Detecting the influence of stakeholders' mental models on emergent collective awareness in instrumented teamwork workshops

by

Kevin P. McDonough

B.S., Computer Engineering, University of Rhode Island (2004)

Submitted to the System Design and Management Program
in partial fulfillment of the requirements for the degree of

Master of Science in Engineering and Management

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2021

Author .....
System Design and Management Program
May 21, 2021

Certified by .....
Bryan R. Moser, Ph.D.
Academic Director & Sr. Lecturer, System Design and Management Program
Thesis Supervisor

Accepted by .....
Joan S. Rubin
Executive Director, System Design and Management Program

Detecting the influence of stakeholders' mental models on emergent collective awareness in instrumented teamwork workshops

by

Kevin P. McDonough

Submitted to the System Design and Management Program
on May 21, 2021, in partial fulfillment of the
requirements for the degree of
Master of Science in Engineering and Management

Abstract

The use of models to represent, investigate, and explain the world extends across disciplines, professions, and walks of life. From supporting learning in the classroom, to aiding organizational decision making, to influencing people's daily lives by informing them about the weather, patterns of disease spread, and climate change, models pervade our lives. In Engineering, models provide a mechanism through which teams may organize, align, and share knowledge, communicate across role and domain of expertise, and develop new insights. Models enable stakeholders to learn and make more informed decisions in the face of complexity and uncertainty. While the value of models for representing complex sociotechnical systems-of-systems has been demonstrated, what is lesser well known is how a user's knowledge and perceptions about the systems being represented mediate the use and efficacy of the models. This work explores one aspect of this phenomena — how the diversity of a team's mental models affects, and is affected by, their use of a system model. The design of a series of instrumented team experiments was developed and a teamwork research platform, using an agent based modeling and simulation framework, was created. A series of three instrumented teamwork workshops was conducted. Twelve teams participated in the workshops, role-playing as expert stakeholders, in the exploration of options for population, transportation, and function in site designs for a conceptual human settlement on Mars. A diversity in mental models distinguishable from postulated generative distributions was detected. This work demonstrates the use of instrumented methods to detect, quantify, and analyze mental models and tradespace exploration by users of a system model.

Thesis Supervisor: Bryan R. Moser, Ph.D.
Title: Academic Director & Sr. Lecturer, System Design and Management Program

Dedication

To my son, Kevin Theodore McDonough ("Beag"), for always inspiring and motivating me to make a difference and leave a lasting impact for the greater good.
—October 26, 2018 2:07PM ET

Acknowledgments

I would like to thank the following individuals and organizations for their support and contributions.

I cannot start my acknowledgements without thanking Katherine M. Carroll for her motivation, support, and collaboration. I am especially grateful for her work in the development of the ABMS software used in our research and her assistance in running the teamwork workshops. Without all of Katie's efforts, the foundations upon which this work was built would not have been as strong.

I would like to thank my professor and advisor, Dr. Bryan R. Moser for his insights, mentoring, and encouragement. All of the conversations and guidance on instrumented teamwork, ABMS, and doing good research were invaluable and without which, much of this thesis would remain part of "the tradespace unseen".

As part of our research efforts, Dr. Keiji Kimura has contributed immensely to the validation of our teamwork research platform through his application of the simulator to a new model problem and his thorough analysis of simulation results.

George and Alexandros Lordos deserve a special acknowledgement and much appreciation for their design of the Star City concept and for permitting and encouraging its use as part of this work.

Additionally, this research would not have been possibly without the sponsorship of East Japan Railway Company, thank you for your interest in, and pursuit of, model based approaches to solving real world challenges. I also extend my gratitude to all the many people who make up the wider MIT community which has supported, shaped, guided, and participated in this work and my time here. In particular, the MIT Global Teamwork Lab, professor Oiliver de Weck and Dr. Afreen Siddiqi of Course 16, and all of the workshop participants from SDM, ESL, EPFL, and beyond.

Finally, I would like to thank my son, Kevin T. McDonough, and my best friend, Karlee R. Markovich for their unending love and understanding over the course of my pursuit of this research and my degree.

Contents

List of Figures

List of Tables

Glossary

design vector The set of variable values representing a particular design configuration for a system model. 13, 47

design walk The ordered enumeration of design vectors and their resulting objective vectors produced by an individual or a team during the use of a system model. 13, 47, 48, 74

objective vector The set of performance metric values produced by a system model, as configured based upon a given design vector. 13, 47

Pareto optimal The state of being in which no metric of evaluation can be improved upon without degrading another metric. In the case of a system design, a design which exhibits Pareto optimality when compared to alternative designs. 13, 86

Pareto Rank The ordinal step in which an entity, from among a set, exhibits Pareto optimality during the step-wise removal of all Pareto optimal entities from the set. 13, 46

problem space The set of possible systems model performance outcomes (objective vectors) achievable from the set of designs within the solution space. One half of the tradespace of a system model. 13, 20, 21, 47, 137

solution space The set of all possible system designs (design vectors) under the actual (rather than known or believed) feasibility constraints of a system model. One half of the tradespace of a system model. 13, 20, 21, 23, 24, 47, 137

tradespace The space of architecturally feasible choices for a system model represented by its solution space and problem space. 13, 47

Acronyms

ABM Agent Based Model. 13, 75, 76

ABMS Agent Based Modeling & Simulation. 6, 9, 10, 13, 36, 64, 72, 74–77, 80–83, 92, 144–146, 149, 150

GTL Global Teamwork Lab. 6, 13, 23, 24, 36, 39, 41, 42, 51, 75, 81, 82, 151–153

I-P-O Input-Process-Output. 13, 28, 151

JRE East Japan Railway Company. 6, 13, 65, 75

Chapter 1

Introduction

This chapter explores the motivation for this research and establishes a common context for consideration of teamwork, collective awareness, and the mental models of stakeholders. The research questions and hypotheses that were studied are then developed, before an outline for the entirety of this work is given.

1.1 Research Motivation

Teamwork and collective decision making are complex social phenomena that have played a fundamental role in the development of human culture, and are core to how humanity accomplishes much of what it does[1]. The complexity of these phenomena comes, at least in part from, the fact that scale, duration, form, and internal function can vary significantly[2]. Within this vast landscape of human teamwork, understanding teams within the context of engineering projects is of particular interest and value when we consider the impact of team performance and collective decision making ability on project outcomes.

As de Weck, et al. framed so well[3], the discipline of Engineering has matured over time. Its focus has expanded from small, deterministic systems within a single technical domain, to much larger, more complex systems that span multiple technical domains and have non-linear and non-deterministic behaviors. Today, engineers stand before the monumental challenge of working with and understanding multiple,

complex systems, having dynamics and interactions not just across technical interfaces, but also social and natural systems. Indeed, the very practice of Engineering is no longer rooted in just the production of an artifact suited to providing a solution to a problem. Instead, modern Engineering is grounded in the co-design of technical, regulatory, and economic elements, and their relationships, as a solution to a set of related problems in the broader societal context, all in an increasingly interconnected and interdependent world. Faced with the ever growing complexity in this challenge, Engineering has sought to leverage formalized approaches to modeling these systems in order to foster shared understanding in cross disciplinary engineering teams, enhance team communications, and control modification of system design[3]. By using models, engineering teams with these capabilities are able to surface and explore the relationships and tradeoffs that exist between the solution space and the problem space of the system — the solution space of a system being represented by the available set of decisions around the composition of elements that comprise the system, and their relationships, whereas, the problem space is represented by particular figures of merit or performance metrics for such a set of decisions, that are of interest to stakeholders.

Engineering has brought to bear scientific and technical advancements throughout human culture and across societies over the centuries, as can be seen by many of our current systems, including communication, transportation, agriculture, and healthcare. Despite the great advances, significant, emergent problems have arisen that are now being faced by society writ large at a global scale — climate change, food & water insecurity, social inequality, and information partisanship. The confluence of these negative, social externalities, driven by the development of our technological systems, has led to the study, and in turn, modeling, of sociotechnical systems; systems that encompass both the technological elements and relationships of a system as well as specific social elements and their relationships with each other and the technological elements. Such modeling and study can help to elucidate the dynamics and emergent behavior of the combination of people and the systems we build[3].

The goal of Engineering is to produce an artifact, in response to a problem, that

provides an external function of value to one or more stakeholders. In order to be successful in this endeavor, the engineering discipline seeks to understand the relationships between the problem and solution spaces, in order to predict, or at least illuminate, the key dynamics and emergent behaviors of the artifact. In the case of a sociotechnical system, the dynamics and emergent behaviors between the artifact and the social context beyond its technical system boundary must be understood as well. All of this is, at its core, an effort to select from a set of potential solutions based upon the tradeoffs that exist in the solution to problem space mapping. However, in selecting from such a set of potential solutions bounded rationality becomes critically important — rational decision making is bounded by the limited knowledge and capabilities of the decision maker(s), which leads to mental simplification of reality and can result in divergence from ideal, rational choices[4]. The value, therefore, of the use of models in Engineering is exactly to address such limitations in the knowledge and capabilities of those designing the artifacts. By more completely representing actuality, models extend human mental representations, enabling decision makers to explore the problem space-solution space relationships in real time, overcoming human cognitive limits and, ideally, facilitating a more robust exploration of potentially desirable designs from which to select.

As previously mentioned, the intention behind model use in Engineering goes beyond simply the unshackling of a single engineer from the constraints of their limited ability to make a rational decision. Rather, models enable *teams* of engineers to expand their collective mental capabilities, awareness, and considered set of potentially desirable solutions, as well as facilitating and enhancing team communication. The working premise of this approach to Engineering is that that these activities by a team, and thus the use of models themselves, fundamentally bring value to an engineering project. Taken together, this goal of model use and the implicit valuation of the practice seems to indicate that the value is realized through the shaping of the mental models of both expert and non-expert stakeholders. In turn, the mental models influence the decisions that are made, and ultimately impacts the outcome of the engineered artifact and resultant sociotechnical system that gets realized, with all

of its positive and negative emergent effects. As such, a body of research has been built around team behavior in relation to the team's interaction with these models[5][6][7][8]. Such studies bring understanding, detection, and formalization of the drivers of team dynamics and the processes of teamwork as represented by collective awareness of systemic effect. This work has been in that same vein, looking at the sociotechnical system that is a team working with a systems model, communicating and making decisions. This work is contextualized in the larger body of teamwork related research more broadly such that the results and applicability of the findings herein can be appropriately generalized.

Thus, in this study of teamwork and systems models, it is not the team's achievement of desired outcomes, as measured by the system's figures of merit, that is of particular, specific interest. Rather, it is the definition, measurement, and analysis of a team's performance in achieving those desired outcomes *and the drivers behind such achievement*, that is being sought. Herein, it will be argued that a team's performance can be defined and quantified through the lens of the team's alignment on understanding of the systems being designed, exploration of the overall problem and solution tradespace represented by a model, discussion and decision making leading to awareness of systemic effect, assessment of desired outcome achievement, and individual and collective learning resulting in the revision of mental models.

1.2 Research Questions

In an effort to progress the state-of-the-art within the teamwork field of study, **this work seeks to detect, measure, and analyze the influence of stakeholder diversity in team collaboration and coordination, on performance when using a systems model to solve a sociotechnical problem**. Often, diversity is considered from an exogenous perspective, that is, through the lens of the external attributes of an individual that are thought to distinguish them, or classify them as part of a larger group (age, gender, race, ethnicity, profession, etc.). However, in this work, diversity has been considered from an endogenous perspective with

respect to the stakeholder and their understanding of the system(s) represented by a model. This understanding can be framed based upon the individual's evaluation of desirability and feasibility for solution space choices (architectures) representable with the model; these measures being representative of their understanding of anticipated performance, fitness of purpose, and achievability. Thus, desirability and feasibility are a proxy for the accuracy of the individual's mental model relative to the optimal performance achievable according to the systems model, and the alignment of their mental model with other stakeholders.

From this framing of diversity and the desire to detect teamwork phenomenon, a series of potential research questions arise: How do we represent and measure stakeholder mental models of desirability and feasibility? How do we quantify and compare such measurement of mental models and against what basis should comparisons be anchored? What influence do these mental models have upon the dynamics and behaviors of teamwork and how can those influences be measured? Does the diversity of stakeholder mental models influence the emergent performance of the team as measured by collective awareness of systemic effect during, and arising from, systems model exploration? How does one measure collective awareness of systemic effect or quantify systems model exploration by a team? Do the stakeholder mental models change as a result of teamwork and what is the influence of the original composition of mental model diversity on such changes and through what processes (negotiation, learning, etc.) by what drivers (discussion, exploration, etc.)? Does the diversity of stakeholder mental models leading into teamwork ultimately influence the final evaluation and selection decisions of the individual stakeholders post teamwork, and in what way?

This research builds upon teamwork literature that goes back as far as the 1920's, as well as the experimental methods and findings of work done out of MIT's Global Teamwork Lab (GTL)[9]. The Global Teamwork Lab, in collaboration with the University of Tokyo, is working to "build the teamwork laboratory of the future" in order to better quantify the results formally captured by social science approaches to teamwork research. With a focus on meso-scale teamwork, defined as the teamwork

of teams, existing above the dynamics derived from the interaction of the attributes and personalities of individuals within a single team, but below the dynamics of enterprise scale behaviors, the GTL's aim is to produce novel, reproducible, scalable, and insightful research through "instrumented" teamwork experiments. In particular, the work of this group has previously demonstrated methods in detecting and measuring teamwork dynamics, behaviors, and emergent performance.

As part of MIT's GTL, Pelegrin et al.[7] looked at how social dynamics and behaviors of the team mediate tradespace exploration and thus learning. Their method employed four pilot workshops followed by a primary experiment using a system-of-systems model and simulation with supporting, instrumented user interface and additional social (audio, etc.) sensors. These workshops looked to correlate indicators of "surprise" captured by the social sensors, with the teams' allocation of attention to system objective tradeoffs and changes in tradespace exploration patterns indicative of increased awareness of systemic effect (per se, learning). Additionally, Pelegrin et al. defined[10], and subsequently Manandhar et al. measured[8], the effect of team strategy in their tradespace exploration. In the work of Manandhar et al., 50 teams of non-expert stakeholders (n=199) modeled and simulated the design of an innovation campus where teams variously received different treatments on their discussion of design exploration strategy. Again, a system model and simulation with supporting, instrumented user interface was used. The impact of the strategy discussions on the teams' allocation of attention to "systemically-significant portions" of the problem and solution space was then analyzed.

It is from within this context of the above referenced work that the following, specific, research questions and hypotheses were framed, and, subsequently used to build out the methods and experimental design covered herein:

Given a team of stakeholders evaluation designs for a sociotechnical system using a systems model —

- **Can we detect and measure the diversity of mental models within a team?**

- **Does diversity of stakeholder mental models within a team affect that team's performance?**
- **Can one detect a pattern of model exploration by a team that is indicative of performance resulting from the emergence of collective awareness of systemic effect?**
- **Does the emergence of collective awareness of systemic effect lead to greater post-system-evaluation alignment of a team's mental models?**

From these questions, the following hypotheses are set forth:

Hypothesis 1. *Teams having a greater diversity of mental models of system architecture among stakeholders are more likely to be high performing teams.*

Hypothesis 2. *High performing teams will exhibit a more diverse exploration of architectures within a systems model.*

Hypothesis 3. *A more diverse exploration of architectures within a systems model will result in greater alignment among stakeholder mental models.*

1.3 Thesis Outline

Chapter 1 starts with a discussion of the research motivation, research questions, and hypotheses explored in this work. Chapter 2 presents a literature review framing the foundation of previous work in the domains of teamwork research (broadly), mental models of individuals and teams, and instrumented teamwork experiments. Chapter 3 describes the research hypotheses, methods, and experimental design used in generating data for analysis. The results of the experiments run and analysis of data generated are presented in Chapter 4. Finally, Chapter 5 closes this work with a discussion of the results from Chapter 4 and their shortcomings, followed by conclusions drawn from the discussion, insights gained from this work, and next steps including future work.

