* **13**, e0194890 (2018).

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

205. Guo, N., Mauter, M. S. & Davis, A. L. A quantitative method for reaching consensus on federal climate change policy in the United States. (2022). *In Review.*

206. Nie, Z., Jung, S., Kara, L. B. & Whitefoot, K. S. Optimization of part consolidation for minimum production costs and time using additive manufacturing. *J. Mech. Des. Trans. ASME* **142**, (2020).

207. Yang, S., Santoro, F. & Zhao, Y. F. Towards a Numerical Approach of Finding Candidates for Additive Manufacturing-Enabled Part Consolidation. *J. Mech. Des. Trans. ASME* **140**, 1–13 (2018).

208. Yang, S., Santoro, F., Sulthan, M. A. & Zhao, Y. F. A numerical-based part consolidation candidate detection approach with modularization considerations. *Res. Eng. Des.* **30**, 63–83 (2019).
209. SAE International. 2015 Formula SAE Rules. 178 (2014).
210. Booth, J. W. *et al.* The design for additive manufacturing worksheet. *J. Mech. Des. Trans. ASME* **139**, 1–9 (2017).
211. Department of Army. *Field Maintenance Manual Including Repair Parts and Special Tools List for Improved Ribbon Bridge (IRB)*. (2013).
212. Traets, F., Sanchez, D. G. & Vandebroek, M. Generating Optimal Designs for Discrete Choice Experiments in R: The idefix Package. *J. Stat. Softw.* **96**, 1–41 (2020).
213. Meyer, R. K. & Nachtsheim, C. J. The coordinate-exchange algorithm for constructing exact optimal experimental designs. *Technometrics* **37**, 60–69 (1995).
214. Colson, A. R. & Cooke, R. M. Cross validation for the classical model of structured expert judgment. *Reliab. Eng. Syst. Saf.* **163**, 109–120 (2017).
215. Cooke, R. M. Uncertainty in risk assessment: A probabilist's manifesto. *Reliab. Eng. Syst. Saf.* **23**, 277–283 (1988).
216. Moore-Berg, S. L., Ankori-Karlinsky, L.-O., Hameiri, B. & Bruneau, E. Exaggerated meta-perceptions predict intergroup hostility between American political partisans. *Proc. Natl. Acad. Sci.* (2020).
217. Bollyky, T. J. *et al.* Pandemic preparedness and COVID-19: an exploratory analysis of infection and fatality rates, and contextual factors associated with preparedness in 177 countries, from Jan 1, 2020, to Sept 30, 2021. *Lancet* (2022). doi:10.1016/S0140-6736(22)00172-6
218. Halvorsen, R. & Palmquist, R. The Interpretation of Dummy Variables in Semilogarithmic Equations. *Am. Econ. Rev.* **70**, 474–475 (1980).

Appendix A Supplemental Information for Chapter 2

A.1 Case Study Setup

The case study system was modified from Bartholomew and Mauter, and the problem is transformed from a MILP into a non-linear integer programming problem (NIP).[1] The problem can be formulated as the following:

At the highest level, the two objective functions are:

$$\mathrm{Cost}^{\mathrm{Financial}} = \left(\mathrm{Cost}^{\mathrm{WaterWithdrawl}} + \mathrm{Cost}^{\mathrm{FreshwaterTransportation}} + \mathrm{Cost}^{\mathrm{WastewaterTransportation}} + \mathrm{Cost}^{\mathrm{Storage}} + \mathrm{Cost}^{\mathrm{WastewaterTreatment}}\right) \quad (1)$$

And:

$$\mathrm{Cost}^{\mathrm{HHE}} = \sum_{e} \mathrm{CE}_e \left(\mathrm{M}_e^{\mathrm{Storage}} + \mathrm{M}_e^{\mathrm{Pipeline}} + \mathrm{M}_e^{\mathrm{Trucking}} + \mathrm{M}_e^{\mathrm{CentralTreatment}} + \mathrm{M}_D^{\mathrm{Disposal}}\right) \quad (2)$$

Where:

$$\mathrm{Cost}^{\mathrm{WaterWithdrawl}} = \sum_{c} \sum_{j} \sum_{k} \sum_{t} \mathrm{WC}_c{*}\mathrm{FWSource}_c{*}\mathrm{FWReq}_{jkt}{*}\mathrm{Reuse}_j{*}\mathrm{Storage}_k \quad (3)$$

$\mathrm{FWSource}_c$ is the decision variable for the freshwater source, WC_c represents the freshwater withdraw cost associated with freshwater source c, Reuse_j is the binary decision variable on if wastewater is reused for future operations, $\mathrm{Storage}_k$ is the binary decision variable on if wastewater is stored for future operations, and FWReq_{jkt} is the freshwater requirement associated with the appropriate reuse/storage scenario at week t.

$$\text{Cost}^{\text{FreshwaterTransportation}}$$

$$= \sum_{c} \sum_{l} \sum_{u} \text{FWImPipe}_{l} * \text{FWImPipeType}_{u} * \text{FWImDist}_{c} * \text{FWSource}_{c} * \text{PipeCapCost}_{u}$$

$$+ \sum_{d} \sum_{l} \sum_{u} \text{FWWellPipe}_{l} * \text{FWWellPipeType}_{u} * \text{FWWellDist}_{d} * \text{PipeCapCost}_{u}$$

$$+ \sum_{t} \sum_{c} \sum_{l} \sum_{u} \text{FWImPipe}_{l} * \text{FWImPipeType}_{u} * \text{FWImDist}_{c} * \text{FWSource}_{c} * \text{PipePumpCost}_{u} * \text{FWPum}$$

$$+ \sum_{t} \sum_{d} \sum_{l} \sum_{u} \text{FWWellPipe}_{l} * \text{FWWellPipeType}_{u} * \text{FWWellDist}_{d} * \text{PipePumpCost}_{u} * \text{FWPumpFlow}$$

$$+ \sum_{t} \sum_{c} \text{FWTruckImpound}_{t} * \text{FWImDist}_{c} * \text{FWSource}_{c} * \text{FWTruckCost}$$

$$+ \sum_{t} \sum_{d} \text{FWTruckWellpad}_{t} * \text{FWWellpadDist}_{d} * \text{FWTruckCost} \ (4)$$

FWImPipe_l represents the binary decision variable on whether a pipeline is to constructed to ship freshwater from the source to the impoundment area, FWImPipeType_u represents the decision variable on the type of pipe to be used for transporting water from source to impoundment, FWImDist_c represents the distance from freshwater source c to the impoundment, and PipeCapCost_u represents the capital cost associated with pipeline type u, FWWellPipe_l is the binary decision variable on whether a pipeline is constructed to transport freshwater from the impoundment area to each well-pad, FWWellPipeType_u represents the decision variable on the type of pipe to be used for transporting water from impoundment to well-pad, FWWellDist_d represents the distance from the impoundment area to each well-pad d, PipePumpCost_u represents the pumping cost associated with pipeline u, and FWPumpFlow_t represents the amount of freshwater pumped at week t. FWTruckImpound_t represents the amount of freshwater transported from the freshwater source to the impoundment at week t, and FWTruckCost represent the unit cost of transporting freshwater.

$$
\begin{aligned}
&\mathrm{Cost}^{\mathrm{WastewaterTransportation}} \\
&= \sum_{\mathrm{CT}} \mathrm{Y_{CT}}{*}\sum_{\mathrm{l}} \mathrm{WWPipe_l}{*}\sum_{\mathrm{u}} \mathrm{WWPipeType_u}{*}\mathrm{PipeCapCost_u}{*}\mathrm{CTDist} \\
&+ \sum_{\mathrm{CT}} \mathrm{Y_{CT}}{*}\sum_{\mathrm{l}} \mathrm{WWPipe_l}{*}\sum_{\mathrm{u}} \mathrm{WWPipeType_u}{*}\mathrm{PipePumpCost_u}{*}\mathrm{WWPumped}{*}\mathrm{CTDist} \\
&+ \sum_{\mathrm{CT}} \mathrm{Y_{CT}}{*}\mathrm{WWTrucked}{*}\mathrm{WWTruckCost}{*}\mathrm{CTDist} \\
&+ \sum_{\mathrm{CT}} \mathrm{Y_{CT}}{*}\rho^{\mathrm{eff}}{*}\mathrm{WWProduced}{*}\mathrm{WWTruckCost}{*}(\mathrm{DPDist\text{-}CTDist}) \\
&+ \sum_{\mathrm{NCT}} \mathrm{Y_{NCT}}{*}\mathrm{WWProduced}{*}\mathrm{WWTruckCost}{*}\mathrm{DPDist}\ (5)
\end{aligned}
$$

$\mathrm{Y_{CT}}$ is the binary decision variable on if central treatment is used to treat wastewater. $\mathrm{WWPipe_l}$ is the decision variable for if a pipeline is used to transport the wastewater from the well-pads to the central treatment facility. $\mathrm{WWPipeType_u}$ is the decision variable for the type of pipeline. CTDist is the distance from the well-pad to the central treatment facility. WWPumped is the amount of wastewater pumped to the central treatment facility. WWTrucked is the amount of wastewater trucked to the central treatment facility. ρ^{eff} is the treatment efficiency of the central wastewater treatment facility. DPDist is the distance from the well-pad to the disposal site. WWProduced is the total amount of wastewater that is produced. $\mathrm{Y_{NCT}}$ represents the inverse of the binary decision variable on if central treatment is used to treat wastewater, such that $\mathrm{Y_{NCT}} + \mathrm{Y_{CT}} = 1$.

$$
\mathrm{Cost}^{\mathrm{Storage}} = \sum_{\mathrm{j}} \sum_{\mathrm{k}} \mathrm{Storage_k}{*}\mathrm{Reuse_j}{*}\mathrm{FWReq_{jk}}{*}\mathrm{StorageCost}(6)
$$

StorageCost is the unit cost of storing freshwater on site.

$$\text{Cost}^{\text{WastewaterTreatment}} = \sum_{CT} Y_{CT} * \text{WWProduced} * \text{TreatmentCost} \quad (7)$$

$M_e^{Storage}, M_e^{Pipeline}, M_e^{Trucking}, M_e^{CentralTreatment}, M_e^{Disposal}$ represent the amount of criteria air pollutant e produced for each activity, CE_e represent the associated unit cost of each criteria air pollutant.

The full set of criteria air pollutant considered for this case study is adopted from Bartholomew and Mauter and can be found in Table A-1. The associated emissions coefficients for each activity and cost coefficients are also used directly from Bartholomew and Mauter.[1] They used a hybrid LCA approach to calculate the coefficients, using both the Carnegie Mellon University EIO-LCA tool and process estimation through electricity production from NEI and eGRID2012. The marginal cost coefficients were found through the AP2 model.