1.4 Summary

This chapter has presented the motivation for this work by framing the importance of teamwork, the use of models in Engineering, and the influence of the intersection of teamwork and model use on team performance outcomes. The development of research questions and hypotheses was then presented and an outline for the entirety of the work given.

Chapter 2

Literature Review

This chapter looks at the existing body of relevant related research. A brief summary of teamwork research more broadly is first given, followed by work on mental models at an individual level and the shared mental models of a team. Next, research on the use of instrumented teamwork workshops and a discussion of the research measuring tradespace exploration are presented. This serves to lay out a roadmap of the theoretical models, frameworks, principles, and methods that guided this research.

2.1 Teamwork Research

According to Cannon-Bowers[9], the formal investigation of teamwork within organizations goes back at least as far as the 1920s. While Sundstrom et al.[11] attributes the limited use of teams before the 1960s, instead favoring individual, specialized, Taylor-esque jobs, to the scant research around the topics of teams and teamwork before this time. Regardless, there is broad agreement that teamwork plays an even more critical role today than in the past, as the globalization of business has increased teamwork related demands, such as coordination and collaboration[12], upon organizations[9][13][14].

In many pieces of teamwork literature the nature and definition of "team" and "teamwork" are the first point of discussion. Sundstrom et al.[12] defines a teams as "interdependent collections of individuals who share responsibility for specific out-

comes for their organizations", while Bowers et al.'s[14] working definition is a bit more nuanced, specifying the nature of interaction between individuals to be dynamic, adaptive, interdependent, and *differentiated*, all in service of a goal that is shared and valued. Maitheu et al.[15] borrows a definition from Salas et al.[16] that is similar to the previous definition, but includes assignment of specific functions or roles to the individuals and recognizes the temporary nature of team membership. Furthermore, some authors make a distinction between a "team" and a "group"[11], and this complexity is then only further compounded by the lack of agreement upon the definition of "group" within the literature[13]. Finally, Cannon-bowers[9] undertook a similar review of various authors' definitions of "team" and came to more or less the same set of properties: interdependence, role specialization, shared responsibilities, the ability to adapt in order to reach an objective that is valued, and external recognition as an entity.

Moving beyond the efforts of simply defining the entity under study there is a large corpus of research across the many attributes of teams and the factors that govern teamwork. Of particular importance are those works that organize these attributes and factors into models and frameworks usable by others for performing research. In 1964 McGrath[17] introduced a framework for the analysis of teams that classified the elements of study into seven categories: group composition, group structure, task and environment, group process, group development, effects on members, and task performance. These seven categories broadly fell into three higher level constructs which then form a model for describing teamwork. The first three categories, composition, structure, and task and environment, represented the elements that serve as inputs into a team. The fourth category, group process, represented the emergent result of those inputs through the activities of a team as mediated by its social interactions, the process of teamwork. Finally, the last three elements, group development, effects on members, and task performance, represented the outputs of the process of teamwork. In the framework described, these elements of output have direct effects upon each other and upon the three elements of input into a team. Taken together, the three higher level constructs make up McGrath's Input-Process-Output (I-P-O) model of

teamwork.

As can be seen in the definitions of "team", at the heart of teamwork is the accomplishment of a shared goal or task. In McGrath's model, this output is but one element of many, and yet without it, the efforts of the process are meaningless. Thus, much of teamwork literature rightly focuses specifically on the task performance element of the model. A key insight into the value to teamwork and the study of performance comes from Dyer's[18] oft cited review of teamwork research. In it, Dyer looked at various perspectives on performance and concluded that for those tasks that require more than one person, teamwork has a positive effect on performance such that the net outcome of the team is more than simply the sum of the individual efforts of the team members. This emergent benefit of teamwork though is not a universal feature of all discussions of performance. Sundstrom et al.[12] defined performance as "acceptability of output to customers within or outside the organization who receive team products, services, information, decision, or performance events". However, they also recognized the relationships captured in McGrath's model between the team process and the team itself, as well as the task performance and the team. In doing so, they framed the output of a team more holistically as "effectiveness" and included in that the "viability" of the team, defined as "member's satisfaction, participation, and willingness to continue working together"[12]. This broad definition included elements of cohesion, coordination, communication, problem solving, norms, and roles[12].

Numerous other topics in the literature on teams and teamwork exist, including work on the classification of teams and tasks, selection and training of teams, and analysis of team processes, qualities, and states. While all of these topics bear relevance in most any study of teams and teamwork, of particular importance to this work is the latter category. Specifically, that research dealing with team processes involving mental models.

2.2 Mental Models

Research on mental models originates from the fields of human psychology, cognition, and neuroscience[19]. Mental models are often distinguished from other similar concepts in these domains of research, such as schemas, scripts, and frames[20][21]. While all of these concepts are thought to be part of human cognition and play various roles in the storage, interpretation, and use of knowledge, Seel[22] provides a key, distinguishing feature of mental models — the the ability to provide for mental simulation. Thus, mental models are a critical cognitive construct for understanding and reasoning about physical and social systems. It is this act of simulation through mental models that allows for the interpretation of anticipated outcomes in order to infer attributes and relationships about the subject of the model, and thus enable understanding and decision making[21][22]. A slightly broader interpretation is taken by Carroll et al.[23] in their look at the role that the mental models of individuals play in shaping organizational decisions, where they draws upon seminal works in the domain to claim that mental models also encode beliefs, aspects of culture, and social protocol.

However, Gentner[21] makes clear in his description of mental models that they are incomplete and often imperfect constructs. Indeed, Simon's[4] notion of bounded rationality implies that human mental models must be limited due, at the very least, to imperfect information and the bounds of mental capability. Carroll[23] goes even further to include a range of dynamics that epitomize systems, such as feedback, delay, and nonlinearity, as additional common failure modes of mental models. Because of these limitations, a range of research has looked at how, as new information and experience is gained, the process of mental model revision occurs. Seel[22] provides a framework (Figure 2-1) for thinking about the revision process and the interplay between schemas and mental models whereby mental model creation and revision occurs in response to the failure to assimilate and as part of the accommodation of new knowledge. Similarly, Buckley[24] presents a model of learning with mental models whereby their use for accomplishing tasks is a process of model validation.

This validation either reinforces the model, causes revision, change and/or expansion of the model, or forces the individual to discard the mental model and form a new one to take its place.

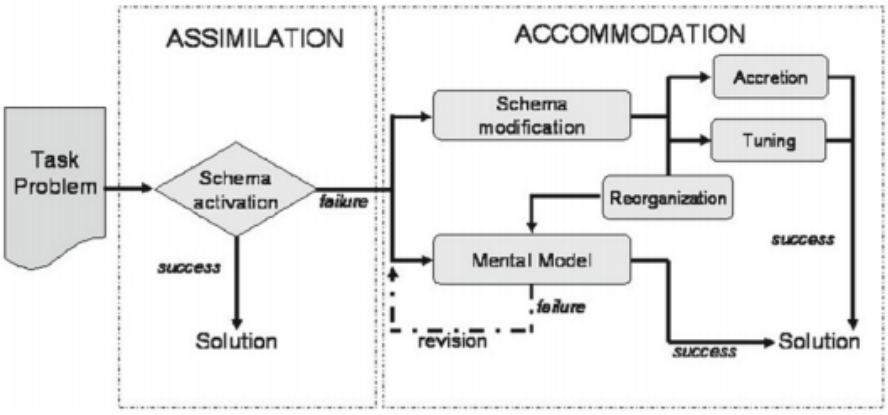


Figure 2-1: A model of mental model revision and schema modification through knowledge assimilation and accommodation. Reused from Springer Science+Business Media under STM and PSP guidelines for academic uses. ©2014 Seel[22]

Beyond just their general application by humans in learning and reasoning about the world, mental models play a key role in teams and teamwork. As Seel[19] states "within any given domain of activity, the richness and flexibility of learners' mental models influences the quality of the task performance." However, as mentioned previously, the outcomes of teamwork are often greater than the total output of the team. By extension, if mental models drive individual performance, some dynamics between team and mental models must exist as well.

2.3 Shared Mental Models

The characterization and formalization of mental models as the way in which an individual represents knowledge about physical and/or social phenomena is of importance in understanding how individuals reason about problems and complex systems. Yet a question exists about whether a similar but parallel construct is necessary when considering collective intelligence and group decision making processes.

In the work of Allard-Poesi[25], it was suggested that the cognitive approach for

conceptualizing an emergent, shared representation of an organization be replaced with a new socio-cognitive framework instead. Allard-Poesi framed the "cognitive paradigm" as an extension of the, appropriate and useful, application of cognitive psychology at the level of individuals to the level of groups and organizations. In particular, that organizational actions are enabled by the collective sharing of ideas or beliefs that exist across, and persist beyond, the members of the organization at a particular point in time; this collective mental state emerges from the interactions and common experiences occurring within the organization. Additionally, in this framework, the individual representations of knowledge are expected to align through the shared experiences of the members of the organization as well as the social forces experienced by the members during repeated exposure to each others individual representations. While seemingly logical, Allard-Poesi pointed out an array of drawbacks limiting this approach, including the comparative lack of definition for a collective representation relative to an equivalent representation at an individual level, the multitude of biases leading from the quantification of collective representations through aggregation of individual representation measurements, the assumed isomorphic equivalence in structure between collective and individual levels of representation, and the lack of adequately robust mechanisms that explain the interpersonal evolution of individual representations and the emergent development of the collective representation from such evolving individual ones.

To address these shortcomings in a purely cognitive approach to collective representation within organization, Allard-Poesi proposes a new perspective in which the frame of reference, with respect to study, shifts from the individual or organization to that of interaction of the individual within an organization. The paradigm of "social cognition" then, is framed by conceptualizing representations, both collective and, in some cases, individual, as the result of processes that occur during interpersonal exchange of ideas, beliefs, and knowledge and the social interactions between those involved. As Allard-Poesi puts it, "cognitive phenomena cannot be reduced to intra-individual processes, and inversely, interactions are influenced by the representations held by group members"[25]. From this, collective representations then become more

than a manifestation of a phenomenon but the result of on-going interactions between and within the members of an organization. These members are thus influenced by, and influencers of, the cognitive conflicts that arise from divergent perspectives and the social dynamics of negotiation that resolve them.

In order to explain the emergent collective representations that result from establishment of consensus within a group, the socio-cognitive framework provides three models of interaction. These models are representative of behavioral patterns arising from the contextual factors of the group, and the resultant effects on the group members individual representations (social response vs private response). First is that of "conformity as conflict control or rejection" whereby the contextual factors of negotiation lead to the adoption of a consensual collective representation without regard to the private acceptance of that representation among the members of the group that held a minority perspective on the task. In this case, the individual representations of the minority may remain intact at the private level but members show conformity at the social level, or, the individual representations may convert to become more aligned with the expressed collective representation of the majority. Second, the "normalization as conflict-avoidance" model represents that case where conflict arising from differentiated perspectives within the group is avoided through self-censure resulting in a collective representation equivalent to the "average" or "lowest common denominator" from among the individual representations. Here, it is assumed that conversion of individual, private representations is unlikely due to the nature of the contextual factors that lead to this model of behavior. Finally, the third model of interaction, "polarization as conflict-creation and resolution", corresponds with true collaboration within the group that results in a collective representation well aligned with the individual representations of the group members, but far from an "average" representation that would result under normalization. For this model, the resulting collective and individual representations will be well aligned between members, as will be their social and private responses. Yet these representations will likely be very different than their starting structures.

While Allard-Poesi provides a useful framework for understanding the social dy-

namics that may influence the expression and formation of shared mental models, the practical question remains — what good do shared mental models do? Salas et al.[26] provides some insight, claiming that it is the construction of shared mental models that, in part, enable high performance teamwork. It is through the use of mental models of not just the task, but also the *team itself* that arises the adaptability, coordination, anticipation of team needs, and the ability to infer the team's state which are necessary to achieve high levels of performance. In light of this, significant research has gone into seeking to detect and measure the influence of such shared constructs on team performance.

2.4 Instrumented Teamwork Experiments

In the late 1980s and early 1990s Bowers et al.[14] was among those who explored the use of low-fidelity models in teamwork research. This work looked at coordination and performance on small teams of two, using commercially available military helicopter flight simulation software to complete "missions". In this work, performance measurement was achieved through video recording and manual rating during subsequent post-processing. Bowers's argument for this approach was two fold — first, they sought to provide a controllable, repeatable laboratory setting capable of producing more realistic workloads and demands for coordination than in traditional laboratory controlled teamwork experiments, and second, they wanted to do so at a lower cost than the more expensive full-mission, field simulations. By providing an environment that was operationally complex and rapidly changing, yet retained experimental controls and eliminated extraneous variables, they believed that the tasks would be more useful for testing underlying hypotheses and the results more applicable to the field.

Almost a decade later, in 2000, Mathieu et al.[15] utilized a similar approach to Bowers in order to explore the formation of shared mental models and their influence on team processes and performance. Seeking to assess the sharedness of mental models, their influence on team processes and thus performance, and their convergence

over time, Mathieu succeeded to empirically distinguish between task and team types of mental models. In doing so, it was found that, indeed, shared mental models of team correlated strongly with team performance while shared mental models of task did so only indirectly through team process. As in Bowers, Mathieu also cites the use of a "low-fidelity" (commercial flight) simulator as "dynamic, interactive, interdependent, and allow(ing) for objective measures of team performance"[15] that achieved a balance between complexity and simplicity of simulated task demands. This study acknowledged the limitations of the true generalizability of the findings to actual combat missions, but highlighted the control maintained over operational aspects considered relevant to teamwork processes and team performance. Such aspects included team member roles, shared goals and values, (inter-)dependence, and interaction. Additionally, Mathieu points out that their sample of participants, being undergraduate students having no pre- or post- experiment team relationship, would likely limit the applicability of findings directly to real world flight crews. Finally, Mathieu also utilized video recording of participant teams with post-process coding of team process, while team performance was measured through a scoring system in the simulation, and mental models were measured through survey of participant assessment of attribute relationships for both tasks and team.

2.4.1 Measurement of Tradespace Exploration

A subset of instrumented teamwork experiments that both builds upon and complements the work of Bowers[27], Mathieu[15], Stout[27], and others, focuses on the topic of tradespace exploration. A tradespace is a view of a system that is the product of the space of design choices available and their mapping, in combination, to the space of multi-dimensional performance criteria of that system[10]. The exploration of a tradespace is the use a system model, by one or more stakeholders, to evaluate such a mapping of design choices to performance in order to better understand the tradeoffs that exist between options and enable decision making[10][28].

Analysis of a system's tradespace can take several forms including set based design evaluation, statistical meta-modeling, and point based design evaluation[10][28][29].

It is in the latter of these approaches, point based design evaluation, that MIT's GTL has focused its research and begun to develop methods for use in instrumented teamwork experiments.

Fruehling and Moser[5] and Tan and Moser[6] both explored the use of instrumented workshops for project design. Fruehling used absolute measures of estimated project performance, cost and schedule, and measures of attention allocation, selection of elements within the visualized project, to investigate the effects of dependency awareness in project planning. By applying clustering methods common in bioinformatics to the project performance measures and measures of attention allocation enumerated during tradespace exploration, the relationship between the patterns of model use and performance outcomes under different treatments were studied. Tan on the other hand, focused on the team decision making process and the effect of its level of coherence on team performance. Tan used a relative measure of project performance, Pareto Rank, to order the tradespace exploration of teams participating in a project design workshop. Once ranked, several measures of the teams' sequences of enumerated designs were then evaluated for predictive power; these measures included the number of designs explored, number of "branches" in the exploration sequence, and the pattern of changes made between designs. In both of these studies, the research included not only the pattern of results within the tradespace but also the pattern of teamwork behavior that induced these results. This focus on both halves of the tradespace and the relationship between the team, its processes, and its emergent performance, are common elements in many of the other studies of tradespace exploration performed by MIT's GTL.

Both Pelergrin et al.[7] and Manandhar et al.[8] demonstrate these same facets of study in their application of tradespace exploration measurement in domains outside of project design. Pelergrin performed a series of quasi-experiments using an ABMS of maritime shipping of fuels, looking for evidence of indicators of learning events, framed as the detection of surprise. Pair-wise comparisons between the performance outcomes of the designs enumerated in each experiment were performed. However, the focus of the research was more on demonstrating methods for finding relation-

ships between the patterns in the teams' allocation of attention, to the design choices and system performance measures, and to indications of surprise in the teams' interactions. Manandhar took a different approach in their study of teams exploring innovation campus designs. A novel teamwork-research platform was developed to enabled the instrumentation of a team's use of a system model. A robust experiment examining the effects of strategy on team performance was conducted. The participant teams were instructed in the use of the research platform and given different treatments with regard to the timing of a team discussion on the strategy of their exploration of the tradespace. Similarly to Tan and Moser[6], teams were evaluated based on the Pareto Rank of their designs. However, Manandhar used an "objective" Pareto Rank based on the fully enumerated tradespace, compared to Tan's use of a "subjective" Pareto Rank based upon only the tradespace enumerated by participants. Once ranked, the influence of treatments was examined.

These examples of existing research on the measurement of tradespaces in instrumented small-team workshop experiments provide insight into several areas key to the field of study: (1) measurement of both the solution space and the problem space halves of tradespace exploration, and the social interactions of a team, allows for a more robust study of the dynamics of learning with a system model, (2) the measurement of the solution space is multifaceted, spanning attention allocation, design choices, pattern of interaction with the model, and exploration path, and (3) the measurement of the problem space is more straightforward but still takes multiple forms such as pair-wise distance comparisons, aggregation of absolute measurement, and both objective and subjective relative measures.

2.5 Summary

This chapter has laid out a roadmap of the theoretical models, frameworks, principles, and methods that guided this work. It has summarized key elements of teamwork research, including mental models for individuals and teams, and discussed instrumented teamwork experimentation and the measurement of tradespace exploration.

Chapter 3

Research Methodology

Given the motivations established in Chapter 1, and the context of the existing body of related research as outlined in Chapter 2, this chapter establishes the methodological framework that was used to explore the underlying research questions and formalized hypotheses of this work. The methods used continue and extend those previously explored within the GTL at MIT with a focus on instrumented teamwork, and fit more broadly within the space of social science teamwork research supported by computational simulation and gaming, in the vein of work done by Bowers[14], Mathieu[15], Stout[27], and Sterman[30]. The development of an experimental design is presented, guided by the research questions and hypotheses, and drawing upon insights and rationale derived from the literature on mental models, teamwork science, systems modeling, simulation and gaming, and instrumented teamwork experiments. The subsequent use of this design in the development of a model problem, and its application in a series of experiments, is then described.

3.1 Research Questions and Hypotheses

As stated in Chapter 1, the fundamental questions that this work asked were:

Given a team of stakeholders evaluation designs for a sociotechnical system using a systems model —

- **Can we detect and measure the diversity of mental models within a**

team?

- **Does diversity of stakeholder mental models within a team affect that team's performance?**
- **Can one detect a pattern of model exploration by a team that is indicative of performance resulting from the emergence of collective awareness of systemic effect?**
- **Does the emergence of collective awareness of systemic effect lead to greater post-system-evaluation alignment of a team's mental models?**

From these questions, the following hypotheses are set forth:

Hypothesis 1. *Teams having a greater diversity of mental models of system architecture among stakeholders are more likely to be high performing teams.*

Hypothesis 2. *High performing teams will exhibit a more diverse exploration of architectures within a systems model.*

Hypothesis 3. *A more diverse exploration of architectures within a systems model will result in greater alignment among stakeholder mental models.*

With the focus of this work being teams and teamwork, it was necessary to define the use of the word "team" before a methodological framework could be laid out. Thus, for the purposes of this work, and as guided by the existing teamwork-research literature mentioned in Section 2.1, the following definition was established:

Definition 1: team - A set of two or more individuals, having specialized skills or knowledge, that depend upon the actions and abilities of one another other in order to achieve one or more shared goals.

3.2 Related Methods

3.2.1 Earlier Work

In the examples of instrumented teamwork experiments discussed in Section 2.4, a number of objectives and limitations relevant to this work appear. First and foremost, the fidelity of a simulation used in evaluating teamwork is not the driver of efficacy in teamwork research. Rather, efficacy results from the ability to provide and maintain experimental control in complex, multifaceted environments. This experimental control enables researchers to test their specific hypotheses more effectively and provide results that are, ideally, more translatable to real-world working environments. However, generalizability of results need not be the goal of such teamwork research either. Instead, the ability to construct the theoretically meaningful structures of teamwork in order to be able to adequately study their drivers and influence has real value.

A platform used for experimentation should then facilitate construction and measurement of the desired attributes of team and process necessary for testing the hypotheses of interest, while care must be taken in the research methodology to ensure that participants have the appropriate background (latently or via training) and motivation to properly engage such teamwork structures. Thus, low fidelity simulation environments need the ability to properly balance facilitating objective measurement, and the complexity of team and individual tasks, the latter being necessary for appropriate participant engagement. Additionally, the use of video recording software and manual review/coding as the sole, or primary, method for evaluation of teamwork is both labor intensive and potentially error prone, subject to the capabilities of the reviewer and rigor of the researchers.

This work argues that the methodology described within this section, in part as an extension of the work being done in the GTL at MIT, is an approach that simultaneously addresses these limitations while achieving objectives in line with previous studies of teamwork, measurement of mental models, and use of low-fidelity simulation and gaming.

3.2.2 Contemporary Work

The GTL at MIT is attempting to build the teamwork-research lab of the future through the development of methods and tools for unobtrusively instrumenting and analyzing teamwork in order to detect behaviors indicative of the phenomena that lead to high performing teams[5][6][7][8]. It is believed that this approach will ultimately enable the instrumentation of real world teams to allow leaders and managers to, in real time, detect patterns of teamwork. In doing so, the development of positive patterns that are likely to lead to higher performance can be fostered, while appropriate interventions can be formulated and applied to address patterns likely to inhibit teams from achieving improved, or even good, outcomes. Of particular relevance to this work are the earlier research efforts out of the GTL from Tan[6], Pelegrin[7], and Manandhar[8]. The work of Tan studied methods of analyzing the design choices and the solution space path of a team exploring project designs. Pelegrin focused on a rich set of instrumentation and successful implementation of a series of teamwork workshops for the detection of teamwork phenomenon. While Manandhar's research effort looked at the use of a teamwork research platform for system model simulation and detection of the systemic effects of strategy, as a team process, upon team outcomes. All three of these efforts developed approaches to, and protocols for, the execution of teamwork experiments using instrumented system model platforms and helped to focus and guide the thinking used in this work.

3.3 Experimental Design

In order to test the hypotheses previously stated, an experiment was designed having three treatment conditions and consisting of three parts: a pre-stimulus survey, the stimulus, a team-based role-play activity using a systems model, and a post-stimulus survey. All three parts of the experiment were designed to be conducted in an instrumented, multi-team workshop. However, given the state of public health limitations on in-person, social interaction in the face of COVID-19 during the time

of this research, the design was implemented using "virtual", multi-team workshops as described in Section 3.4.

3.3.1 Form & Function

Given the hypotheses stated above, the attributes of team and team process to be measured were identified as (i) diversity of mental models, (ii) team performance, and (iii) exploration of architectures, while the phenomena of interest were (i) emergence of collective awareness of systemic effect, and (ii) change in mental models. It was then necessary to define both the form and function of the elements of an experiment capable of unambiguously measuring these attributes and detecting the phenomena of interest.

Mental Model Diversity & Change in Mental Models

According to Gentner[21], "a mental model is a representation of some domain or situation that support understanding, reasoning and prediction," and mental models are based on qualitative, rather than quantitative, relationships that allow for such reasoning and prediction via mental simulation. Gentner also gives guidance on the study and representation of mental models, indicating that survey methods are commonly used, but likely insufficient if used alone. Additionally, Gentner notes that, due the use of qualitative relationships in the construction of mental models, ordinal scales are a good match for their measurement.

Given the above, in the context of sociotechnical system design evaluation using a systems model, the mental models in question are the stakeholders' mentally simulatable representations of the sociotechnical system. This should included the set of system attributes, operational characteristics, and constraints, both known and *assumed*. Mental simulation of the system by stakeholders could then be used to estimate the performance of the system under a particular configuration of attributes (system design) and such an estimate compared relative to another design using an ordinal evaluation scheme.

However, differences likely exist between two stakeholders in the underlying knowledge or assumptions that comprise the relationships used to build their mental models. This would lead to different performance outcomes, and thus evaluations, for the same system design when mentally simulated. Additionally, stakeholders, as members of a team, have one or more shared goals, by Definition 1. For this shared goal, it is likely, or may be established by design, that the stakeholders also share some common perspective regarding evaluation of this goal. Yet, stakeholders, as individuals, may also hold, or may have established by design, a latent set of perspectives (ex. role constraints, individual goals, etc.) that additionally influence their evaluation of the shared goal at any point in time[25]. Because of this, the subjective evaluation of the performance of a system, derived from simulation via a mental model, may differ between two stakeholders, even for equivalent mentally simulated performance outcomes.

Connecting these insights to the need to measure diversity of mental models and to detect changes in those models over time, a two-part survey was established as part of the experimental design. The first part of the survey was to be given before the team activity stimulus, and is therefore referred to as the pre-team-activity survey, while the second part, the post-team-activity survey, was given after the stimulus in order to produce repeated measurement data samples. The data collected aimed to represent mental model state in a way capable of capturing changes due to the stimulus.

Therefore, for the elicitation of mental model state, the use of two measures, desirability and feasibility, were selected for the surveys. In order to correspond to the two mechanisms previously described by which stakeholder mental models may lead to differing evaluations for a given design or performance outcome, the two measures were given the following meanings for this work:

Definition 2: desirability (of a design) - The degree to which a design meets a stakeholder's *personal* goals.

Definition 3: feasibility (of a design) - The degree to which a design meets the

shared goal(s) of a team.

It is argued that, for any given design, these two measures indicate the degree of alignment between two stakeholders' mental models both in the sharedness of their underlying knowledge structures (feasibility) and their polarization, as Allard-Poesi[25] would say, arising from their latent perspectives (desirability). This stems from the fact that, for a shared goal, the evaluation perspective of that goal should be shared by both stakeholders and only an underlying mismatch in knowledge should cause a difference in the assessment of feasibility. Conversely, the latent perspectives of the stakeholders may potentially be influenced by the shared goal and shared perspective for its evaluation.

The variation between stakeholders' evaluations of a design, when compared using these elicited measures of desirability and feasibility, is then a measure of the diversity of mental models among the stakeholders. Additionally, this work argues, for a given design, this approach allows for the measurement of similarity/dissimilarity between mental models, for a single stakeholder over time, within a stakeholder group at one or more points in time, and across different stakeholder groups at one or more points in time. However, it must be noted that neither of these measures actually capture the true correctness of the stakeholders' mental models relative to the performance outcomes of the modeled, or actual, sociotechnical system. Only the alignment of the stakeholders' mental models can be compared using the method suggested.

Given Gentner's assertion of mental models being built upon qualitative relationships and thus the fitness of ordinal scales for their representation, the surveys were designed to use individual Likert item questions. These questions asked participants to score the desirability and feasibility of various reference designs for the sociotechnical system to be evaluated by the team during the team activity.

Team Performance

A common approach to measuring the performance of a team is by considering the degree, in acceptability or proficiency, to which the team's *tasks* are completed with respect to the desired outcomes of the team's organization[12][31]. However, equating

team performance to task performance seems insufficient when considering the positive correlation of shared mental models on task performance[31], or when one considers, in addition to pure performance outcomes, the value of team viability[12] and the scociocognitive models where shared and individual mental models conflict[25]. With this expanded perspective, it then makes sense to extend team performance to be taken as a measure of both the team's task performance *and* the degree of alignment in the team's mental models. The latter aspect potentially being an indicator of the team's viability as well as indicative of the team's ability to have better performance on the same, or similar, tasks in the future.

Having previously established the measure of mental model alignment, the measure of task performance in the experimental design was set. The measure selected was the Pareto Rank of the system design indicated by the team as their preferred design after having explored the systems model.

Exploration and Emergence of Collective Awareness of Systemic Effect

Chapter 1 set out the context of this research to be an engineering team using a systems model to solve an engineering challenge. Given this framing for the research questions and hypotheses above, the remaining elements to be covered by the experimental design were measurement of the exploration of architectures and the detection of emergent collective awareness of systemic effect. These elements correspond to the primary activity of a team and the core phenomenon hypothesised to be influenced by the composition of the team and, in turn, mediate the performance of the team.

Chapter 1 posits that the purpose of modeling in Engineering is, essentially, to facilitate learning about a system at a team or organizational scale, if not at the individual scale. The use of simulations and games can be effective mechanisms for producing learning if they encourage and enable a cycle of active mental model revision through experimentation and reflection on outcomes[23]. Therefore, the measurement of exploration should seek to detect this cycle of mental model revision, and discriminate based on the level of experimentation being done by a team. This leads to the following definition of "exploration", adequate enough to then define a method

of measurement within the experimental design:

Definition 4: exploration - The act of a team using a computational model to experiment with attributes of that model in order to learn about the system the model represents.

Previous work[7][8] has sought for, and found, evidence of learning when examining the pattern of designs enumerated by a team and the performance of those designs on some set of system evaluation metrics. The enumeration of designs by a team gives insight into the design decisions, and choices made for those decisions, to which team has allocated attention relative to the overall *solution space*. The solution space being the set of all possible designs under the actual constraints of a system model's modifiable attributes and parameters (variables). Each design may then be represented as a *design vector*, with each element being one of the variables in the system model. The values of each element then represent the choice made for each variable from among the options possible, given the choices made in the other variables (active constraints). Similarly, each design enumerated by a team produces a calculated performance outcome for that design across the metrics of evaluation for the system model. This set of performance outcome values then makes up the *objective vector*, and the set of all such possible objective vectors produced from all of the design vectors within the solution space is the *problem space*. The performance outcomes for a given design provide insight into the information available to a team for evaluation, discussion, and revision of mental models before proceeding to enumerate another design. Combined, the two spaces just described make up the *tradespace* in which the user(s) of a system model must make trades in both the design choices (solution space) and expected performance outcomes of the system (problem space). The sequence of enumeration of the design vectors, and the resulting objective vectors, is known as the *design walk* of the team's exploration of the tradespace.

Therefore, the final, core aspect of the experimental design, the team activity and stimulus for affecting change within the mental models of the team, was set to be the enumeration of a subset of solution space design vectors and the subsequent

evaluation of the resulting subset of problem space objective vectors, for the purpose of identifying a "preferred design" that best delivered value to the stakeholders (team members). The identification and selection of a "preferred design" by the team was established as the team's shared goal for the experimental design, satisfying Definition 1. The design and objective vectors were to be recorded during the team activity, unobtrusively so as not to disrupt from the team process of design enumeration itself. The vectors were then to be post-processed in order to measure the team's exploration and detect any emergence of collective awareness.

It was decided that the measurement of exploration was to then be quantified by both the number of total unique design vectors enumerated, as well as the information content of changes made to design vector elements over the design walk. For the latter, greater diversity of element values having a more even distribution were weighted to indicate greater exploration. The detection of the emergence of collective awareness of systemic effect was then defined based upon the measurement of exploration.