Table A-1 - Pollutant and associated marginal air emission damages estimated from the AP2 model

Pollutant	Cost ($2015/metric ton)
NH_3	131,000
NO_x	5,540
$PM_{2.5}$	118,000
SO_2	44,500
VOC_s	11,300
CO_2	40

The model also consists of a series of constraints, ranging from mass balance constraints, to flow capacity constraints, to scheduling constraints. While these constraints reduce the total number of possible design space to ~2,200, this is still infeasible for participants to iterate through the entire solution space under a time constraint.

To design the system, participants make a series of decisions to meet their individual objectives. These decisions include the freshwater source, the transportation options for freshwater and wastewater, treatment options, and storage options.

A.2 Participant Briefing Material

The proposed users of the tool, industry experts, are vastly different than the sample population of CMU students. Experts differ from layperson in two important aspects: deep knowledge, and how the information is structured in their minds.[2] This includes how they think, their problem-solving methods, forms of perception, and modes of communication.[2,3] However, there are ways to mitigate this difference between experts with the sample population. Using existing literature that examined scientific communication between experts and layperson, we can reverse-engineer interventions that will reduce the gap between the two groups.

First, we address the knowledge gap by providing all the relevant information to the participants. This included the overall case study setup, general overview of unconventional gas exploration, specific domain knowledge such as pipeline capacity and transportation distance, and historical system performance that will provide a useful marker for the participants. In addition, we address the information structure gap by immersing the participants into the perspective of the expert, since the "process of becoming an expert entails immersion into a certain perspective".[2] This includes communicating the motivation behind each expert, the implicit rules in their domain (for example, the relationship between air emission and different economic activity), the mental model of how the system behaves, and assumptions about relationships they have with other experts. To simulate expert interactions more realistically, we added *privileged information* to each participant's briefing material to ensure each participant possessed information unique to themselves.[4] When combined, the briefing material is designed to reduce the gap between the overall population and our sample population.

The advantage of using student population also allowed us to evaluate the effect of CADS in a robust manner, since we can recruit a larger sample than if we limited ourselves to expert participants. We believe this trade-off between robustness in our results and use of students instead of experts is justified, given our mitigation efforts through the design of the briefing material.

The briefing material for each of the experts can be found below.

A.2.1 Role: well-pad operator

Individual Goal: Minimize the well-pad operational cost, which encompasses freshwater impoundment cost, wastewater storage cost, and frack fluid storage cost.

Team Goal: Minimize the overall financial cost to the entire system, including freshwater and wastewater transportation cost, wastewater treatment cost, disposal cost, and freshwater withdrawal cost.

You are the well-pad operator of your company, and your main goal is to minimize the financial cost of the system to ensure your company's overall profitability. Your individual goal is to minimize the well-pad operational cost (in the form of storage cost, freshwater impoundment cost, and frack fluid storage cost). Since the fracking schedule has already been set, you do not need to consider the production side of the operation. From past experience, you know that a well-designed system of a similar scale should cost no more than $50 million for the entire project lifetime, but you want to ensure you minimize the financial cost as much as you can.

Decisions:

You are responsible for two decisions for the overall system:

1. Reuse wastewater for fracking fluid: You can choose to reuse the wastewater generated from previous week's operation. Around 10-15% of the cumulative fracking fluid is returned to the surface as wastewater each week. Reusing this wastewater will reduce the amount of freshwater you need to sustain the operation. In addition, reusing the wastewater will also reduce the amount of wastewater that will need to be treated/disposed. There is a cost to store the fracking fluid, in the form of impoundment cost. However, only 30% of the fracking fluid can be composed of the wastewater. The rest will need to be either treated or disposed in offsite disposal well.

2. Store wastewater onsite: You can choose to store the wastewater for future uses. You can store up to 8,000 barrels a week. If you choose both to reuse the wastewater and to store the wastewater, you can use the stored wastewater for future fracking uses.

Relationship with other experts:

You rely on both the freshwater and wastewater experts to select the most appropriate transportation technique to move the freshwater onsite and the wastewater offsite. In addition, the wastewater expert will decide whether wastewater will be treated before being disposed in an offsite location or disposed directly.

You rely on the environmental regulator to select the freshwater source the firm can withdraw water from. There are different costs associated with different available freshwater sources, and that information is held by both the freshwater expert and the environmental regulator.

Key Parameters:

Parameter	Value
Fracking Fluid Demand	~240,000 barrels/week
Freshwater Demand	Varies, depending on expert decision. Ranges around ~80,000 – 240,000 barrels/week
Maximum Wastewater Fracking Fluid Composition	30%
Wastewater Storage Cost	$0.525/barrel-week
Freshwater Impoundment Cost	$800,000 capital cost + $1.20/barrel

A.2.2 Role: freshwater expert

Individual Goal: Minimize the financial cost to transport freshwater from the freshwater source to the well-pads.

You are the freshwater expert of your company, and your main goal is to minimize the financial cost to transport freshwater from the freshwater source to the well-pads. This includes freshwater withdraw cost, pipe construction cost, pipe variable cost, and trucking cost.

There are three freshwater sources that can be used to withdraw freshwater, but the environmental regulator will make the decision on which source will be used. Freshwater will need to first be transported from the source to the freshwater impoundment, where they then can be sent to each individual well-pad as set by the fracking schedule. The distance from the freshwater source to the impoundment site is dependent on the selected source, and the distance from the impoundment site to each well-pad ranges from approximately 0.5 to 6 miles, with the majority between 1 – 2 miles. Since the fracking production schedule has already been set, you do not need to decide on which well-pad the freshwater needs to be transported to, you only need to decide on the mode of transportation. You can choose to transport the freshwater by either constructing new pipelines or use a third-party trucking contractor.

Decisions:

You are responsible for three decisions for the overall system:

Construct pipeline to transport freshwater from source to impoundment: You can choose to transport the freshwater from source to impoundment either by constructing a pipeline or a third-party trucking contractor. If you choose to construct the pipeline, there is a large capital cost associated with that option.

If you do not choose to construct a pipeline to transport freshwater from source to impoundment, you can hire a third-party contractor where you will only need to pay on a per-volume-distance basis. The key financial parameters of each case will be found in the parameter table below.

Use pipeline to transport freshwater from impoundment to well-pad: You can choose to transport the freshwater from impoundment to well-pads either by constructing a pipeline or a third-party trucking contractor. If you choose to construct the pipeline, there is a large capital cost associated with that option.

If you do not choose to construct a pipeline to transport freshwater from impoundment to well-pads, you can hire a third-party contractor where you will only need to pay on a per-volume-distance basis.

Pipeline options: If you choose to construct a pipeline for either options listed above, there are three pipeline options that you can choose from. They range in different maximum flow capacity, capital cost, but have the same variable cost. If you choose a pipe option that has does not meet the freshwater demand, the excess freshwater will be transported via trucks. The financial parameters of each case will be found in the parameter table below.

Relationship with other experts:

You will act as a primary advisor to the well-pad operator, and the well-pad operator will provide you with an estimate of how much freshwater you will need to transport to the well-pads. The environmental regulator will set which freshwater source will be used, and you will need to take that information into consideration while you make your decision. You know there is a trade-off in air emissions damages between trucking and constructing a pipeline, but you do not know the exact values for that trade-off. In addition, since air emissions damages are calculated through an economic input-output method, the higher the financial cost of the various components, the higher the air emissions cost will be. However, you can cycle through your options to check which option will have higher air emissions damages.

Key Parameters:

Parameter	Value
Trucking Cost	$0.053/barrel-mile
Pipeline Variable Cost	$0.005/barrel-mile
Pipeline A Capacity	200,000 barrels/week
Pipeline B Capacity	1,000,000 barrels/week
Pipeline C Capacity	1,200,000 barrels/week
Pipeline A Capital Cost	$240,000/mile
Pipeline B Capital Cost	$320,000/mile
Pipeline C Capital Cost	$400,000/mile
Distance from Stream	10 miles
Distance from Lake	15 miles
Distance from River	35 miles
Freshwater Withdraw Cost – Stream	$0.50/barrel
Freshwater Withdraw Cost – Lake	$0.30/barrel
Freshwater Withdraw Cost – River	$0.10/barrel

A.2.3 Role: wastewater expert

Individual Goal: Minimize the financial cost of treating, disposing, and transporting wastewater from the well-pads

You are the wastewater expert of your company, and your goal is to minimize the financial cost related to wastewater. This includes transportation cost (either in the form of pipeline construction or trucking cost), central treatment of wastewater, and disposal fee paid to the disposal wells.

The amount of wastewater generated is dependent upon the decisions of the well-pad operator, and that person will have an estimate of amount of wastewater that will be transported. You can choose to transport the wastewater to a centralized treatment facility through either constructing new pipelines or use a third-party trucking contractor. However, wastewater can only be transported to disposal wells through the third-party trucking contractor.

Decisions:

You are responsible for three decisions for the overall system:

Use central treatment facility to treat wastewater: You can choose to treat the wastewater generated from the well-pads using a central treatment facility, either by constructing a pipeline or use a third-party trucking contractor. 85% of all wastewater treated in the central treatment facility will reach a quality that can be directly returned to the water sources. This 85% will not incur any additional cost. However, central treatment facility will charge a fee per wastewater barrel treated. The remaining 15% will need to be disposed in the disposal well. If you do not choose to treat the wastewater centrally, 100% of the wastewater will need to be transported to the disposal wells.

Construct pipeline to transport wastewater from well-pad to centralized treatment facility: If you choose to treat the wastewater centrally, you can choose to transport the wastewater from well-pads to centralized treatment facility either by constructing a pipeline or use a third-party trucking contractor. If you choose to construct the pipeline, there is a large capital cost associated with that option. If you choose to hire a third-party contractor, you will only need to pay on a per-volume-distance basis. The key financial parameters of each case will be found in the table below.

Pipeline options: If you choose to construct a pipeline, there are three pipeline options that you can choose from. They range in different maximum flow capacity, capital cost, but have the same variable cost. The financial parameters of each case will be found in the parameter table below. If you choose a pipeline option that has a too low of capacity, the remaining water will be transported via trucking.

Relationship with other experts:

You will act as a primary advisor to the well-pad operator, and the well-pad operator will provide you with an estimate of how much wastewater you will need to transport from the well-pads. You know there is a trade-off in air emissions damages between trucking and constructing a pipeline, with the fact that trucking usually generates higher air emissions damages than constructing a pipeline. Furthermore, there is also air emissions damages associated with central treatment, but you do not know the damages associated with each option. However, you do know that due to the economic input-output model used, air emissions damage generally increases with increased financial cost.