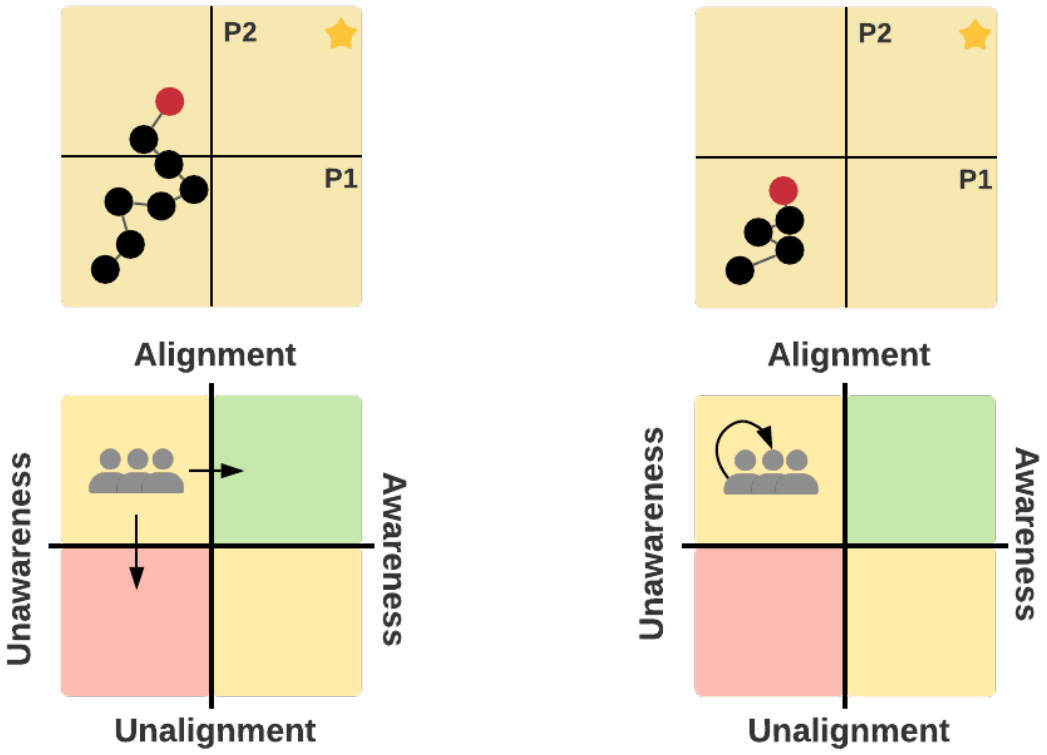
As a team explores a subset of the solution space, the regions of the problem space within which the resulting enumerated objective vectors lie, will vary. Awareness of systemic effect would indicate that the team understands the drivers of the modeled system's performance outcomes. This, in the context of the team's shared goal, would in turn allow them to consistently produce unique designs with performance outcomes that are objectively desirable and feasible. Additionally, in the case where a particular design resulted in a significantly worse performance outcome than previous designs, the team could quickly return to designs with desirable and feasible performance outcomes without reiterating a previous design. The design walk would have consistency and intent. On the other hand, a team without awareness of systemic effect is likely to produce a design walk that is more random in nature with respect to the observed performance outcomes, both in general and between sequential designs, due to the lack of understanding in the underlying mechanisms that govern performance. Consequently, under this work's hypotheses, when a team without awareness of systemic effect begins exploration of a tradespace, the team's pattern of exploration, and thus performance, are mediated by the team's mental

model alignment (Figure 3-1, Figure 3-2).

Given the use of the objective vectors for the evaluation of designs by a team, the Pareto Rank of the designs were taken as the objective measure of systemic effect such that awareness would lead to consistently similar, high-value, or constantly improving, Pareto Rank over the exploration.

Use of Role-play

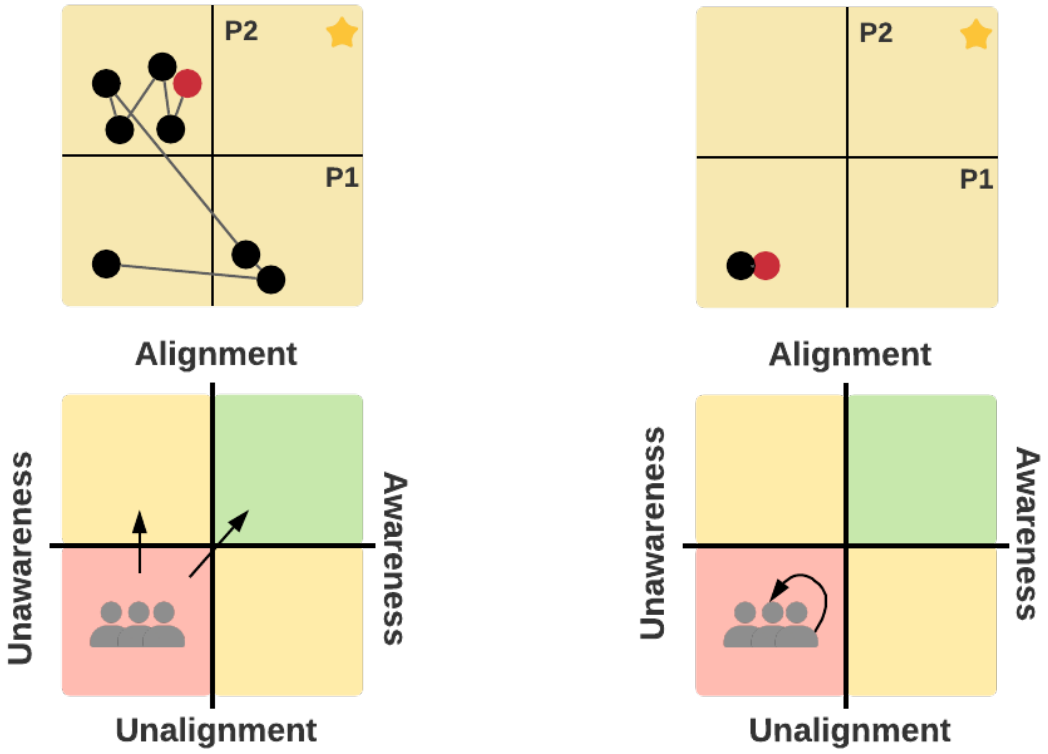
As highlighted in subsection 3.2.1, establishing a context for the participants and providing them with sufficient motivation to realistically engage with the structures



(a) Incremental exploration.

(b) Premature consensus.

Figure 3-1: Hypothesized, notional effects of exploration diagrams — aligned mental models. Teams with aligned mental models may exhibit only incremental tradespace exploration or reach a premature consensus during exploration. Incremental exploration may lead to a gaining of awareness or cause a loss of alignment. Premature consensus may limit exploration and prevent a gaining of awareness and a loss of alignment. (Teams starting unaligned and/or aware not shown.)



(a) Architectural exploration.

(b) Dysfunction.

Figure 3-2: Hypothesized, notional effects of exploration diagrams — unaligned mental models. Teams with unaligned mental models may exhibit architectural tradespace exploration or experience dysfunction during exploration. Architectural exploration may lead to a gaining alignment only, or, a gaining alignment and awareness. Dysfunction may limit exploration and prevent a gaining of alignment and/or awareness. (Teams starting aligned and/or aware not shown.)

and conventions of a team has been cited as a limitation to the generalizability for previous research. The use of role-play has been previously demonstrated as an effective method for supporting learning by supplementing acquisition and use of knowledge through the promotion of affective engagement[32] (Figure 3-3). In particular, Rooney-varga et al.[33] states about the use of role-play with a particular computational simulation, "...it creates an immersive, social learning experience..." and "...a risk- and cost-free environment in which participants learn from iteration, without accumulating the repercussions of failed decisions." These are the same conditions this work seeks to create in order to provide an environment similar to real-world model based engineering teamwork. As such, the experimental design included a role-playing scenario to be developed and used to contextualize the team activity and participant motivations.

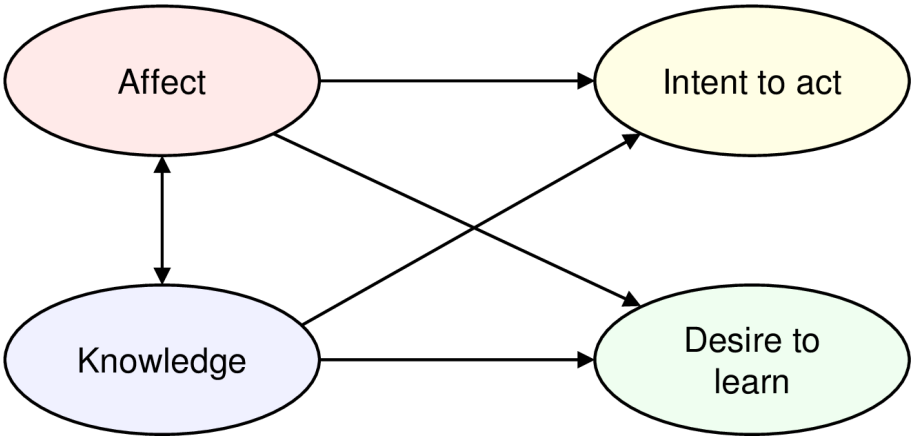


Figure 3-3: A theoretical model of learning under role-play and simulation. Reused under the Creative Commons Attribution License, ©2018 Rooney-Varga et al.[32]

3.4 Application of Experimental Design

In order to operationalize the experimental design presented in the previous section, a model problem was developed that could be used in an instrumented, multi-team workshop. Insights from the MIT's GTL previous work, as well as in the small group experiments literature more broadly, were used to inform the development of the model problem and workshops. As previously mentioned, due to exogenous

constraints imposed by COVID-19, these workshops ended up being designed to be run "virtually" using an internet based video conferencing system such as Zoom. A series of three workshop events were then run with this model problem, using Zoom, across three distinct and non-overlapping populations of participants.

3.4.1 Workshop Structure

The overarching event for each experiment was a two to two-and-a-half hour, multi-team workshop. A single model problem was developed for all three experimental runs of the workshop — a site design exploration for Star City, a conceptual city on Mars. The prototypical workshop began by introducing the scenario and background for the event, followed by a walk-through of the software to be used during the main team activity, team formation during which the pre-team-activity survey was administered, then, between one hour and one hour and twenty minutes of the instrumented team activity, and finally, ended with a short debrief and administration of the post-team-activity survey.

Scenario and Background

In order to support the role-play aspect of the experimental design, a scenario was created and shared at the beginning of each workshop. The scenario was meant to establish a standardized background and common set of motivations across all participants and to contextualize the team activity in relation to the model problem (i.e. why teams were exploring Star City site designs).

The scenario (Appendix A) oriented the participants to a higher order, shared goal at the project level of a fictional organization. This was achieved by framing the teams' work as being part of a company wide astronautic mission being run by the worlds largest, fictional, aerospace company, and setting their team goal to be choosing a site design (system architecture) for Star City's first settlement that met mission needs. The mission needs were not made explicit beyond the need to find a design that would lead to a successful colony. Rather, mission needs were implicitly

shared through the goals and value functions embedded in the stakeholder roles each member of the team would role-play. Suspension of disbelief was achieved by establishing the imagined scenario 20 years in the future and clearly emphasising that no new technological development or validation was needed for mission success. Knowledge related to the shared goal was provided by a short, supplementary background on the Star City concept. The Star City concept[34] was the winner of the Mars Society's 2019 Mars Colony Design Competition and materials from that competition's design pitch were included, with the authors' permission, as part of the workshop background[35].

In addition to establishing the setting for the role-play, the team level details of the scenario were also provided. This included a description of the size and composition of the team, and a statement of their team level task. Teams consisted of a stakeholder from each of the following three roles: Mission Health Engineer, Mission Power Budget Engineer, and Mission Performance Engineer. Details of the stakeholder roles were not available during this portion of the workshop but providing during Team Formation. The team level task was stated as follows: to "use the provided systems model to explore potential architectures and select your preferred design" with "preference" being decomposed into the two aspects of evaluation previously discussed, desirability and feasibility, and these terms defined.

A summary of the model problem and demonstration of the workshop software was then presented to the participants. This included a high level description of the systems model (Figure A-2) with which the teams would be working during the team activity, a description of the design choices available to them, a summary of key elements of the model simulation (Figure A-3), and a summary of the system performance evaluation metrics available to the team. Presentation of this background material provided a natural segue into the workshop software demonstration that followed. The design choices and evaluation metrics defined as part of the model problem are discussed in further detail in this section while the systems model and key elements of the model simulation are discussed further in Section 3.5.

Model Problem — Star City Site Design

The model problem selected for the team activity was the exploration of site designs using a systems model of Star City. Lordos & Lordos[34] conceptualized Star City as a set of five distinct human population "villages", built within the rim of a Martian crater. Each village is made up of a series of tunnels used as living, working, and recreating spaces, as well as connecting tunnels within and between adjacent villages in the crater rim, and a connection to a "central hub" built in the center of the crater. Given the scale and complexity of the full Star City concept, the model problem was scaled down to focus on an interesting set of design decisions and emergent performance outcomes within the scope of a single village. Three types of design decisions (Table 3.1) were selected to establish the solution space of the model problem. Section 3.5 includes details of the systems model developed to represent and simulate the model problem described here.

The first type of design decision selected was related to the population of Martian settlers inhabiting Star City. Due to the unique approach envisioned for Star City in composing the skill sets of the human settlers, as well as the space constraints inherent in the establishment of a habitation within the rim of a Martian crater, the set of decisions was the allocation of five specific sub-populations of settlers (farmers, industrial workers, life support workers, healthcare workers, and educators) to a fixed number of residences. The sub-populations represented a subset of the population responsible for five of the primary professional occupations defined for the Star City concept, each was also assigned a relevant secondary (personal) occupation[34].

Each sub-population's occupations correspond with a particular activity type (Table 3.2) that must be present within the village in order for settlers with that occupation to engage in work. Therefore, a set of activity-type-to-space allocation decisions, for a subset of three activity types, was selected as the second type of design decision. These decisions take advantage of the unique structure of Star City settlers having primary and secondary occupations and the temporospatial dynamics caused by the overlap in the defined activity types for the selected sub-populations.

Design Decision Group	Decision	Options
Population Allocation	Farmers	Residence 4A, Residence 4B, Residence 4C, Residence 4D, Residence 5
	Industrial Workers	Residence 4A, Residence 4B, Residence 4C, Residence 4D, Residence 5
	Life Support Workers	Residence 4A, Residence 4B, Residence 4C, Residence 4D, Residence 5
	Healthcare Workers	Residence 4A, Residence 4B, Residence 4C, Residence 4D, Residence 5
	Educators	Residence 4A, Residence 4B, Residence 4C, Residence 4D, Residence 5
Connection Mode Allocation	Level 1	Walk, Moving Sidewalk, Scooter
	Level 2	Walk, Moving Sidewalk, Scooter
	Level 3	Walk, Moving Sidewalk, Scooter
	Level 4	Walk, Moving Sidewalk, Scooter
	Level 5	Walk, Moving Sidewalk, Scooter
Functional Space Activity Allocation	Level 1	Educational, Healthcare, Cultural
	Level 2	Educational, Healthcare, Cultural
	Level 3	Educational, Healthcare, Cultural
	Level 4	Educational, Healthcare, Cultural
	Level 5	Educational, Healthcare, Cultural

Table 3.1: Design decisions available in the model problem. Options within each decision were mutually exclusive, otherwise, no constraints existed for any given decision or between decisions.

Sub-population	Size (people)	Primary Occupation	Secondary Occupation
Farmers	75	Agricultural	Cultural
Industrial Workers	50	Industrial	Educational
Life Support Workers	25	Industrial	Healthcare
Healthcare Workers	25	Healthcare	Cultural
Educators	25	Educational	Public Engagement

Table 3.2: The model problem's activity types for the selected Star City sub-populations' primary and secondary occupations.

Finally, given the relative distances settlers would be moving through each village, and Star City in general, the limited space and energy available in a Martian colony, and the need for efficient use of scarce human time resources, a set of mobility decisions were included as the third type of design decision available to the teams. These choices were unconnected to aspects of the sub-populations or activity types and so they provided an interesting and meaningful alternative axis of design options for the model problem.

In addition to these design decisions, the metrics that would be used by a team for evaluating its site designs needed to be defined. Seven metrics were chosen to provide a rich set of perspectives for evaluation: transit time, on-site time, diversity, energy use, occupancy, interaction, and utilization. The definitions of these metrics (Table 3.3) were provided to the workshop participants as part of the Scenario and Background. However, it should be noted that, by their definition alone, the evaluation metrics offer no insight into the subjective value they are assigned and how they were to be perceived by the stakeholders and teams; this is discussed further in Team Formation.