Key Parameters:

Parameter	**Value**
Estimated Generated Wastewater	Less than 80,000 barrels/week
Central Water Treatment Cost	\$5.50/barrel
Disposal Well Cost	\$1.50/barrel
Distance to Treatment Facility	50 miles
Distance to Disposal Well	150 miles
Trucking Cost	\$0.053/barrel-mile
Pipeline Variable Cost	\$0.005/barrel-mile
Pipeline A Capacity	200,000 barrels/week
Pipeline B Capacity	1,000,000 barrels/week
Pipeline C Capacity	1,200,000 barrels/wee
Pipeline A Capital Cost	\$240,000/mile
Pipeline B Capital Cost	\$320,000/mile
Pipeline C Capital Cost	\$400,000/mile

A.2.4 Role: environmental regulator

Individual Goal/System Goal: Ensure that the wastewater management system produces the least amount of human health damages through air emissions.

You are local community environmental regulator, and your main goal is to minimize the human health damages as a result of the air emissions generated from the overall water management system. Each activity has their associated damages that can be estimated on a dollar basis. Based on the decisions of other actors, the tool will provide you with the overall damages of the operation. However, from prior experience, you know that a well-designed system should not have more than $15 million in human health damages to the community.

The human health damages are calculated through an economic input-output model, that generally means that the greater the financial cost of the system, the higher the emissions damages. In addition, trucking emissions damages are calculated based on distance traveled, with increased damages on a per mile basis.

Out of the three freshwater sources under consideration, the community has provided a preference on Stream to be the freshwater source. You will then want to ensure the firm uses the Stream. However, the Stream has the highest freshwater withdraw cost, and it might lead to higher financial costs. If other participants are willing to compromise on other design decisions that would lower environmental cost, you would be open for them to use the Lake and River options.

Decisions:

You are responsible for one decision for the overall system:

Freshwater source for the operation: You will decide which freshwater source will be used for the firm's operation. The three water sources are: a stream, which is 10 miles away from the fracking operation; a lake, which is 15 miles away; and a river, which is 35 miles away.

Relationship with other experts:

Since your main goal is to minimize air emissions damages, you will want to push for the other experts to iterate through their design decisions to find the design with the lowest damages. However, you know that the operation has already received approval, so preventing the fracking operation entirely is impossible. You can only minimize the damages associated with the operation.

Key Parameters

Pollutant	**Cost per ton**
NH3	$131,000
NOx	$5,540
PM2.5	$118,000
SO2	$44,500
VOC	$11,300
CO2e	$40

A.3 Residual Analysis

To conduct statistical hypothesis testing generate reliable confidence intervals, we need to confirm the residuals are normally distributed.[5] In addition, OLS regression requires that the residuals are approximately mean zero and they have equal variance (homoskedasticity).[6]

The residuals for both the log-transformed financial cost model and the log-transformed environmental cost model are approximately mean zero and normally distributed, however they exhibit heteroskedastic behaviour. As seen in Figure A-1 and Figure A-2, after performing the log-transformation on the dependent variable (**B**), the studentized residuals fits the QQ-plot significantly better after the log transformation. In addition, after the log transformation, the jackknife residuals are more closely clustered around zero.

We also verified this result by applying a gamma log link function to the model and found the result to be extremely like the log transformed model results. The regression table for the gamma log link function can be found in Table A-4.

However, there is a clear fan shape in the jackknife residuals, where the absolute values of the residuals increased as the fitted values increased. This shows that the error term does not have a constant variance. The Levene's Test for equal variance[7] shows that the null hypothesis of equal variance is rejected at $p < 0.12$ for financial cost and at $p < 0.04$ for environmental cost. Therefore, as a matter of prudence, heteroskedastic robust standard errors were reported for both models.

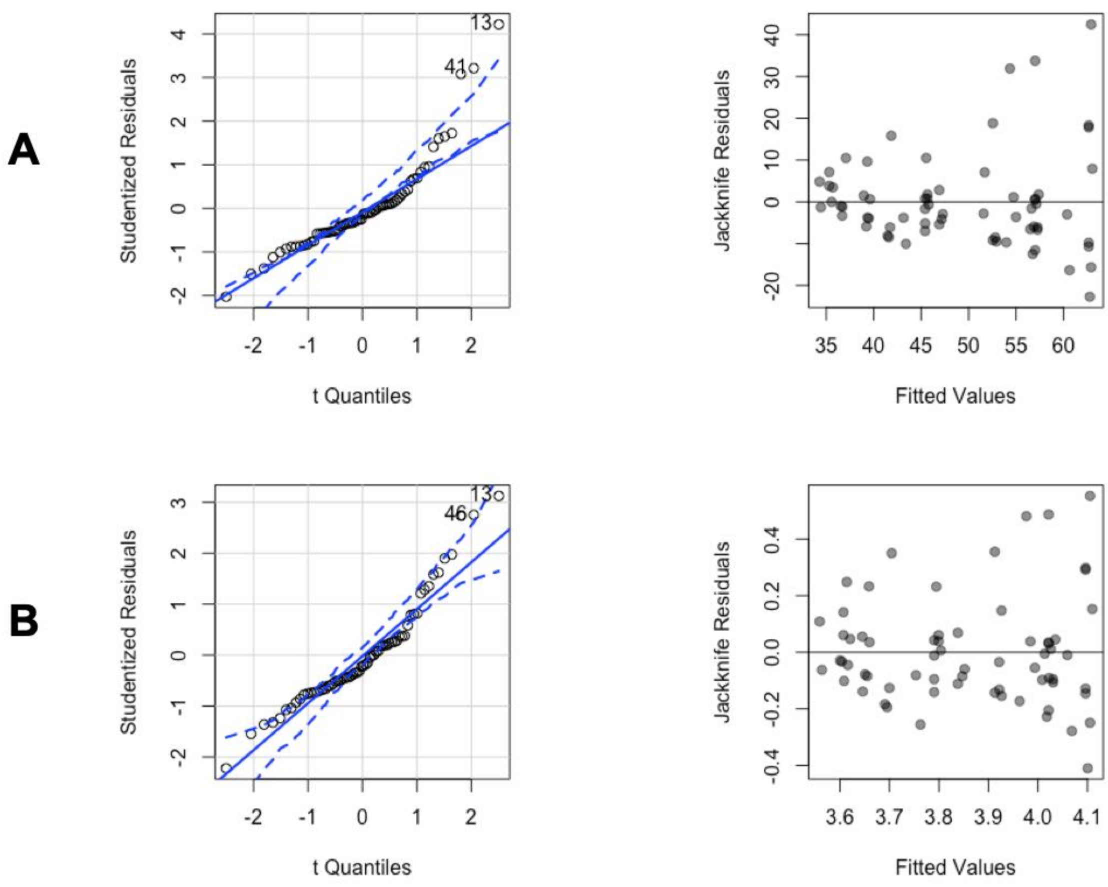


Figure A-1 - Residual analysis of the financial cost objective, with (A) representing the untransformed model and (B) representing the log-transformed model. The left plot for each panel represents the QQ-plot of the studentized residuals, and the right plot for each panel represents the jackknife residuals measured against the fitted values.

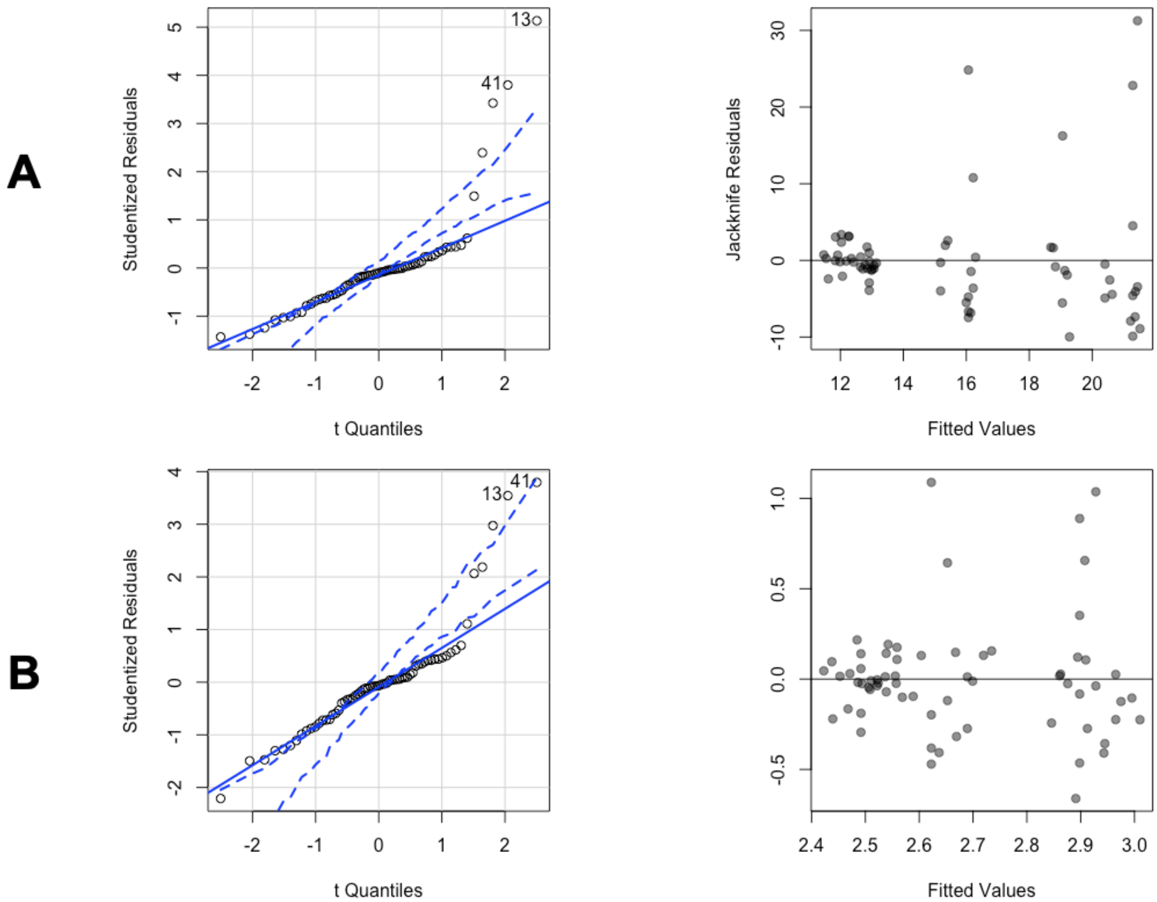


Figure A-2 Residual analysis of the environmental cost objective, with (A) representing the untransformed model and (B) representing the log-transformed model. The left plot for each panel represents the QQ-plot of the studentized residuals, and the right plot for each panel represents the jackknife residuals measured against the fitted values.

Finally, the full regression results of all the models constructed for both Financial Cost and Environmental Cost are presented below in Table A-2 and Table A-3. It is worth noting that the standard errors for each coefficient (and therefore their respective p-values) are heteroskedastic robust and clustered around each group, creating a more rigorous statistical conclusion about their effects.

Table A-2 Regression of Log Transformed Financial Cost Objective

Regression of Log Normal Financial Cost Objective

	Regression	**Regression + Covars**	**Order Interaction**	**Order Interaction + Covars**
Intercept	4.02***	3.92***	4.04***	3.97***
	(0.06)	(0.26)	(0.08)	(0.25)
Collaboration Task	-0.18*	-0.18**	-0.07	-0.07
	(0.08)	(0.06)	(0.08)	(0.08)
CADS Task	-0.31***	-0.31***	-0.31**	-0.31**
	(0.07)	(0.07)	(0.09)	(0.09)
Order Factor Variable			-0.06	-0.09
			(0.11)	(0.11)
Collaboration Task Third			-0.29*	-0.29*
			(0.11)	(0.11)
CADS Task Second			-0.01	-0.01
			(0.15)	(0.15)
Graduate Students %		0.14		0.18**
		(0.07)		(0.06)
Validation Score		-0.00		-0.05
		(0.29)		(0.27)
R^2	0.25	0.29	0.40	0.47
Adj. R^2	0.22	0.24	0.35	0.41
Num. obs.	66	66	66	66
RMSE	0.23	0.23	0.21	0.20

***$p < 0.001$, **$p < 0.01$, *$p < 0.05$

Table A-3 Regression Results on the Log Transformed Environmental Cost Objective

Regression of Log Normal Environmental Cost Objective

	Regression	**Regression + Covars**	**Order Interaction**	**Order Interaction + Covars**
Intercept	2.92***	3.12***	2.94***	3.15***
	(0.09)	(0.41)	(0.12)	(0.41)
Collaboration Task	-0.30*	-0.30**	-0.28	-0.28*
	(0.12)	(0.09)	(0.15)	(0.12)
CADS Task	-0.41***	-0.41***	-0.41**	-0.41*
	(0.09)	(0.11)	(0.13)	(0.15)
Order Factor Variable			-0.06	-0.05
			(0.18)	(0.18)
Collaboration Task Third			-0.08	-0.08
			(0.20)	(0.17)
CADS Task Second			-0.02	-0.02
			(0.20)	(0.20)
Graduate Students %		-0.09		-0.07
		(0.09)		(0.10)
Validation Score		-0.16		-0.18
		(0.44)		(0.43)
R^2	0.23	0.24	0.25	0.25
Adj. R^2	0.21	0.19	0.18	0.16
Num. obs.	66	66	66	66
RMSE	0.33	0.33	0.33	0.34

***$p < 0.001$, **$p < 0.01$, *$p < 0.05$

Table A-4 Gamma Log Link Function Results for Financial and Environmental Cost Objectives

Regression of Financial Cost and Environmental Cost Using Gamma Log Link Function

	Financial Cost	**Environmental Cost**
Intercept	4.03***	3.13***
	(0.23)	(0.45)
Collaboration Task	-0.10	-0.28
	(0.08)	(0.16)
CADS Task	-0.35***	-0.50**
	(0.08)	(0.16)
Order Factor Variable	-0.10	-0.11
	(0.10)	(0.18)
Collaboration Task Third	-0.28*	-0.13
	(0.14)	(0.26)
CADS Task Second	0.03	0.03
	(0.14)	(0.26)
Graduate Students %	0.19**	0.02
	(0.07)	(0.14)
Validation Score	-0.09	-0.10
	(0.25)	(0.48)
AIC	496.41	416.01
BIC	516.11	435.72
Log Likelihood	-239.20	-199.01
Deviance	2.41	7.42
Num. obs.	66	66

***p < 0.001, **p < 0.01, *p < 0.05

Appendix B Supplemental Information for Chapter 3

B.1 Study Design Attribute Information

The full set of policy attributes used in the study can be seen below. Participants were provided with this table for reference in their briefing material for both their individual surveys and for their group discussion. These levels are taken from different climate policies proposed by various candidates during the Democratic Presidential Primary.[8–13] The attributes were selected due to recent polls showing large disagreement on these issues.[14–18] This is to ensure that participants are not already in agreement prior to the group discussion.

Table B-1 Policy characteristics and description

Policy Characteristic	Characteristic Description	Characteristic Levels
Nuclear Power in the Power Grid	Nuclear power is the largest carbon-neutral energy source in the United States. About 20% of the electricity in the U.S. comes from nuclear power. Many people worry about the safety of nuclear power and nuclear waste. Some candidates support new nuclear power plants, while others want to remove many existing plants.	Keep current nuclear power plants but do not build any new plants.
		Keep current nuclear power plants and provide money to increase nuclear power by 14% over the next eight years.
		Shut down nuclear power plants that are not making money, reducing nuclear power by 14-29% than today. Do not build nuclear plants in the next eight years.
Price on Carbon ($/ton)	Putting a price on carbon means polluters would pay when they release greenhouse gas (GHG) into the air. Economists believe a carbon price is a cost-effective way to reduce GHG emissions. However, a high carbon price could lead to higher prices on many goods such as gasoline. Candidates support a wide range of carbon prices, from $0 per ton to $150 per ton of CO2.	0
		30
		60
		90
		120
		150
Fossil Fuel Exploration Rules on Federal Land	A large amount of our fossil fuel is produced on federal lands and waters. Some candidates want a full ban on fossil fuel exploration in these areas. Other candidates want tighter regulations that increases fracking safety standards and publicly say what	Unregulated access to federal lands and waters for fossil fuel exploration.
		Tighter fracking regulation on public lands that increases storage safety standards and transparency of what chemicals are used.

	chemicals fossil fuel companies are using. Finally, some candidates believe producing fossil fuel from federal lands will lower prices and want to keep the current situation.	Fully ban fossil fuel exploration on public lands.
Clean Energy Standard Target	A clean energy standard sets the amount of electricity that must be generated using an approved clean energy source such as solar, wind, hydro, or nuclear. Because the United States electricity sector accounts for ~33% of our GHG emissions today, a clean energy standard will lower the total GHG emitted. However, this will force many existing fossil fuel power plants to close and could increase prices on goods. Candidates have different clean energy targets, ranging from 100% clean energy by 2030 to 100% clean energy by 2100.	Starting from 2021, reach 100% clean energy by 2035 with a yearly increase of 7.5% of clean energy in the grid.
		Starting from 2021, reach 100% clean energy by 2050 with a yearly increase of 3.5% of clean energy in the grid

B.2 Sample User Interface for Individual Survey

Participants used a R Shiny web app to complete their stage 1 survey. An example choice task can be seen below (Figure B-1).

Pairwise Comparisons

You will be asked to choose between two alternatives: A and B. The two alternatives will have different values for some of the attributes (though not all of them). Please use the drop-down menu on the right of each question to select the alternative that you support the most. You must select an alternative for each question. Once you finish all questions, click on Submit Your Answer. You will receive a message letting you know when it is safe to close the browser page. It is important to note that your responses here will determine whether you will be invited back for part 2 of the study, so it is crucial that you answer all questions with your undivided attention and with the best of your ability. It is also important to remember that there is no right answer and we want to understand what your true preferences are.
Warning: Do not close your browser window immediately after clicking on the submit button. Please wait for a green checkmark to appear beside the button before you close the browser. Closing the browser prior to the green arrow may cause the system to not save your answers.

Question 1

Characteristics	Alternative A	Alternative B
Nuclear Power in the Power Grid	Shut down nuclear power plants that are not making money, reducing nuclear power by 14-29% than today. Do not build nuclear plants in the next eight years.	Keep current nuclear power plants but do not build any new plants.
Price on Carbon($/ton)	120	150
Fossil Fuel Exploration Rules on Federal Land	Tighter fracking regulation on public lands that increases storage safety standards and transparency of what chemicals are used.	Fully ban fossil fuel exploration on public lands.
Clean Energy Standard Target	Starting from 2021, reach 100% clean energy by 2100 with a yearly increase of 1.3% of clean energy in the grid.	Starting from 2021, reach 100% clean energy by 2050 with a yearly increase of 3.5% of clean energy in the grid.

Your Choice

Question 2

Characteristics	Alternative A	Alternative B
Nuclear Power in the Power Grid	Keep current nuclear power plants but do not build any new plants.	Keep current nuclear power plants and provide money to increase nuclear power by 14% over the next eight years.
Price on Carbon($/ton)	0	30
Fossil Fuel Exploration Rules on Federal Land	Unregulated access to federal lands and waters for fossil fuel exploration.	Tighter fracking regulation on public lands that increases storage safety standards and transparency of what chemicals are used.
Clean Energy Standard Target	Starting from 2021, reach 100% clean energy by 2100 with a yearly increase of 1.3% of clean energy in the grid.	Starting from 2021, reach 100% clean energy by 2050 with a yearly increase of 3.5% of clean energy in the grid.

Your Choice

Figure B-1 Example choice task for stage 1 individual survey

B.3 Stage Two Cluster Analysis Sample Statistics

Summary statistics of the estimated individual utility function parameters can be found below:

Table B-2 Individual parameter estimation, divided by clusters.

Parameters *(Median, Mean, Standard Deviation)*	**Indifferent**	**Fast Decarbonization**	**Against FF Exploration**	**Against Status Quo**
Intercept	-0.38, -0.68, 4.81	-3.51, -3.95, 3.70	-20.0, -16.7, 7.37	-2.33, -1.97, 4.55
Increase Nuclear	-0.07, -0.76, 4.12	0.81, 1.16, 2.88	-0.22, -0.66, 3.14	-1.46, -2.60, 5.59
Decrease Nuclear	0.10, 0.87, 3.83	-0.55, -1.36, 4.01	-0.21, -0.43, 1.18	-0.10, 0.11, 3.14
CO2 Price	0.04, 0.18, 1.07	0.52, 0.65, 1.35	-0.02, 0.14, 1.22	0.07, 0.27, 1.34
Tighter FF Exploration Regs	0.18, 0.22, 0.71	1.38, 1.69, 2.90	20.6, 15.8, 8.44	2.56, 2.88, 2.71
Ban FF Exploration on PL	0.12, 0.06, 0.75	1.14, 1.16, 3.59	21.2, 15.8, 9.04	2.22, 2.43, 1.82
2050 100% Clean Energy Goal	0.12, 0.46, 4.86	2.32, 3.31, 4.65	1.14, 2.35, 5.18	0.61, 1.15, 3.01
2035 100% Clean Energy Goal	0.09, -0.37, 3.33	4.03, 3.98, 3.56	1.13, 3.92, 7.42	1.16, 0.71, 3.63

This result shows that the range of estimated parameters values are descriptively different for each parameter between all clusters, once again showing that participants from each cluster have different preferences prior to the group discussion. However, it is important to note that there is still a wide dispersion of estimated values for each parameter within each cluster (note the high standard deviation values). While this shows that we cannot say with statistical significance (at $p = 0.05$ level) that all clusters are different for all attributes, we believe this is still sufficient to ensure there is some preference heterogeneity within each group during the group discussion.