The seven evaluation metrics, in combination with the set of design decisions across population, activity, and transportation, make up the tradespace in which the teams worked.

Site Design Evaluation Metric	Definition
Transit Time	The amount of time people spend moving through the site as opposed to being engaged in a functional space activity.
On-site Time	The amount of time people spend, per-day, moving through the site and engaging in activities.
Diversity	The relative measure of population diversity within spaces on the site.
Energy Use	The amount of energy being consumed by people moving through the site and engaging in activities.
Occupancy	The number of people present within the site, or one of its spaces or connections, at a given time.
Interaction	The measure of the likelihood two people within a space will serendipitously "interact" due to their proximity to each other.
Utilization	The percentage of spaces within the site with people actively engaging in a functional space activity at a given time.

Table 3.3: Definitions provided to workshop participants for the seven site design evaluation metrics

Software Walk-through

Once the role-play scenario and background information had been shared with the workshop participants, a demonstration was provided for the systems modeling and simulation software the teams would be using. This walk-through lasted roughly 20 minutes and served as the formal training to familiarize participants with the user interface and use of the software. A specific "baseline scenario" site design was shown and simulated, which corresponded to the software's default values for design decisions. This approach was similar in intent to the training procedure described by Maitheu et al.[15] — to establish and standardize across all participants a common baseline of background experience for the team task and overall process. However, it did not attempt to focus on individual tasks or specific team processes, nor did it follow any specific team training guidelines.

Workshop participants were shown how to launch the software, load the systems model if necessary, use of the design decisions and simulation results interfaces, and how to run a simulation. Further details of these aspects of the software are discussed in Section 3.5.

During the walk-through, workshop participants were not instructed to follow along or practice in any way. Indeed, as the software prompted users to enter a team number upon launching, participants were limited in their ability to use the software having not yet been assigned to a team. This was an intentional choice meant to limit premature exploration of the systems model tradespace that could result in learning or awareness of systemic effect, and therefore influence the formation of participant mental models, before the application of the designed team treatments and team activity.

Team Formation

During all three workshops, attendees were randomly grouped into teams of three. Due to the variable nature of participant count among experiments, and in some cases, during a given experiment, some teams ended up with either four team members or

with only two. These teams were identified during the workshop for proper handing of results in later analysis. The teams were then divided into one of three treatment groups and workshop materials were provided. The workshop materials consisted of treatment specific instructions, descriptions of four canonical site designs, and stakeholder profiles for each of three roles: Mission Health Engineer, Mission Power Budget Engineer, Mission Performance Engineer.

The design of the three stakeholder profiles created separate, distinct sets of goals and value functions for each of the team members (Table 3.4) that was meant to result in a dispersion of mental models across the four canonical designs when assessed on desirability and feasibility. For example, the Power Budget Engineer was instructed to focus on reducing the power consumption of the site, and given knowledge of the power consumption of the three choices of transportation mode available. Thus, this stakeholder was likely to find Canonical Design 1 to be the most desirable, where walking, the least power intensive mode of transportation, was used exclusively. On the other hand, the Mission Performance Engineer was instructed to focus on the amount of time the population was engaged in activity as opposed to simply moving from one space to another. Thus, this role was more inclined to find Canonical Design 3 as the most desirable, as the sub-populations were allocated to separate residences and as close as possible to their primary occupation activities, and scooters were the sole transportation mode. The Mission Performance Engineer role had no knowledge of transportation mode power consumption, nor any objective related to site power use. An example stakeholder profile and an example canonical site design can be found in Appendix A. Through the use of the stakeholder roles in the role-play scenario, and by creating disparity in their goals and value functions, the intention was to shape the subjective evaluation and perception of the model problem's design decisions and evaluation metrics, and thus participant mental models of the system. Then, through application of different experimental treatments, the effects of varying degrees of mental model alignment were to be measured.

The teams within each workshop were divided among the following treatment groups as evenly as possible, and favoring groups A and B when necessary.

			Value Function		
Stakeholder	**Primary Goal(s)**	**Secondary Goal(s)**	**High**	**Acceptable**	**Minimal**
Mission Health Engineer	Ensure social engagement of sub-populations.	Ensure active lifestyles, Ensure professional/personal fulfillment	Diversity >65%, Interaction >3.5%	Diversity 55-65%, Interaction 3-3.5%	Diversity <55%, Interaction <3%
Mission Performance Engineer	Ensure productivity of the site.	Ensure efficient use of the site.	Transit Time <305min, Utilization >32%	Transit Time 305-310min, Utilization 31-32%	Transit Time >310min, Utilization <31%
Mission Power Budget Engineer	Ensure effective use of power resources.	*None.*	Energy Use <1,000 MJ/day	Energy Use 1,000-1,250 MJ/day	Energy Use >1,250 MJ/day

Table 3.4: Role-play scenario stakeholder profiles and their associated goals and values.

Group A - Aligned Mental Models

These teams were instructed to randomly assign each stakeholder role to a team member. Each individual was then to read the stakeholder profile for their assigned role (and assume the responsibilities and objectives of that role) before coming back and, *as a team*, reviewing and discussing the four canonical designs provided. An emphasis was to be placed on considering the desirability and feasibility of these designs. Once this discussion was finished, the team was then instructed to have each member individually complete the pre-team-activity survey.

The intention of this treatment group was to provide the team with a dedicated opportunity to have a discussion around goals and values in order to uncover biases, reveal differing objectives, and ultimately align mental models before survey and team activity completion.

Group B - Unaligned Mental Models

These teams were instructed to randomly assign each stakeholder role to a team member. Each individual was then to read the stakeholder profile for their assigned role (and assume the responsibilities and objectives of that role) before *individually* reviewing the four canonical designs provided and assessing each design on its desirability and feasibility. Once an individual had completed his/her review of the designs, he/she was to complete the pre-team-activity survey.

The intention of this treatment group was to provide the same set of background information to the participants as in treatment group A, but without the opportunity to have a specific period in which to align around the desirability and feasibility of the canonical designs. This was meant to prevent the team from having a discussion around goals and value before exploration began. It was also meant to allow for the independent establishment of each individual's mental model of the system, based upon their assigned role's goals, value functions, and with limited knowledge, and thus assumed constraints and limitations.

Group C - Control

For these teams, no assignment of stakeholder role was instructed, although the profiles were provided in the materials. Rather, each team member was to *individually* review the four canonical designs provided, assess each design on its desirability and feasibility and then complete the workshop pre-team-activity survey.

The intention of this treatment group was to allow each individual team member to establish their mental model of the system independently, but without any specific goals, value functions, or knowledge from an assigned stakeholder role. While this may result in pre-existing knowledge or biases influencing their assessments of the desirability and feasibility of the canonical designs, given that it was without stakeholder profile assignment, that would allow for comparison to group B teams. Group B teams were likely to have similar pre-existing knowledge or biases. Such a comparison was meant to elucidate whether or not the stakeholder profiles indeed helped to shape the mental models of teams in groups A and B.

In the case of teams in treatment groups A or B with two or four members, the teams were instructed to randomly assign one team member with an additional role, or to have an additional team member assigned to one of the roles, respectively. The stakeholder profiles provided to all three treatment groups were the same for each role. The set of four canonical designs provided were also the same across all three treatment groups. In order to encourage a more realistic stakeholder engagement process during the teamwork activity, the teams in groups A and B were also instructed not to directly share their stakeholder profiles, but rather to engage in role-play during group discussions in order to share knowledge and articulate their role's goals and value functions.

Once all of the individuals on a team had completed their pre-team-activity survey, the teams were instructed to move on to the team activity.

Team Activity

The main portion of the workshop, and the experiment, was the use of the systems model by the team to explore the tradespace of site designs given the design decisions available in the software. During this time, the teams worked independently with no further instruction beyond any technical support necessary for completing the team activity and clarifications on the provided workshop materials.

Exploration consisted of making one or more changes to the option values for the design decisions presented in the software. Once a set of choices was made, these values would become the design vector and be used in the systems model by the simulator to produce the objective vector of evaluation metric values. The software then allowed these values to be plotted against one another. Further details regarding these features of the software are discussed in Section 3.5. The teams were instructed to discuss each simulated site design in terms of its desirability and feasibility and to record any notes regarding that design's results for later reference.

Finally, the teams were instructed to spend the final five minutes of the team activity period discussing the set of site designs that they had simulated and to select their team's preferred site design. This design, identified by simulation number, was to be noted down for inclusion in their individual post-team-activity survey responses.

Debrief

The final segment of the workshop consisted of a short debrief in which comments, feedback, and questions were fielded. All participants were instructed to complete the post-team-activity survey and to submit an export of their simulation results. No discussion of the details of the team activity, including discussion of individual results, simulation behaviors noted, patterns of results, team processes, approach to exploration, or differences between team instructions occurred until all participants had completed the survey. Upon completion of the debrief the workshop was ended and all participants excused, this concluded the experiment.

3.4.2 Participant Surveys

Each workshop participant was instructed to complete two surveys. The first, the pre-team-activity survey, was to be completed after having received the scenario and background information that contextualized and explained the workshop's main team activity, but before having done any exploration of the systems model tradespace. While the second survey, the post-team-activity survey, was to be completed immediately upon the conclusion of the team activity portion of the workshop, after having explored some amount of the systems model tradespace. Both surveys were to be completed individually by each participant and sought to assess the state of the taker's mental model of Star City by measuring their perception of desirability and feasibility for various site designs. This approach mirrors that used in Mathieu et al.[15] where mental models were measured by giving participants a matrix of system attributes and asking them to rank the influence of one system attribute on the other system attributes.

The surveys sought to establish a per-participant assessment of their mental models, both before and after their team activity, based on what attributes of Star City's site design were considered valuable (desirability) and important (feasible). This was achieved by asking each participant to independently rank the desirability and feasibility of the four canonical designs provided during team formation. Ranking was done on a scale of one to five with one being "very low" and five being "very high". It is important to note that these Likert item questions *were not* composed in such a way as to produce a true Likert scale. Rather, a single item question was used for each site design to be assessed. Additionally, free text input was solicited from each participant in response to asking for the system attributes that drove their assessment, both the most and the least, of desirability and feasibility. Finally, each participant was asked to rank order at least one, but up to three, of the systems model evaluation metrics as the most important in their assessment of desirability and feasibility. The order of questions was carefully chosen in order to flow logically but avoid influencing the response of the participant by providing them suggested system attributes before

evaluation of canonical designs or entering free text responses.

The post-team-activity survey mirrored the pre-team-activity survey but added an additional set of questions regarding the preferred site design selected by the participant's team. These added questions were ordered after the free text response questions and before the evaluation metric rank ordering questions. The additional questions asked participants to identify their team's selected preferred design (by simulation number), indicate whether they agree with the selection as being the most desirable and feasible, and to provide a free text response for the system attributes that contributed most to their *individual* evaluation of desirability and feasibility.

The surveys were anonymous but did ask each participant to identify their team, the role(s) that they had been assigned, and, whether they had previously worked with any other members of their randomly assigned team, or participated in a previous workshop.

3.5 Workshop Software

All workshop participants were provided with an installer for the software necessary to participate in the team activity portion of the workshop and encouraged to use the software to explore the systems model tradespace. However, teams were instructed to select a single team member to serve as the team's official simulator whom would input the team's desired design decision choices, label and run the simulation, and be responsible for submitting the team's simulation data at the end of the workshop. Thus, each team needed only one instance of the software to be run, by one team member, in order to complete the team activity. The software provided consisted of a customized "game" interface, the systems model, and an engine for Agent Based Modeling & Simulation (ABMS).

3.5.1 Systems Model

The systems model was a specialized representation of a single village from Lordos & Lordos's[34] Star City concept — a human settlement built inside the rim of a

Martian crater. The model was built using a novel modeling framework (Figure 3-4) developed as part of a collaborative research effort at MIT sponsored by East Japan Railway Company (JRE). The framework used was designed to facilitate the modeling of populations engaging in transportation and non-transportation activities within a complex site; therefore, the portion of the systems model that represents the complex site is referred to as a "site model" within this section, while the portion that represents the population is referred to as the "population model". This specialization of the modeling approach captures unique aspects of the model problem's architecture and allows for representing decisions well suited to exploring the principles underlying the Star City concept[36].

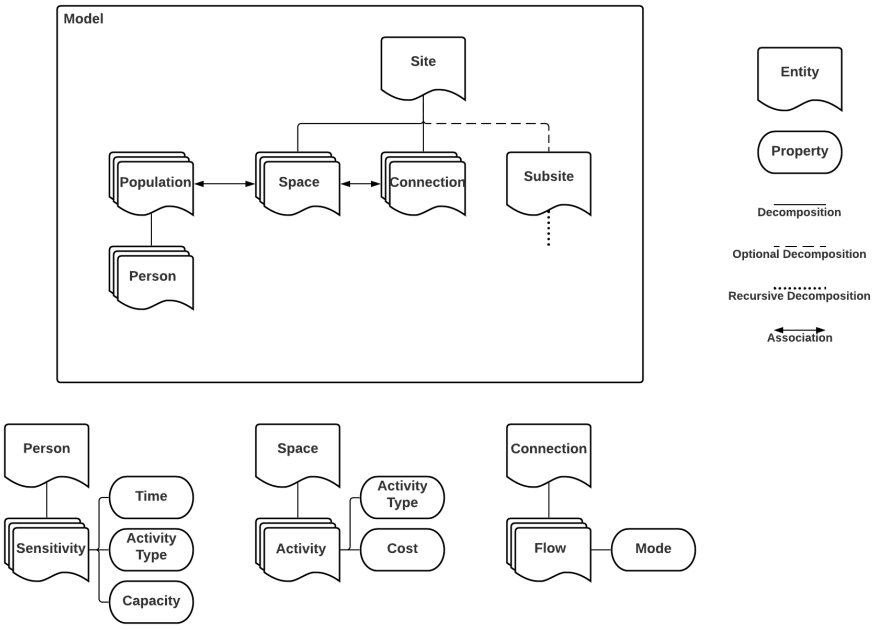


Figure 3-4: Simplified representation of systems modeling framework.

The population model consisted of five sub-populations, one for each of the represented demographics of Martian settlers. Each of the sub-populations was defined by the number of people which composed it, the distribution with which they would "wake up" at the beginning of a simulated day, and their temporal sensitivity to the activities available within the site (Table 3.5). The temporal sensitivity to activity type was used to define the behavioral structure related to primary and secondary

occupations in which individuals within each sub-population were most likely to engage during each day. These sensitivities were defined as the probability of selecting a given activity relative to all activity types to which the demographic was sensitive within a given temporal interval; these probabilities were also modeled as distributions. In addition to the probability of selection, each sensitivity defined a "budget" of time available for engaging in the given activity during a particular interval. This budget represented an internal behavioral constraint and was again defined using a distribution. For the Star City model problem, all demographic distributions were Gaussian of differing means and variances. All other meaningful properties of each demographic (i.e. walking speed) were fixed values across all sub-populations. No representation or visualization of the population model was available through the workshop software.

The site model consisted of five "levels" of spaces and connections. The spaces were sized and arranged to mimic a floor plan layout that would result from the use of tunnels partitioned by activity type. Each level had between one and three "tunnels", with each tunnel segmented into three spaces, and between two and four "community dome" spaces representing structures that would be on the external surface of the Martian crater rim. Every level also had a set of connections and junction spaces that facilitated movement between adjacent tunnels, and between the outermost tunnel(s) and the community domes. Vertically adjacent levels were joined via a single vertical "stairwell" connection between one of the central junction spaces in each level. (Figure 3-5)

Every space was a distinct model object, able to have its attributes, such as maximum occupancy and activity type, independently defined. Table 3.6 outlines the design of the attribute values of the site model's spaces. The circulation activity was necessary in each space in order to facilitate simulation of the movement of people within the spaces of the site, a further discussion of this can be found in subsection 3.5.3.

Interval		Sensitive Activity Types	Probability of Selection (Avg., Var.) [%]	Time Expenditure Budget (Avg., Var.) [min]
Early Morning	[0000h,0600h)	Residential	100, 0	30, 10
Morning	[0600h,0800h)	Residential	98, 2	60, 30
		Circulation	2, 1	30, 10
Workday	[0800h-1700h)	*Primary Occupation Activity*	72, 10	480, 30
		Public Engagement	9.6, 10	120,60
		Circulation	9.6, 5	30, 10
		Residential	5, 1	60, 60
		Educational	1.27 each or 1.9 each, 1	60, 30
		Healthcare		60, 30
		Cultural		60, 30
Evening	[1700h-2100h)	*Secondary Occupation Activity*	46, 20	180, 30
		Public Engagement	9 each or 18, 10	60, 30
		Circulation		60, 30
		Residential	27, 20	60, 30
		Educational	3 each or 4.5 each, 1	60, 30
		Healthcare		60, 30
		Cultural		60, 30
Late Evening	[2100h-0000h)	Offsite	70, 10	20, 0
		Residential	20, 5	20, 10
		Circulation	10, 2	20, 5

Table 3.5: Generalized, definitional sensitivity matrix for the Star City model problem sub-populations. "Primary Occupation Activity" and "Secondary Occupation Activity" represent the actual activity types of the primary and secondary occupations for each sub-population demographic. In the case where either the primary or secondary activity type was Educational, Healthcare, or Cultural, that activity type was not repeated in the same interval and the larger of the two percentages were assigned to each of those activity types remaining. In the case where the secondary activity type was Public Engagement, that activity type was not repeated in the same interval and the larger of the two percentages were assigned to Circulation.

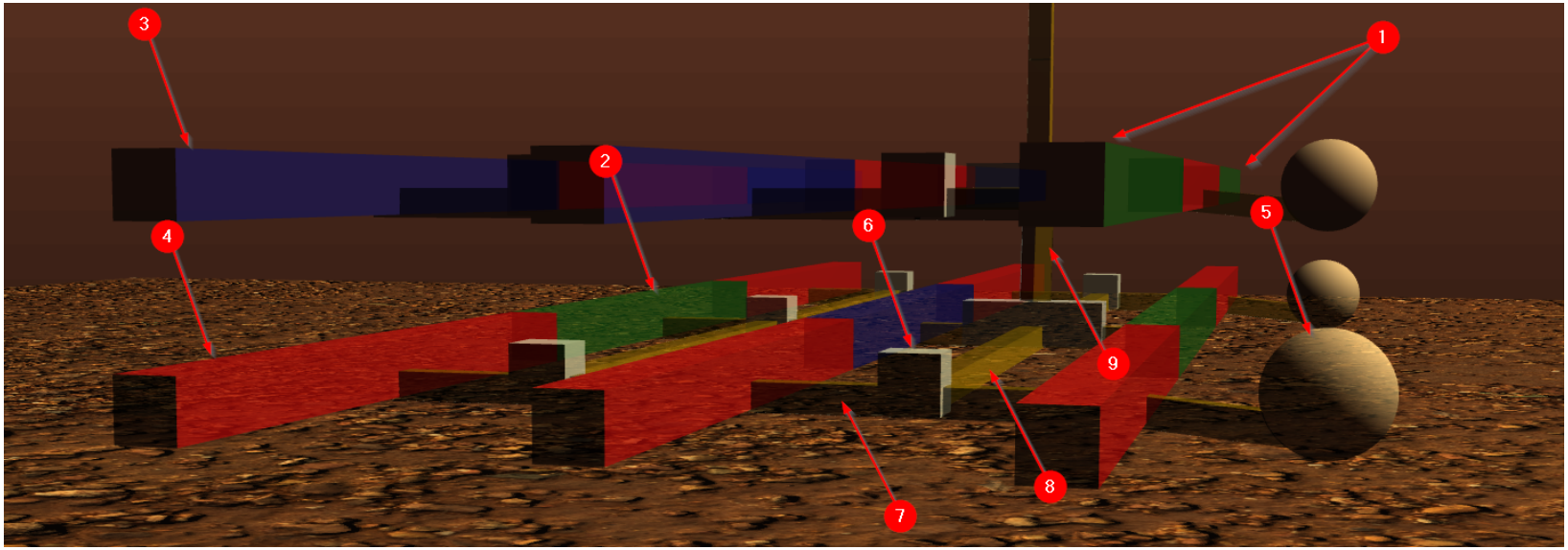


Figure 3-5: Annotated site model visualization. (1) tunnel, (2) space, Agricultural activity type, (3) space, Industrial activity type, (4) space, assignable activity type, (5) community dome, (6) space, junction, (7) connection, walking, (8) connection, assignable mode, and (9) vertical "stairwell" connection, walking. (Residential activity type spaces not shown.)

Model Object Type		**Maximum Occupancy Capacity**	**Activity Types**
Space	Tunnel	100 (200 for Residential Activity)	Circulation + $x \subset \{\text{Residential, Agricultural, Industrial, Educational, Healthcare, Cultural}\}, \lvert x \rvert = 1$
	Public Dome	100	Circulation + Public Engagement
	Junction	100	Circulation

Table 3.6: Space attributes used in the design of the site model.

While the modeling framework supports complex representations of a space that supports multiple activity types, the site model was designed to include just a single activity type, in addition to a circulation activity, for each non-junction space. Junction spaces had only a circulation activity. This choice was made primarily in response to the relatively short amount of time participants would have to work with and understand the model. The simpler representation reduced the cognitive complexity of the site model itself and simplified the dynamics in the simulated results to allow for easier access to awareness of systemic effect.

Similarly, each connection in the model was a distinct model object that could have its attributes, such maximum occupancy and supported mobility mode, independently defined. Table 3.7 outlines the design of the attribute values of the site model's connections.

Each of the modes was associated with a movement factor and an energy-cost-per-use value. The movement factor acted as a multiplier for the speed with which a person would travel when using that mode to transit a connection, while the energy-cost-per-use value acted as a multiplier for the energy required from the overall system for each individual transit of a given connection using that mode.

The site model was provided to the workshop participants as part of the packaged workshop software installer and users did not interact with the site model directly, but could interact with a visualization of the site model through the workshop software's "game" interface.

In addition to reviewing the site model during the Scenario and Background portion of the workshop, a job aid was also provided to the teams that included a more detailed, labeled, level-by-level view of the site model. Each stakeholder profile also contained information related to the systems model. The Mission Health Engineer was provided the sub-population sizes and their primary/secondary occupations, the Mission Performance Engineer was provided the sub-populations primary/secondary occupations and a partial, course granularity, generalization of their sensitivity schedule, while the Power Budget Engineer was provided the transportation mode speed multiplier and energy use.

Model Object Type		Maximum Occupancy Capacity	Supported Modes
Connection	Horizontal	100	$x \subset \{Walking, MovingSidewalk, Scooter\}, \lvert x \rvert = 1$
	Vertical	100	Walking

Table 3.7: Connection attributes used in the design of the site model.

Mode	Speed (Relative)	System Energy Cost (kJ/use)
Walking	1	0
Moving Walkway	1.53	129.2
Scooter	7.34	396

Table 3.8: Connection transportation modes used in the design of the site model, and their seep factor and energy use.

3.5.2 User "Game" Interface

The "game" interface (Figure 3-6) was the portion of the workshop software with which participants directly interacted, a labeled version can be found in Appendix A.

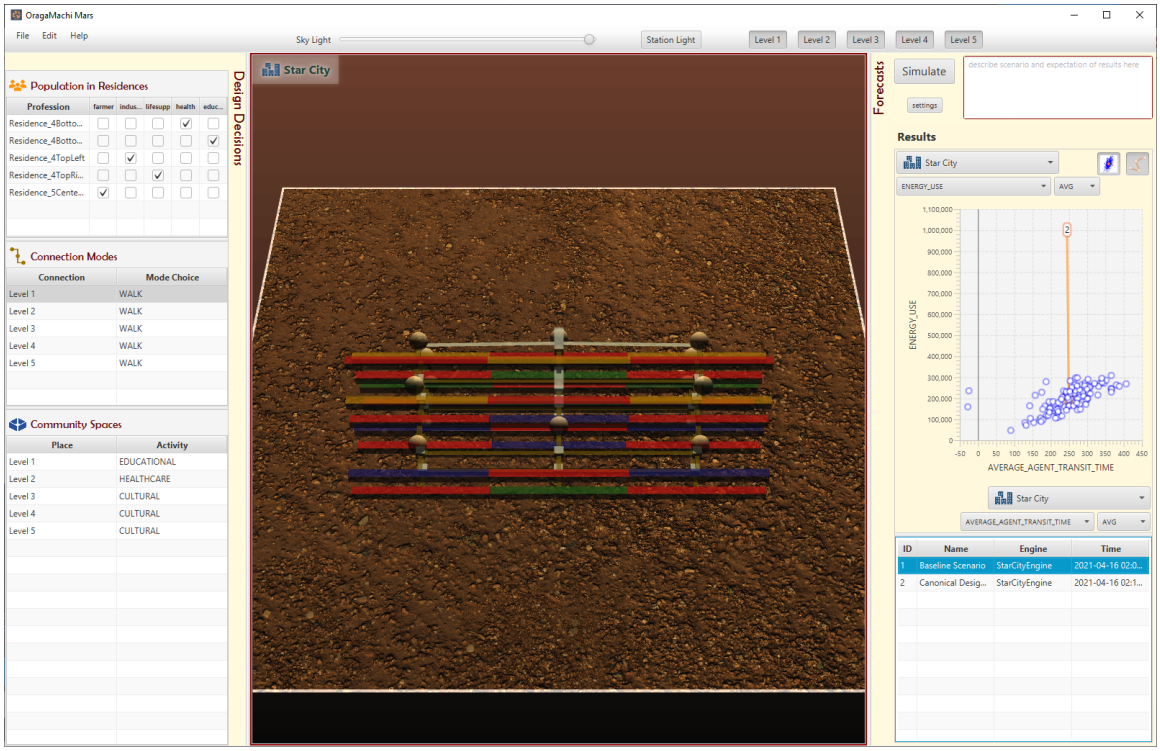


Figure 3-6: Workshop software user interface

The left hand pane of the window contained the systems model design decisions interface, the right hand pane the simulation results interface, and in the center was an interactive systems model visualization. The user was able to load a systems model and save the simulation results using the menu at the top of the window. However, for the version of the software provided to participants the Star City systems model was automatically loaded by default upon launching the software, and users were automatically prompted to save their simulation results when exiting the application. Users were not able to modify or manipulate the systems model in any way other than through the design decisions interface provided. Thus, the attributes set for the model objects (spaces, connections, etc.) in the systems model were fixed, except for those objects with attributes modifiable through via the design decisions interface.

Additionally, the "game" interface served as the mechanism by which users could execute a simulation of the systems model using the bundled ABMS engine.

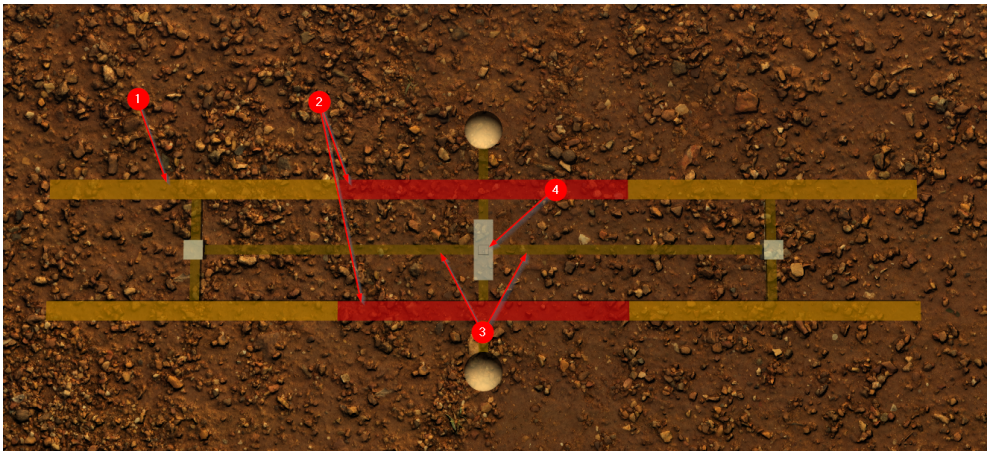
Systems Model Design Decisions

The systems model design decisions interface offered the user a customized set of design decisions, and their option values, for the loaded systems model. Changes made to the option values in the design decisions modified the configuration of the systems model, allowing for the simulation of different designs. For the Star City model, three types of design decisions were presented to the user: population allocation, per-level transportation mode selection, and per-level space activity selection.

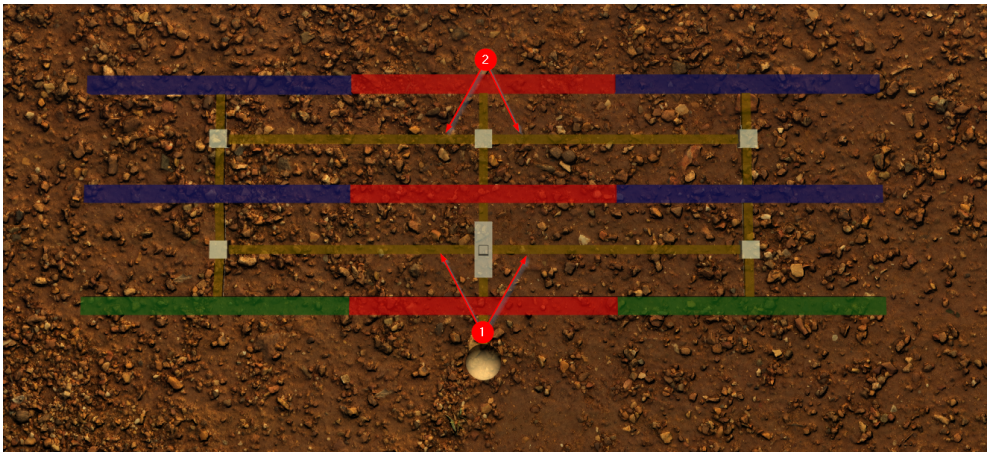
The population allocation decision allowed users to select in which spaces, from among the four spaces on the fourth level and one space on the fifth level having an activity type of Residential (Figure 3-7a), each of the five sub-populations would begin and end their simulated days. For each of the $N \in \{1, 2, 3, 4, 5\}$ spaces to which a sub-population was assigned, $\frac{1}{N}\%$ of that that sub-population would be allocated to each of those spaces at the start of each simulated day.

The transportation mode decision allowed users to select, for each level, which of the supported modes would be used by people transiting two of the horizontal connections on that level during a simulation. The two connections modified were those that ran parallel to the tunnels and out from the central junction space having the vertical "stairwell" connection (Figure 3-7b). All other horizontal connections and all vertical connections between levels were not modifiable and therefore had a fixed mode of "walking".

The space activity selection decisions allowed users to select, for each level, the non-circulation activity, from among Educational, Healthcare, or Cultural, assigned to a subset of the tunnel spaces on that level (Figure 3-7a). The subset of spaces that could be modified consisted of those spaces on each level that were assigned an activity type of Educational, Healthcare, or Cultural. Thus, for the entire systems model, all spaces assigned an activity type of Residential, Agricultural, or Industrial were fixed and could not be modified.



(a) Level 4. (1) space, Residential activity type, (2) space, assignable activity type, (3) connections, modifiable modes, and (4) vertical "stairwell" connection, walking mode . All other connects not annotated have a fixed moded of walking.



(b) Level 2. (1) connections, modifiable modes, and (2) connections, fixed modes (walking). All other connections not annotated have a fixed mode of walking.

Figure 3-7: Annotated site model level visualizations. (1) connections, modifiable modes, and (2) connections, fixed modes (walking). All other connections not annotated have a fixed mode of walking.

Simulation Results Interface

The simulation results interface provided the user with two separate functionalities — executing a simulation and viewing the evaluation metrics of a previously executed simulation.