B.4 Group Discussion Briefing Material

During the group discussion, the following script was read to the participants:

"Thank you again for coming today to participate in the second stage of our experiment. In front of you is a consent form, and before we start, please read through the consent form carefully and acknowledge the three voluntary confirmation statements at the bottom of the page. If you have any questions, please let me know and I would be happy to answer them.

<Consent form>

Thank you all for agreeing to participate in the study today.

Your goal today is to identify the most appropriate metal part in this Pontoon bridge assembly for MAM. The possible set of parts are the same ones you saw during your initial survey, and they will have the same attributes. There will be three stages in today's group exercise.

Stage 1: You will be presented with 15 comparisons (similar to the individual surveys you filled out earlier), and you will have 2 minutes to discuss the two alternatives as a group. Before the end of the two minutes, you have to vote on which alternative you believe the group judges to be the best part for MAM. There is no need for a consensus, but you are encouraged to share your perspective and work together. You will enter your vote on the screen in front of you, select the appropriate option in the dropdown menu and click on the associated submit button.

Stage 2: Once this is done, you will be presented with a recommendation on what part the group judges to be the best for MAM. You must provide an up or down vote on whether you think the group should choose this alternative. You will also have 2 minutes to discuss your responses.

Stage 3: At the end of stage 2, you will be asked to make 10 further comparisons. Each comparison will always include the recommended part from stage 2, and you will be asked to judge the recommended part against another part on its suitability for MAM. There will only be 30 seconds for each comparison.

Once you complete stage 3, that will conclude our study.

Here are a few final notes: 1) For all three stages, if the group makes a decision before the end of the 2 minutes (stages 1 and 2) or 30 seconds (stage 3), you cannot move onto the next question until the timer expires. 2) During your discussions, we encourage you to share with each other how and why you believe one alternative is better than the other, especially if there is disagreement amongst your group.

Are there any questions?"

B.5 Estimated k Parameter Values for Mean-Variance Model During Stage Three Group Discussion

The estimated k values of groups with the mean-variance models range from -0.10 to 0.09, with most groups clustered around 0 (Figure B-2). This result suggests that these groups' social welfare function most closely matched social welfare functions with equal weights ($k \approx 0$). Filtering the k values to only groups that reached consensus on the recommendation demonstrated similar result (Figure B-3).

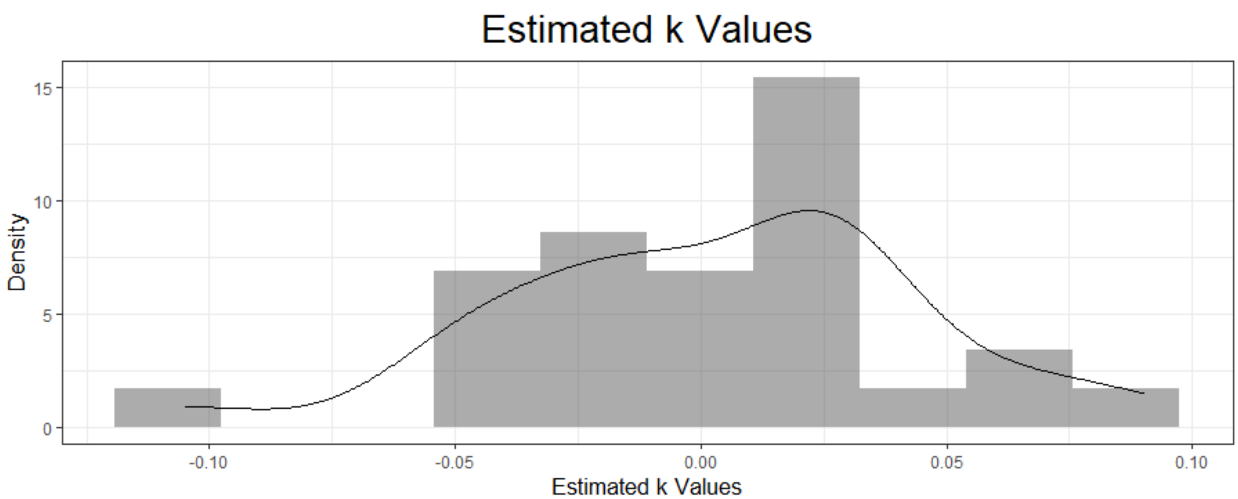


Figure B-2 Estimated k values of groups given recommendations using the mean-variance model. k values range from -0.10 to 0.09, with most values clustered around 0.

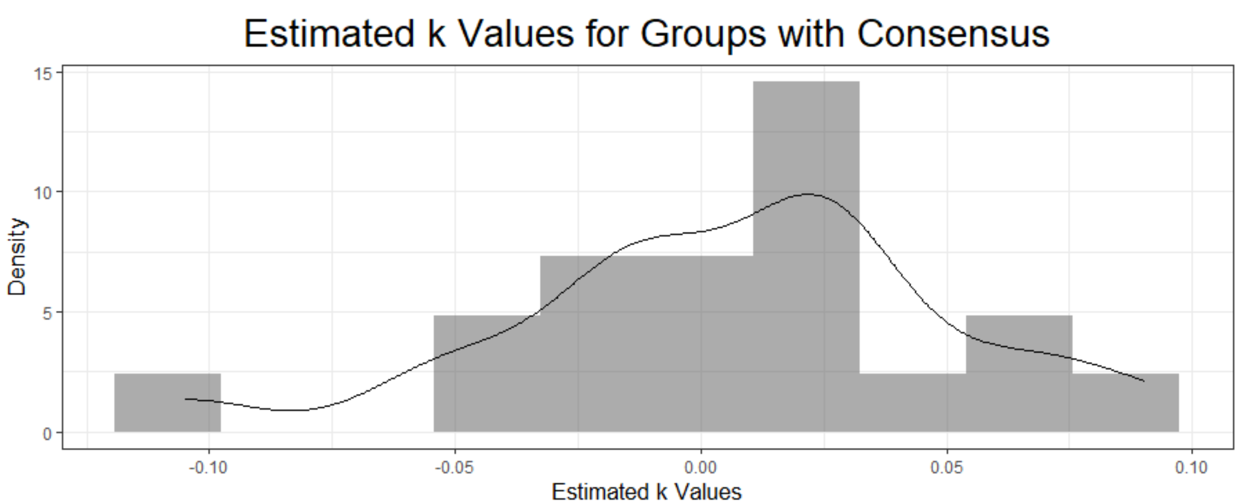


Figure B-3 Estimated k values of groups that reached consensus on recommendations using the mean-variance model. k values range from -0.10 to 0.09, with most values clustered around 0.

Given how close k values were to 0, we calculated the correlation between the final group social welfare function utility values estimated using the mean-variance model and the group social welfare function utility values assuming an equal weight functional form for each group. The correlation between the two social welfare function utility values for each group ranged from 0.84 to 0.99 for all groups receiving the mean-variance recommendation. For 17 out of 27 groups, the top recommendation was the same between the mean-variance model and the equal weights model, and when examining the top five recommendations, the average number of recommendations that are the same for both social welfare functions is 2.6 (range from 0 to 5). This result suggests that while the mean-variance model and the equal weight social welfare functions are similar, there is still some differences in the final social welfare function utility values for all alternatives.

B.6 Group Discussion Simulation Results

Simulation results for the weighted sum model to select the recommendation. We tested four different approaches: information maximization, upper confidence bound (UCB), hybrid approach with 10 recommendations given using information maximization and five using UCB (Hybrid 5), and a hybrid approach with two UCB with (Hybrid 2). UCB approach suggests a recommendation by calculating a new social welfare function based on the estimated parameters and its variances. Information maximization suggests a recommendation that would provide the highest information (d-optimality). We simulated the performance of these approaches by first simulating four participants with known attribute parameters, and then constructed the group utility values using the calculated utility function values for each individual for all policy alternatives. We then simulated the individual votes in the group discussion using their individual utility functions. Based on those simulated votes, we then calculated the final group recommendation based on those results. To evaluate the performance of each approach, we calculated the correlation of each simulated group ranking against the ground truth of the group preferences. Out of the four approaches, all of them with the exception of the Hybrid 5 approach had high correlation, and the Hybrid 2 approach had the least variance in its correlation, making it more likely that the model could select a recommendation that will reflect the ground truth of the group preferences.

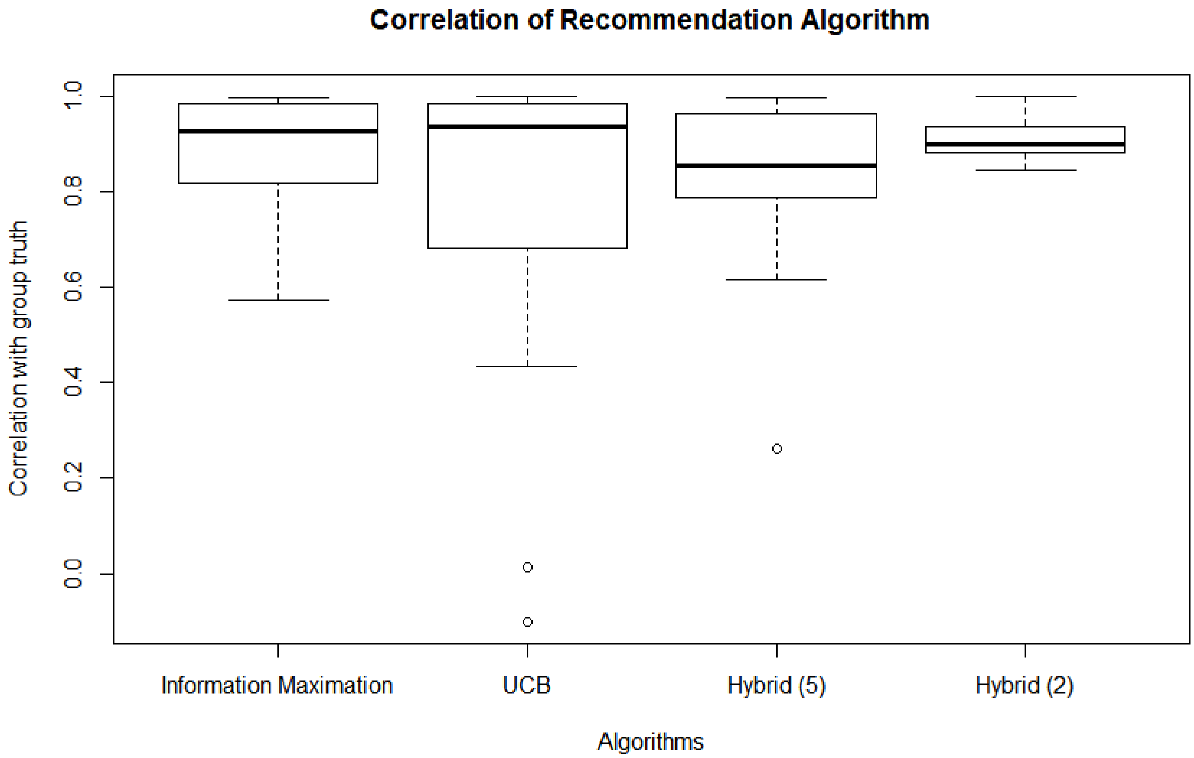


Figure B-4 Correlation of the recommendation algorithm with the four different approaches.

B.7 Individual Mixed Logit Model Validation

To test our pre-registered analysis on an individual level, we used two mixed logit models to test a) whether participants in the two treatment arms differed in their acceptance of the recommendation, and b) whether participants in the two treatment arms differed in the probability of them staying with the recommendation when given random alternatives.

The two model specifications in log-odds forms are:

$$\log\left(\frac{P(L>R)_c}{1\text{-}P(L>R)_c}\right) = \widehat{\gamma_c} + \hat{\theta} \times 1[\text{Treatment Arm} = \text{MV}]_c \ (1)$$

and

$$\log\left(\frac{P(L>R)_{ic}}{1\text{-}P(L>R)_{ic}}\right) = \widehat{\omega_i} + \widehat{\tau_c} + \hat{\zeta} \times 1[\text{Treatment Arm} = \text{MV}]_{ic} \ (2)$$

where the group-level random intercepts (gamma) follow a multivariate normal distribution to allow for dependence of the choice between members of the same group. $\widehat{\omega_i}$ and $\widehat{\tau_c}$ are modeled using a multivariate normal distribution to capture any group-level and participant-level dependence of the choices, and $\hat{\zeta}$ and $\hat{\theta}$ are the treatment effects for the mean-variance group for the two models. The model specifications can be found below (Table B-3):

Table B-3 Model summary statistics for both mixed logit models. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.005$

Coefficients	**Mixed Logit Models**			
	Recommendation Agreement (a)		Validation (b)	
	Estimate	Standard Error	Estimate	Standard Error
Intercept	2.03	0.96*	1.74	0.23***
Treatment (Mean-Variance Model)	1.99	1.12	0.62	0.33

The models showed that participants who received the mean-variance model were more likely to both accept the recommendation, and to stay with the recommendation during the validation stage. While the treatment variables were not statistically significant at $\alpha = 0.05$ level, the size of the coefficient estimate suggests that the choice of model was significant in the outcomes.

In addition, using the DHARMa package[19], we examined the residual diagnostics for our generalized linear mixed models. By examining the residuals for both models, we can identify if there are any potential dispersion, heteroskedasticity, or misspecification problems. We can use the QQ plot of the expected versus observed residuals, along with the Levene Test[7] for homogeneity of the variance of the treatment variable to identify any potential issues.

For both models, both plots showed that there is no obvious model misspecification. KS test, dispersion test, and outlier test values are all significantly above 0.05 for both models (see Figure B-5 and Figure B-6). The Levene Test for homogeneity also shows that the assumption of equal variances holds for both models. This suggests that there are no significant model misspecification issues for both models.

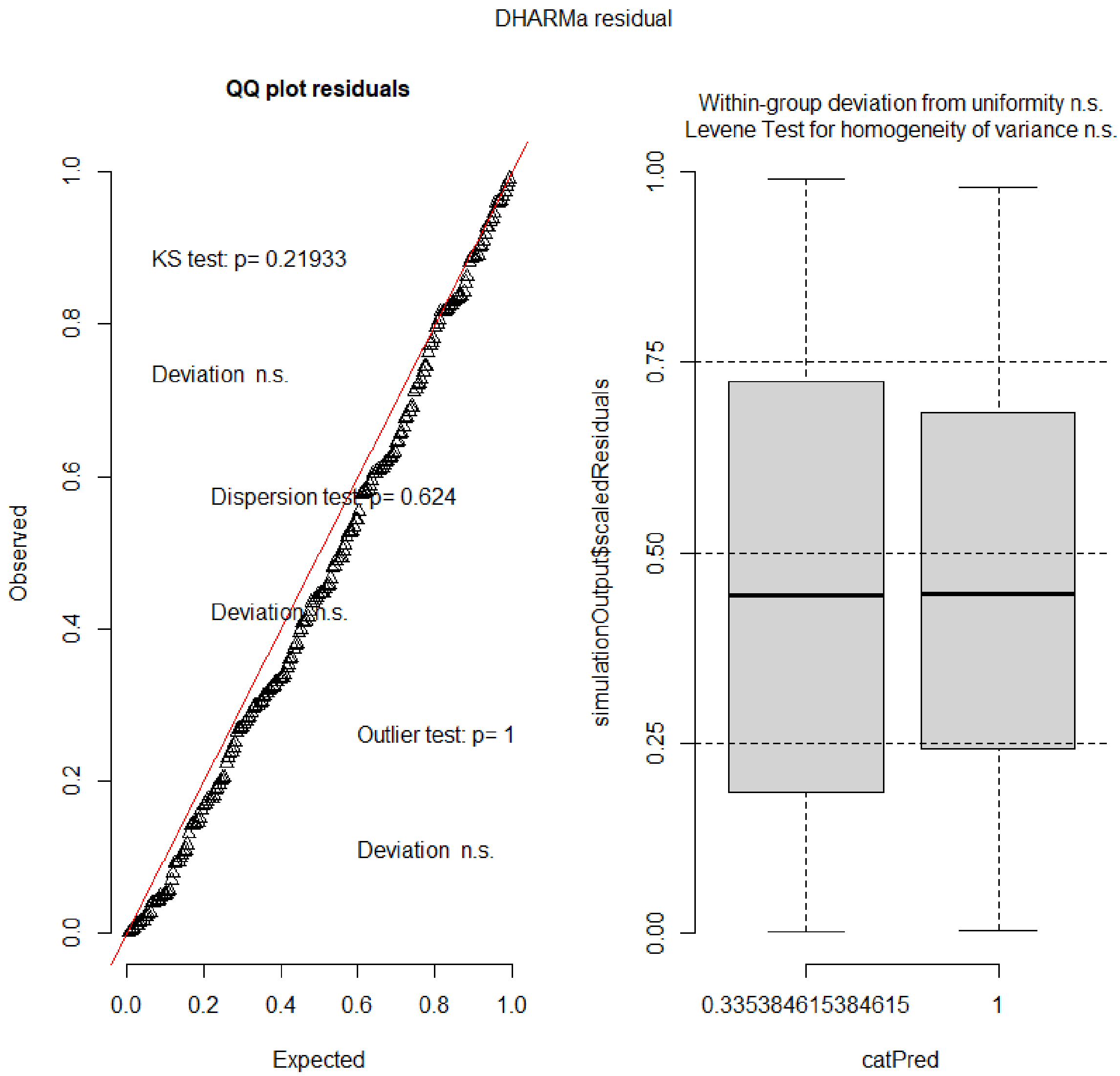


Figure B-5 Recommendation acceptance model. QQ plot of the expected versus observed residuals is on the left, while the Levene Test for homogeneity of variance is on the right.

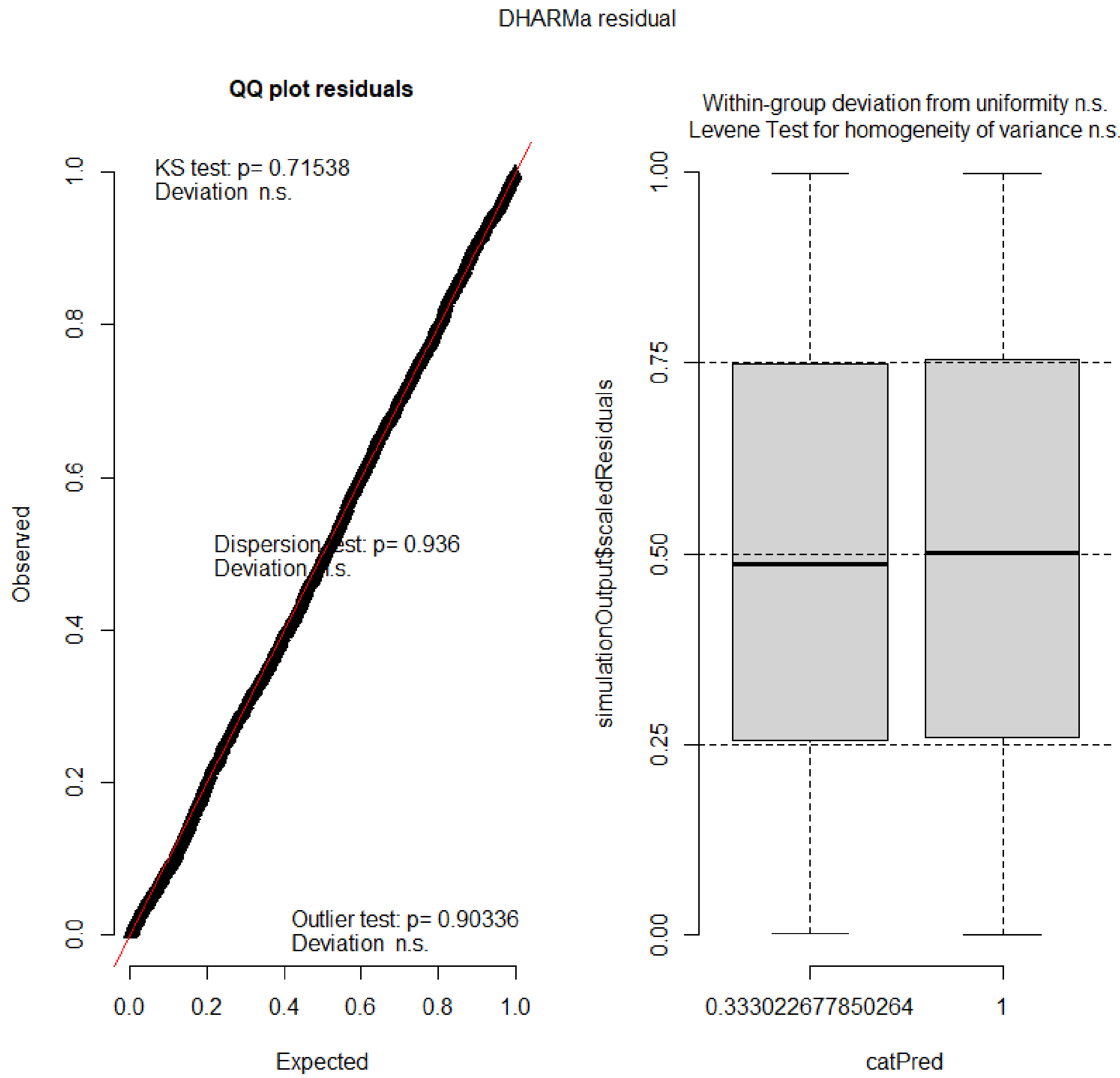


Figure B-6 Individual validation model. QQ plot of observed versus expected residuals is on the left, while the Levene test for homogeneity of variance is on the right.