To execute a simulation, the user would enter an optional text description identifying the design before executing the simulation. Executing the simulation would cause the systems model, as modified by the selections made in the design decisions interface, to be provided to the ABMS engine for execution. Upon completing its execution, the ABMS engine would return the results of the simulation which was then was presented to the user in a chronological list within the simulation results interface. The results returned from the ABMS consisted of a set of values for each evaluation metric, calculated for a subset of the model objects in the systems model.

Selecting one of the simulation results form the chronological list would allow the user to plot the values of the evaluation metrics returned for that result against each other using the tradespace plot in the simulation results interface. The tradespace plot allowed the user to select a systems model object and one of its corresponding evaluation metrics for both its X and Y axes. Once evaluation metrics had been selected, a point for each simulation that had been executed was displayed in the tradespace plot. The results from the ABMS consisted of an average value for each evaluation metric and the set of values, one per Monte Carlo run simulated by the ABMS engine, used to calculate those average values. The points plotted in the tradespace plot were the average values of the set of values for the selected metrics, for each simulation that had been execution. Additionally, using the tradespace plot, users could render a line connecting the plotted points in order of simulation (their design walk) as well as the cloud of points corresponding to the set of values for the selected metrics from the simulation selected in the chronological list.

Interactive Systems Model Visualization

Within the center of the "game" interface, the users were provided an interactive visualization of the systems model. The visualization allowed users to click on a systems model object (i.e. space, connection, etc.) to highlight it, and its name would also be shown in the upper left corner of the scene. The user could also rotate and zoom the visualization as well as hide and show each of the five levels within the model, and adjust the lighting of the scene. This provided users with a mechanism for inspecting the systems model in order to understand the visual representations of its objects and gain insight into their structure and physical relationship. Additionally, when users selected a level in either the transportation mode or space activity decisions within the systems model decision inputs interface, a representative entity on the selected level would become highlighted. This was included to allow users to form an intuitive understanding of which types of entities, on which levels, were being affected by each decision.

3.5.3 Agent-Based Modeling & Simulation

Underlying the user facing "game" interface of the workshop software was a novel ABMS engine. This simulator was developed in conjunction with the site modeling framework discussed in subsection 3.5.1 as a collaborative research effort at MIT sponsored by JRE. Unlike previous research done out of MIT's GTL which used either existing agent-based simulators[7] or custom software utilizing deterministic methods of simulating systems models[8], this work chose a stochastic, agent-based simulation approach using a novel modeling and simulation framework.

Generally, ABM approaches offer a method of uncovering complex, emergent behavior that arise from the dynamics of combining numerous, diverse, but simpler entities within a model[37]. This differs from a parametric approach which attempts to directly model the complex features of entities, or, mathematical expressions of dynamic systems that may capture high level causal structures[38]. Both of these approaches suffer from being top-down methods and often lack the ability to capture

the nuances of lower levels of behavior or heterogenous populations, especially ones that change over time. In both cases systemic emergence may occur that drives, or is driven by, complex, dynamic interactions between a minority sub-population and their environment[37]. Due to the nature of the assumptions made in non-agent based approaches, these types of interactions are often unable to be captured[39]. In the model problem selected, a set of five distinct sub-populations of Martian settlers are living and working together in a self-contained settlement where the physical and social well being of individuals, and the overall success of the settlement, are dependent upon the built environment and the interpersonal behaviors that result from it. Given the nature of this problem, an ABM approach was deemed appropriate.

The ABM was developed specifically to work with the modeling framework used to construct the Star City model problem. The duration of simulation could be varied, however, for the workshop experiments used in this work, all simulations were executed over 24 simulated hours and, due to the non-deterministic nature of ABMS, a Monte Carlo approach to generating results was taken. Each execution of the simulation resulted in 12 or 25 independent runs of the simulator (depending on the version being used in the workshop). The choice of these values was made due to a tradeoff existing between the robustness of simulation results and runtime, with the trade being made in favor of a shorter runtime. Here, the robustness of simulation refers to the convergence of the average value observed for an evaluation metric returned by the simulator when compared between two simulation executions using the same design vector. Because the evaluation metric calculations were non-deterministic, the smaller the number of Monte Carlo runs, the greater the disparity may be between results returned for a given design. Each simulator run produced one value for each of the evaluation metrics. Thus, the simulation results returned to the user interface consisted of a set of 12 or 25 values, one for each run, plus an average value based upon that set, for each of the evaluation metrics.

The process of simulation consisted of creating a runtime representation of the systems model as the simulated environment into which agents would be placed, move, and engage in activity. Each agent within the simulation represented one per-

son generated from one of the five sub-populations. The number of agents placed into the environment within each of the intervals was defined by the sub-populations. However, the time of their placement into the simulated environment was a uniform distribution across the time window of each interval. Once placed into the simulated environment, agent action was driven by a behavior control structure having a simplified "observe, select, act" representation (Figure 3-8).

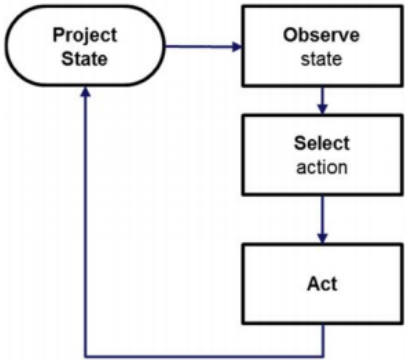


Figure 3-8: A simplified ABMS Observe-Select-Act behavioral model. Reused with permission of the author. ©2012 Moser[40]

The observation step in this loop was governed by the simulated environment. Each space within the model emitted a signal indicating the activities available within that space. These signals may be observed by an agent within that space, and in the case of the Star City model, all signals present in a space were always observed by an agent. Additionally, because agents represent Martian settlers whom live on-site and are intimately familiar with their surroundings, it was decided that signals from every space in the model would propagate, through the modeled connections, for an unlimited distance. This provided each agent with a "map" to any given activity available. As signals propagated through connections, in any given space, an agent could observe the signals for all activities available in that space, as well as signals "pointing" in the direction of activities further away. The only exception to the propagation of signals was for Circulation type activities. These signals would not propagate an unlimited distance but were instead limited to propagating a maximum distance of one space. This allowed an agent to "linger" within the space they were already in

without engaging in a non-circulation activity, or "wander" to adjacent spaces simply to circulate there, but prevented agents from moving a significant distance simply to circulate.

After having observed the signals available, one would be selected by an agent to be acted upon. If the agent was in the process of acting upon a demand from a previously observed signal, the agent would move on in the behavioral loop to acting upon that existing demand based upon the observed signals. However, when an agent was not in the process of acting upon a demand from a previously observed signal it would select an observable signal from which to create a new demand. The process of selection was governed by activity type sensitivities of the demographic from which the agent was instantiated. As activity sensitivity values for each agent were independently drawn from the distribution defined by its demographic at the beginning of each simulation run, it was extremely unlikely that any two agents were the same within the runs of the simulation execution. It should also be noted that when these values were drawn from the distribution, they were not re-normalized to equal 100% of the agent's probability of selection. Rather, this re-normalization was done during the process of signal selection, and based only on those activity types for which the agent had observed a signal (all other sensitivities were removed from consideration when making that selection). In the case where the agent was not sensitive to any of the observable signals, a signal was selected at random. Once the agent had selected a signal upon which to act, the agent would create a new demand based on that signals activity type and space of origin.

The final step of the agents' behavioral loop was to act upon the demands of the agent from previously selected signals. The act portion of the behavioral loop could take four forms — first, the agent could choose to engage in an activity in the space in which it was located, second the agent could choose to move within its current space to either a an activity matching its demand, or, to the edge of the space, third, it could transit a connection to another space if it was at the edge of its current space, and finally, the agent could replace its current demand with a new demand based on an observed signal for the same activity type as its current demand. This final

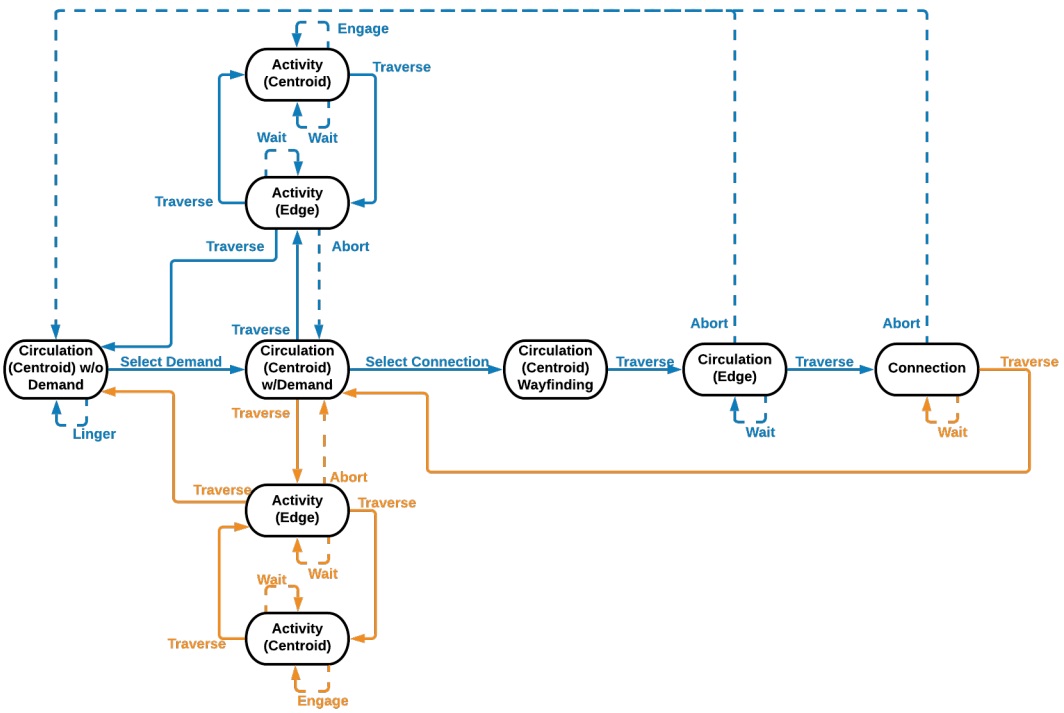


Figure 3-9: Agent state transition diagram, conceptual. Created collaboratively with Katherine M. Carroll[36].

form of action only occurred when an agent followed a propagated signal. The first three forms of action represent the state transition logic of an agent (Figure 3-9) as it moves through the simulated environment in order to engage in a demand from a selected signal.

As agents moved through the site and engaged in activities, the simulator sampled values from various model objects within the simulated environment. These values were used to calculate the seven types of evaluation metrics returned by the simulation. The metrics calculated for a model object depended on its type (Table 3.9), in combination this resulted in 570 individual evaluation metrics. As noted in Section 3.6 the workshop software was modified between the first and second experiments in order to reduce the size of the solution space available to the participants. This resulted in the aggregation of Space type model object metrics into metrics on the "Level" scale of the systems model, and a final count of 78 total metrics in each objective vector generated during a simulation.

Model Object Type	Metrics
Population	Average Agent Transit Time (Avg.), Average Agent Time Onsite (End)
Site	Average Agent Transit Time (Avg.), Average Agent Time Onsite (Avg.), Site Average Utilization (Min., Max., Avg.), Energy Use (Avg.), Average Interaction Probability (Avg.), Average Diversity Score (Avg.)
Space	Site Occupancy (Min., Max., Avg., End), Site Interaction Probability (Min. Max., Avg., End), Site Diversity Score (Min., Max., Avg., End)
Connection	*None*

Table 3.9: Model objects and the metrics for those objects tracked by the simulator.

The ABMS used in this work is further described and analyzed by Carroll[36], a co-collaborator in its development. Kimura[41] also uses the same modeling framework and an ABMS derived from the one used in this work, in a case study, with extensive validation, applied to Odawara, Japan.

3.6 Experiments

3.6.1 Experiment Series

Workshop 1 - March 18, 2021

Participants

This workshop was held for the 2021 cohort of MIT's System Design and Management (SDM) program, as one of two model based learning workshops included in the SDM curriculum. The workshop was held during a specially scheduled two hour recitation time. Approximately 73 students participated resulting in 10 teams of three assigned to treatment group A, 10 teams of three assigned to treatment group B, and three teams of three plus one team of four assigned to group C.

Notes

This workshop was the first run of the experimental design and acted as the prototype for identifying improvements to the workshop materials and workshop software for the later experiments. This was also the first public demonstration and use of the workshop software including its underlying ABMS.

For this event, the the number of systems model decisions within the user "game" interface was larger than what has previously been described, with the transportation mode selection and functional space allocation being a per-connection and per-space, rather than per-level, decision, respectively. It was noted during this workshop that this number of design decisions was overwhelming to users.

In the version of the workshop software provided to participants in this event, the ABMS used 12 independent runs per user simulation execution to produce the returned simulation results.

Due to the special scheduling of the event and the nature of academic recitations, participation fluctuated over the course of the workshop. Some teams had members drop part-way through the team activity, while other teams had members added as students joined the recitation late, both due to scheduling conflicts.

No walk-through of the workshop software was given as part of this event. Due to technical difficulties in packaging and distributing the software leading up to the scheduled team activity period, installation and use of the workshop software was delayed. Teams ended up with approximately 40 minutes with which to install, learn, and use the software.

Workshop 2 - April 17, 2021

This workshop was advertised to students active in MIT's GTL, MIT's Engineering Systems Lab (ESL), MIT's Course 16 students more broadly, and several students associated with MIT's AeroAstro department from Switzerland's EPFL; although the event information was passed on informally, via email, beyond these groups.

Participants

15 individuals pre-registered for the event, of which only nine participated in the event, resulting in three teams of three, one in each treatment group A, B, and C.

Notes

The user "game" interface for this workshop was modified to reduce the transporation mode allocation and functional space activity allocation to be on a per-level basis as previously described in Section 3.5.

In the version of the workshop software provided to partipcants in this event, the ABMS used 12 independent runs per user simulation execution to produce the returned simulation results.

Workshop materials were refined prior to this workshop including the addition of the job aids and revisions to the canonical design in order to align them with the changes made to the user "game" interface for the design decisions. However, the value functions for each of the stakeholder profiles were not yet defined or added for this event.

A walk-through of the workshop software was given to all participants in place of an intended instructional toy problem. The walk-through was meant to have each team step through simulation of a shared baseline design. However, due to the ordering of steps in the workshop structure, team formation had not occurred and therefore teams were unable to launch the workshop software, having not been assigned a team number.

Workshop 3 - April 23/24, 2021

This workshop was offered to students affiliated with Prof. Tetsuya Toma and MIT's GTL at the University of Tokyo and Keio University.

Participants

32 individuals pre-registered for the event, of which 28 participated, resulting in four teams of three assigned to treatment group A, two teams of three and one team of four assigned to group B, and two teams of three assigned to treatment group C.

Notes

The majority of this workshop, including the workshop software walk-through, but *not* the Scenario and Background, was presented in Japanese by Prof. Bryan Moser PhD. and the user "game" interface was internationalized to render in Japanese.

In the version of the workshop software provided to partipcants in this event, the ABMS used 25 independent runs per user simulation execution to produce the returned simulation results.

The workshop materials were again refined for this workshop. Canonical Design 2 was modified in order to make it distinctly appealing to the Mission Health Engineer stakeholder profile and the value functions associated with the stakeholders were added to all three profiles.

As originally assigned, treatment group B had two teams of two rather than one team of four. However, one team failed to have any individual that could install the software and was merged with the other team of two to create a team of four.

One team had difficulty uploading their simulation results data at the end of the workshop and instead tried to share the results file over Zoom. Unfortunately, the file shared was empty, resulting in data loss for that team. Post workshop follow-up with the team failed to recover the data.

Several teams had difficulty understanding the intent of the participant surveys and also questioned their ability to answer the pre-team-activity questions having not had any experience with the systems model.

3.7 Summary

This chapter began by building out the argument for an experimental design to test the hypotheses of this work and explore the underlying research questions. Rationale for the decisions made in this design were given and a model problem applying the experimental design was then described. The use of the model problem as part of an instrumented teamwork workshop experiment was then detailed, including the implementation of workshop software that was used to run a series of such workshop experiments, summaries of which were provided. Through this, the research methodology used to gather the data presented and analysed in Chapter 4 has been presented.