Appendix C Supplemental Information for Chapter 5

C.1 Semi-Structured Expert Interview Questions

C.1.1 Part Selection Interviews

We are trying to understand the best ways of determining whether components in a part, system, or subassembly are suitable for consolidation using metal additive manufacturing. For example, if we take a pair of scissors, would the blade and the handle be appropriate for consolidation using metal additive manufacturing? We want to know, according to your viewpoint, what the general principles, rules, and methods are for determining whether components should be consolidated.

1. First, I'd like to start off by covering the most recent experience you've had with consolidation in metal additive manufacturing. When was the last time you worked on such a problem? What was the context?

   *Establishes recency.

2. First, I'd like to start off by covering the most recent experience you've had with consolidation in metal additive manufacturing. When was the last time you worked on such a problem? What was the context?
   a. Follow up questions based on the response:
      i. What was the component/assembly
      ii. What was the material
      iii. What was the manufacturing process
      iv. What were its dimensions

   *Establishes Context

3. Were there any tools or programs that you used to help you make the assessment? Computer assisted drawings? Optimization model outputs?
   *Establishes the type of tool for their assessment

4. What were the key considerations, objectives, and criteria in the consolidation approach? How did you perform your assessment?

a. Provide specific examples of the rules or models you used

b. Provide specific criteria

i. How did you determine the outcome of the criteria

5. For the following characteristics, describe what they mean to you during your assessment of a component (use as a followup - provide an example to show what we are looking for these bullet points).

    a. Criticality
    b. Thermal load - cyclic
    c. Thermal load - non cyclic
    d. Mechanical load - cyclic
    e. Mechanical load - non cyclic
    f. Overhangs
    g. Tolerance
    h. Total cost

6. Rate the above characteristics' importance to your decision from 1-10, with 1 being the least important, and 10 the most important.

7. Now think back to the list of objectives, rules, and criteria you listed back in question 4, are there additional rules, objectives, or criteria that you would consider if the component was evaluated in an Army context.

    a. Follow up prompt if expert is unable to think of any:

        i. How about objectives such as mission readiness?

C.1.2 Part Consolidation Questions

We are trying to understand the best ways of determining what components are suitable for metal additive manufacturing (MAM) process. For example, would a turbine blade in an aircraft engine be suitable for MAM? We want to know, according to your viewpoint, what the general principles, rules, and methods are for determining whether components would be a suitable candidate for MAM.

1. First of all, I'd like to start off by covering the most recent experience you've had with determining whether a component was suitable for metal additive manufacturing. This could be either in a research or industry setting. We want to know, according to your viewpoint, what the general principles, rules, and methods are for determining whether components are good candidates for MAM.

*Establishes recency

2. For that most recent experience, can you give a general description of the component that you evaluated? What assembly/sub-assembly was that component in?
    a. Follow up questions based on the response:
    b. What was the component/assembly
    c. What was the material
    d. What were its dimensions
3. Were there any tools or programs that you used to help you make the assessment? Computer assisted drawings? Cost model outputs?
4. What were the key considerations, objectives, and criteria in evaluating whether the component was suitable for MAM? How did you perform your assessment?
    a. Provide specific examples of the rules or models you used
    b. Provide specific criteria
        i. How did you determine the outcome of the criteria

C.2 Part Selection Survey Candidate Information

The set of candidates from the M16 Ramp Bay and M17 Interior Bay of the Army's Improved Ribbon Bridge assembly can be found in Table C-1. There are 15 candidate parts and their attribute information was found through a combination of expert interviews with Army experts and LogiQuest Lite database.[20]

Table C-1 Part selection survey candidates and their associated levels. Attribute levels are determined through survey with Army experts and through the LogiQuest Lite database.[20]

Sub-Assembly	Part	Criticality	Overhang	Mechanical Cyclic Load	Production Price	Mission Criticality	Material	Failure Rate	Supply Shortage
Travel Latch and Receptacles	Manual Control Level	No	Yes	Yes	256	No	Yes	Low (Less than 2 failures in the last deployment)	No
Travel Latch and Receptacles	Strike Catch	Yes	No	No	61	Yes	Yes	Low (Less than 2 failures in the last deployment)	No
Pump and Reservoir	Directional Control Linear Valve	No	Yes	No	88	Yes	No	Low (Less than 2 failures in the last deployment)	Yes
Pump and Reservoir	Manual Control Lever	No	Yes	No	120	Yes	Yes	Low (Less than 2 failures in the last deployment)	No
Interior Bay and Foldlock	Roadway to Bow Ponton Foldlock	Yes	No	Yes	199	Yes	Yes	Low (Less than 2 failures in the last deployment)	No
Interior Bay and Foldlock	Spring Support Pin	Yes	No	Yes	4	Yes	No	High (More than 2 failures in the last deployment)	Yes
Interior Bay and Foldlock	Flat Washer	Yes	No	No	0.01	No	No	High (More than 2 failures in the last deployment)	No
Ramp Bay Inner Pontons Foldlock	Manual Control Lever	No	No	Yes	417	No	Yes	Low (Less than 2 failures in the last deployment)	No

Ramp Bay Inner Pontons Foldlock	Mounting Block	No	Yes	No	100	No	Yes	High (More than 2 failures in the last deployment)	Yes
Ramp Bay Inner Pontons Foldlock	Self-Locking Hexagon Nut	Yes	No	No	1	Yes	Yes	Low (Less than 2 failures in the last deployment)	Yes
Upper Couplings and Receptacle Blocks	Cotter Pin	Yes	No	Yes	1	No	Yes	High (More than 2 failures in the last deployment)	No
Upper Couplings and Receptacle Blocks	Helical Extension Spring	Yes	Yes	Yes	2	No	No	High (More than 2 failures in the last deployment)	No
Upper Couplings and Receptacle Blocks	Lock Release Lever	No	No	No	299	No	Yes	High (More than 2 failures in the last deployment)	No
Upper Couplings and Receptacle Blocks	Double Angle Bracket	No	Yes	Yes	30	Yes	No	High (More than 2 failures in the last deployment)	No
Upper Couplings and Receptacle Blocks	Friction Receptacle Catch Stud	No	Yes	No	695	No	Yes	Low (Less than 2 failures in the last deployment)	No

A screenshot of the discrete choice task can be found in Figure C-1.

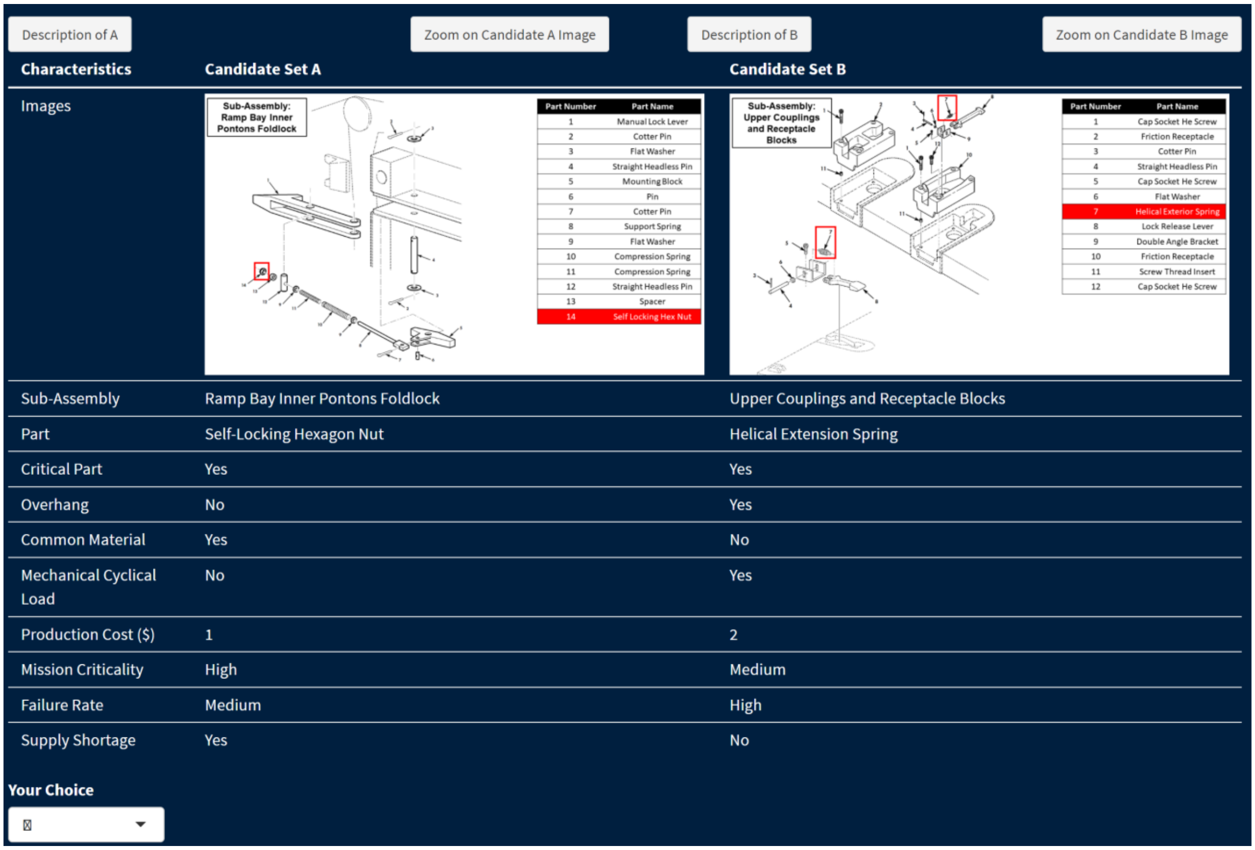
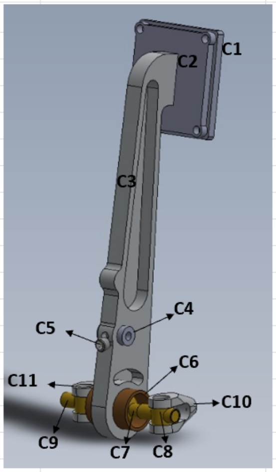


Figure C-1 Screenshot of a sample discrete choice task for the stage one individual survey.

C.3 Part Consolidation Candidate Information

The 11 components in the brake pedal assembly have a total of 2048 possible consolidation combinations. However, after filtering out redundant and infeasible candidates, we reduced total number of combinations to 96. The set of possible consolidation combinations can be seen below in Figure C-2.

	Components	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11
C1	pedal face		F1									
C2	pedal base spacer	F1		F2								
C3	brake pedal arm		F2		F3	F4	F6					
C4	brake pedal axle			F3		F5						
C5	axle hold screw			F4	F5							
C6	balance bar sleeve			F6				F7				
C7	balance bar						F7		F8	F9		
C8	balance bar clevis joint 1							F8			F10	
C9	balance bar clevis joint 2							F9				F11
C10	balance bar clevis 1								F10			
C11	balance bar clevis 2									F11		



Number of all possible candidates: $2^{11} = 2048$
After filtering: 2048 → 96

1. Redundancy?

F3=1	F4=1	F5=1	No
F3=1	F4=1	F5=0	Yes
F3=1	F4=0	F5=1	Yes
F3=0	F4=1	F5=1	Yes

2. Assembly Access Issue

F3=0 F4=1 F5=0 : Not possible
F3=0 F4=0 F5=1 : Not possible

3. Movable Interface (like joint)

F7=1 : Not possible
F8=1 : Not possible
F9=1 : Not possible

Figure C-2 Possible part consolidation candidates after filtering out infeasible consolidation candidates.

A screenshot of the survey UI can be seen below in Figure C-3.

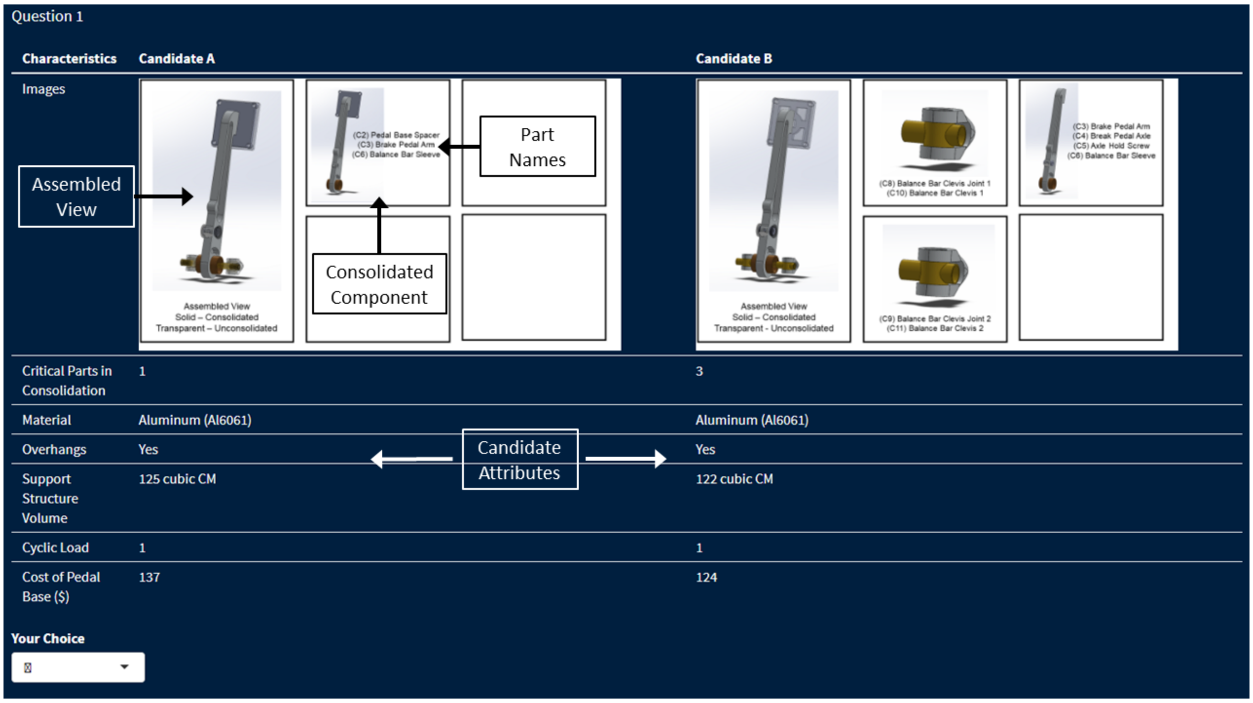


Figure C-3 Screenshot of part consolidation survey discrete choice task.

C.4 Part Selection Candidate Ranking

After applying the aggregate expert model fixed effects, we can determine a ranked order of all known candidates.

Rank	Sub-Assembly	Part	Criticality	Overhang	Mechanical Cyclic Load	Production Price	Mission Criticality	Material	Failure Rate	Supply Shortage
6	Travel Latch and Receptacles	Manual Control Level	No	Yes	Yes	256	No	Yes	Low (Less than 2 failures in the last deployment)	No
9	Travel Latch and Receptacles	Strike Catch	Yes	No	No	61	Yes	Yes	Low (Less than 2 failures in the last deployment)	No
8	Pump and Reservoir	Directional Control Linear Valve	No	Yes	No	88	Yes	No	Low (Less than 2 failures in the last deployment)	Yes
5	Pump and Reservoir	Manual Control Lever	No	Yes	No	120	Yes	Yes	Low (Less than 2 failures in the last deployment)	No
7	Interior Bay and Foldlock	Roadway to Bow Ponton Foldlock	Yes	No	Yes	199	Yes	Yes	Low (Less than 2 failures in the last deployment)	No
11	Interior Bay and Foldlock	Spring Support Pin	Yes	No	Yes	4	Yes	No	High (More than 2 failures in the last deployment)	Yes
15	Interior Bay and Foldlock	Flat Washer	Yes	No	No	0.01	No	No	High (More than 2 failures in the last deployment)	No
2	Ramp Bay Inner Pontons Foldlock	Manual Control Lever	No	No	Yes	417	No	Yes	Low (Less than 2 failures in the last deployment)	No

4	Ramp Bay Inner Pontons Foldlock	Mounting Block	No	Yes	No	100	No	Yes	High (More than 2 failures in the last deployment)	Yes
12	Ramp Bay Inner Pontons Foldlock	Self-Locking Hexagon Nut	Yes	No	No	1	Yes	Yes	Low (Less than 2 failures in the last deployment)	Yes
13	Upper Couplings and Receptacle Blocks	Cotter Pin	Yes	No	Yes	1	No	Yes	High (More than 2 failures in the last deployment)	No
14	Upper Couplings and Receptacle Blocks	Helical Extension Spring	Yes	Yes	Yes	2	No	No	High (More than 2 failures in the last deployment)	No
1	Upper Couplings and Receptacle Blocks	Lock Release Lever	No	No	No	299	No	Yes	High (More than 2 failures in the last deployment)	No
10	Upper Couplings and Receptacle Blocks	Double Angle Bracket	No	Yes	Yes	30	Yes	No	High (More than 2 failures in the last deployment)	No
3	Upper Couplings and Receptacle Blocks	Friction Receptacle Catch Stud	No	Yes	No	695	No	Yes	Low (Less than 2 failures in the last deployment)	No

Appendix D Appendix References

1. Bartholomew, T. V. & Mauter, M. S. Multiobjective Optimization Model for Minimizing Cost and Environmental Impact in Shale Gas Water and Wastewater Management. *ACS Sustain. Chem. Eng.* **4**, 3728–3735 (2016).
2. Jucks, R. & Bromme, R. Perspective taking in computer-mediated instructional communication. *J. Media Psychol.* **23**, 192–199 (2011).
3. Bromme, R., Jucks, R. & Runde, A. Barriers and biases in Computer-Mediated Expert-Layperson-Communication. in *Barriers and Biases in Computer-Mediated Knowledge Communication: And How They May Be Overcome* (eds. Bromme, R., Hesse, F. W. & Spada, H.) 89–118 (Springer Science & Business Media, 2006).
4. Keysar, B., Barr, D. J. & Horton, W. S. The Egocentric Basis of Language Use. *Curr. Dir. Psychol. Sci.* **7**, 46–49 (1998).
5. Stein, C. M. Estimation of the Mean of a Multivariate Normal Distribution. *Ann. Stat.* **9**, 1135–1151 (1981).
6. Nelder, J. A. & Wedderburn, R. W. M. Generalized Linear Models. *J. R. Stat. Soc. Ser. A* **135**, 370 (1972).
7. Schultz, B. B. Levene’s Test for Relative Variation. *Syst. Biol.* **34**, 449–456 (1985).
8. Plan for Climate Change and Environmental Justice | Joe Biden. Available at: https://joebiden.com/climate-plan/. (Accessed: 22nd September 2021)
9. Issues | Bernie Sanders Official Website. Available at: https://berniesanders.com/issues/. (Accessed: 22nd September 2021)
10. Tackling the Climate Crisis Head On | Elizabeth Warren. Available at: https://elizabethwarren.com/plans/climate-change. (Accessed: 22nd September 2021)
11. Amy Klobuchar Views on 2020 Issues: A Voter’s Guide - POLITICO. *Politico* (2020). Available at: https://www.politico.com/2020-election/candidates-views-on-the-issues/amy-klobuchar/. (Accessed: 22nd September 2021)
12. An Evergreen Economy for America | Jay Inslee for Governor. Available at: https://www.jayinslee.com/issues/evergreen-economy. (Accessed: 22nd September 2021)
13. Michael Bennet Views on 2020 Issues: A Voter’s Guide - POLITICO. *Politico* (2020). Available at: https://www.politico.com/2020-election/candidates-views-on-the-

issues/michael-bennet/. (Accessed: 22nd September 2021)
14. Kaufman, A. Voters Back Ban On Fracking, New Poll Finds | HuffPost. *Huffington Post* (2019). Available at: https://www.huffpost.com/entry/fracking-ban-polling_n_5d701602e4b09bbc9ef95ef9. (Accessed: 7th September 2021)
15. Energy Policy Institute. New Poll: Nearly Half Of Americans Are More Convinced Than They Were Five Years Ago That Climate Change Is Happening, With Extreme Weather Driving Their Views | EPIC. *EPIC NEWS* (2019). Available at: https://epic.uchicago.edu/news/new-poll-nearly-half-of-americans-are-more-convinced-than-they-were-five-years-ago-that-climate-change-is-happening-with-extreme-weather-driving-their-views/. (Accessed: 16th January 2022)
16. Reinhart, R. 40 Years After Three Mile Island, Americans Split on Nuclear Power. *Gallup* (2019). Available at: https://news.gallup.com/poll/248048/years-three-mile-island-americans-split-nuclear-power.aspx. (Accessed: 7th September 2021)
17. Newport, F. Americans Tilt Toward Protecting Environment, Alternative Fuels. *Gallup* (2017). Available at: https://news.gallup.com/poll/206159/americans-tilt-toward-protecting-environment-alternative-fuels.aspx?g_source=link_newsv9&g_campaign=item_225053&g_medium=copy. (Accessed: 16th January 2022)
18. Swift, A. Opposition to Fracking Mounts in the U.S. *Gallup* (2016). Available at: https://news.gallup.com/poll/190355/opposition-fracking-mounts.aspx. (Accessed: 16th January 2022)
19. Hartig, F. DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models. (2022).
20. Terabase Corporation. LogiQuest Lite. Available at: https://lqlite.com/.