Chapter 4

Results & Analysis

Using the experimental design described in Chapter 3 a series of three experiments were run over the course of March and April, 2021 as described in Section 3.6. This chapter presents and individually analyzes the data collected during each of the three experiments.

4.1 Analysis Approach

The data collected during each experiment attempted to measure two distinct phenomena, the evolution of a team's "stakeholder" members' mental models over time, and the team's tradespace exploration. This work seeks to detect the influence of one upon the other in that it is argued, all else being equal, the pre-exploration mental models of the stakeholders are likely to influence the nature and pattern of tradespace exploration by the team, which in turn influences the mental models of the stakeholders (Figure 4-1).

To measure this process, three things were considered: (1) the ability to compare mental models within and across treatment groups over time, (2) the quantification of a team's exploration in terms of objective overall performance (i.e. how good was the team's outcome given the high-level goal they were given), and (3) the meaningfulness of a team's overall pattern of systems model use, and/or the instantaneous changes within that patter (i.e. exploration, learning, awareness of systemic effect). The goal

was to qualitatively and quantitatively classify teams by their level of mental model alignment and awareness of systemic effect within the systems model (Figure 4-2).

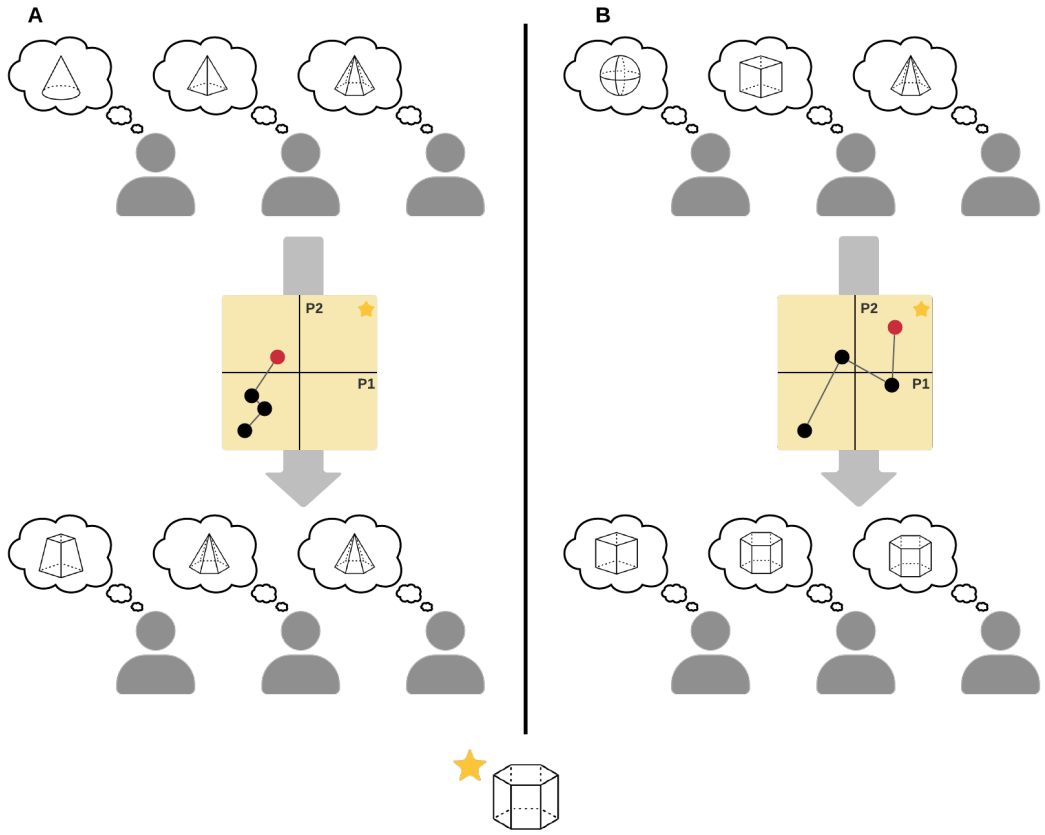


Figure 4-1: The hypothesized process of mental models affecting tradespace exploration, which in turn affects mental models. *A* represents similar mental models leading to a pattern of exploration that limits understanding of systemic effect, whereas *B* represents dissimlar mental models leading to the opposite effect. P1 and P2 are selected performance metrics of the system used to represent the tradespace's objective space. A Pareto Optimal design, a hexagon, is denoted by the star.

This work explores a straightforward method of quantifying both of these two qualities — alignment being measured by the differences in a team's subjective evaluations of reference system-designs, and, the ranking of a team's designs against Pareto optimal designs as a proxy for awareness of systemic effect arising from the team's mental models.

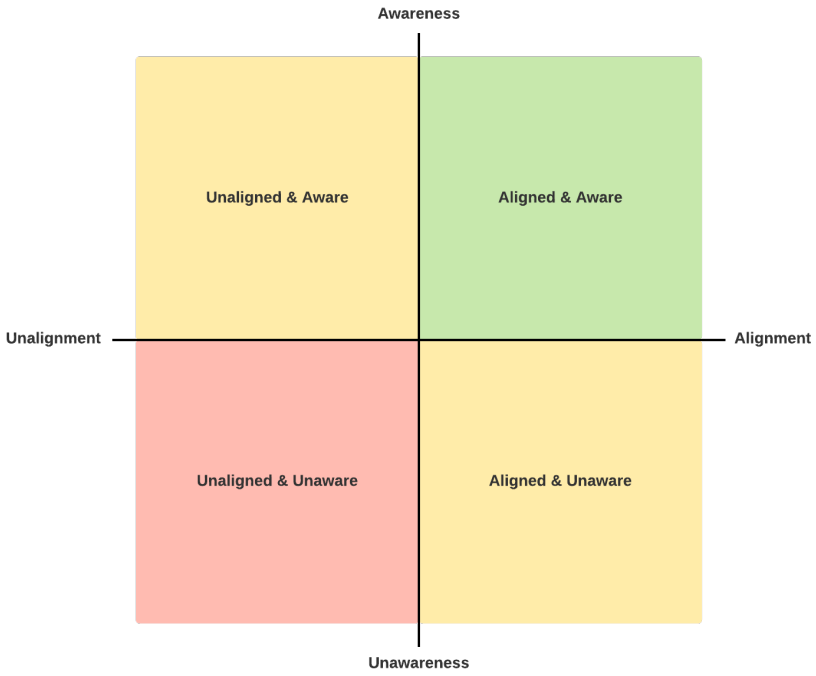


Figure 4-2: A representation of the quadrants of classification for stakeholder teams across alignment of mental models and awareness of systemic effect.

4.1.1 Mental Models

As discussed in the previous chapter, the measurement of mental models was captured through pre- and post-team-activity surveys that collected data on the participants' evaluation of desirability and feasibility for four canonical site designs. This data consisted of independent Likert item values that were *not* structured to create an actual Likert scale. Because of this, the data had to be treated as a true ordinal, rather than interval, dataset during analysis. Therefore, the primary metrics for evaluation of this data were the median, range, and sum, as well as the "spread" of the team's responses for a given question. Median and sum values, when used as a repeated measurement, help to illuminate the presence and nature of mental model changes, while range and spread provide insight into the level of alignment between team members' mental models. Use of the latter metric, spread, was suggested by Dawson[42] as a measure of social diversity when working with ordinal data.

The individual Likert item questions answered by a team, when taken as a set,

capture two related, but distinct, aspects of mental model alignment — their "distance" and their level of agreement. Distance refers to how far apart the assessed measures of desirability or feasibility are within a team, for a given design. Smaller distances representing an overall lower difference in the underlying design evaluation. The range of a team's responses is then a direct measure of this concept of distance. However, distance does not capture the amount of agreement on a team. Take, for example, the following two sets of survey responses for a given system design, each coming from a hypothetical team of five members:

$$\text{Example Team 1 Responses: } \{1, 2, 3, 4, 5\}$$

$$\text{Example Team 2 Responses: } \{1, 5, 5, 5, 5\}$$

For both teams, the range of responses have a value of four, despite the second team, where one dissenter has an extremely divergent evaluation from the others, clearly having greater agreement, and therefore alignment, of mental models. An alternative metric, spread, is therefore needed that takes into account this aspect of agreement.

Spread is calculated based upon Teachman's index for nominal data, but modified, as suggested by Dawson[42], through the multiplication of the index value by the data's range. As Teachman's index is functionally just a measure of Shannon entropy in the natural log base, its value increases the more disparate the values (the less agreement is present) in the set used for its calculation. Scaling such a value by the range then provides a composite value that captures both the measurement of distance between the team's evaluations and the level of agreement on the team, with lower values of spread indicating greater overall alignment.

Before attempting to quantify the alignment of a team's mental models, an analysis of the raw, pre- and post-team-activity Likert item scores was made in order to demonstrate that the data did in fact capture a representation of the mental models of workshop participants. The raw survey scores were analyzed using a chi-squared goodness of fit test against two hypothesised generative models which could realistically have represented the mechanism by which the survey data was created.

The first model tested was that the survey data was generated by the workshop participants at random and thus would be distributed among the Likert item values with a uniform distribution. The second model tested was that survey takers exhibit a central value tendency bias when no mechanism for being able to select a value closer to either extreme end of the scale otherwise exists. Thus, survey data would have been generated with a normal distribution centered at the middle Likert item score of three and have only a small deviation from this central value.

Based upon these models, two null hypotheses were set forth against which the survey data was tested. Rejection of both of these hypotheses would then indicate that the survey scores were likely generated from some other distribution, one that this work argues holds as a representation of participant mental models. The null hypotheses set forth were as follows:

H_{0a} (Random Selection) : The desirability and feasibility scores are sampled from an underlying distribution that is a discrete uniform random distribution taking values in the range $[1, 5]$.

H_{0b} (Central Tendency) : The desirability and feasibility scores are sampled from an underlying discrete distribution for which the cumulative distribution function approximates that of a normal distribution of the form $N(3, 1)$.

Following this, the four metrics, median, range, sum, and spread, were used to investigate correlations among teams within and between the treatment groups.

Lastly, a score of mental model alignment was calculated from the average spread of a team's responses on desirability and feasibility for the four canonical designs. Teams were then ranked based upon this score and correlation between mental model alignment and tradespace exploration metrics investigated. The pre-team-activity data was analyzed with respect to each individual team with the resulting team metrics then used to measure treatment group effect. The rationale for this approach is that, despite the shared treatment by two teams, the evaluations of desirability and feasibility by the individuals on one of the teams is entirely independent of the other — there is no mechanism by which inter-team alignment of desirability and feasibility

can necessarily be reached. While each team is represented by the same stakeholders, all of whom share the same knowledge, the interpretation of that knowledge relative to the system is independent and unmediated. Within treatment group A, two teams may independently come to different evaluations of a given canonical design through team processes that align their assumptions (mental models) within, but not between, the two groups. Within treatment group B, it is only the shared stakeholder profiles that exist to align knowledge, and have been designed to push the stakeholders to the extremes of evaluation on three of the four designs. Post-team-activity data, on the other hand, has a potential mediation mechanism present in the team activity that would appear, under the hypotheses of this work, to align evaluations between individuals within a team but still not necessarily align evaluations between teams.

4.1.2 Exploration

The measurement of a team's tradespace exploration consisted of recording the design vectors and the resulting objective vectors produced by the simulator using those designs for the systems model. These vectors were then independently analyzed in order to attempt to quantify the overall level of exploration engaged in by a team. An evaluation of the *type* of exploration (i.e. concentrated, diffuse, incremental, architectural, etc.) engaged in by a team was considered, but not used for analysis, and is instead discussed in Section 5.1.

Solution Space

Design vectors are composed of a set of nominal data, each element being the value chosen for a given variable in the systems model. Therefore, the metrics selected for analysis of this data were the count of unique vectors in the tradespace exploration, the average number of design vector values changed between each simulation, and the entropy of the exploration and its efficiency. Count and average number of changes are indicators of the volume and nature of exploration, respectively, while the use of entropy measures attempts to further quantify the nature of the exploration from an

information theory perspective.

Given the solutions space as a matrix where each row vector corresponds to a single design, each column vector would then correspond to the full enumeration of values possible for one of the design variables. For unconstrained (independent) design variables, the proportion of occurrences for each value that variable can take would be equal in such a column vector, while for constrained design variables, the proportions would vary as determined by the nature and number of constraints. For unconstrained variables, such a column vector then represents the complete information space represented by the design variable and a measure of its Shannon entropy gives a quantification of the amount of information it contains. Additionally, given that Equation 4.1 holds as strictly equal for independent events, in this case the values of unconstrained design variables, the sum of such column vector entropies in an unconstrained solution space would be the entropy of the entire solution space.

$$H(X, Y) \leq H(X) + H(Y) \tag{4.1}$$

Similarly, the design vectors of a team's design walk, arranged as row vectors in a matrix, can then be used to evaluate the average amount of information captured for each design variable, which compose its column vectors, over the team's exploration. The entropy measure captures the variety of possible values explored for each variable. Efficiency is then the ratio of entropy of the design variable under the design walk to the total information space of the design variable under the solution space. This value is a representation of how well the variety of a variable as explored by a team covers the total variety possible for that variable. This holds only in the case of unconstrained design variables, which make up all of the model problem design decisions in this work.

Problem Space

Objective vectors consisted of the 78 systems model evaluation metrics produced through simulation as described in subsection 3.5.3. Each team's objective vectors

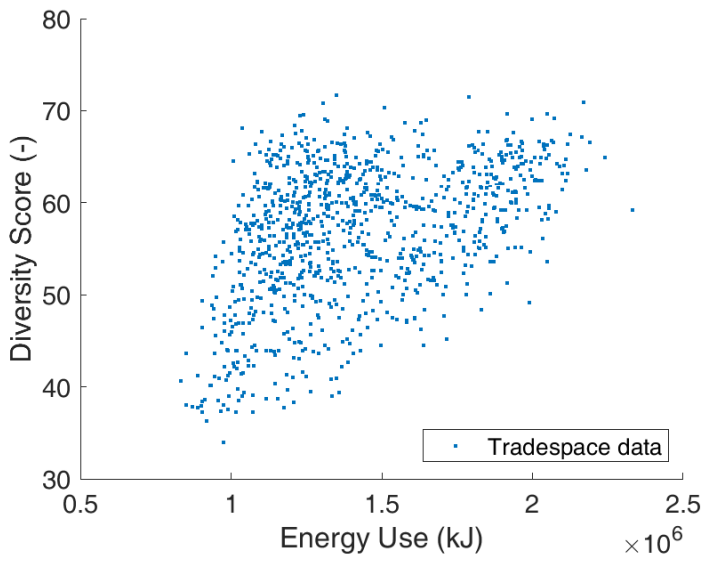
were analyzed by calculating its Pareto Rank within the overall problem space. It should be noted, as described in subsection 3.5.3, the ABMS software provided to the teams intentionally limited the number of Monte Carlo runs to either 12 or 25 per design vector, depending on the version used in a particular workshop. Thus, the objective vectors comprising the team's exploration data were those as seen by the team and used for decision making during the workshop. However, the enumeration of the problem space used for calculation of Pareto Ranks was formed from objective vectors produced by a more robust simulation utilizing 50 Monte Carlo runs per design vector. Therefore, due to the non-deterministic calculation of the objective vectors' values, the results for a given design as seen by a team could diverge from the results of that design within the enumerated tradespace. Given that this represented a real difference in the information provided by the software and used by the team during the workshop, the Pareto Rank of a team's objective vectors is calculated base on the team's "as perceived" system performance, and not the performance for their design as produced from the more robust simulation used for tradespace enumeration.

Additionally, due to limitations in the data collected during the workshop experiments, discussed further in Section 4.2, only a subspace of the full problem space was enumerated and used for Pareto Rank calculations in a demonstration of methods. This reduced problem space enumeration (Figure 4-3) was generated from 1000 design vectors sampled from the solution space and used to calculate the Pareto Ranks using four of the objective vectors' 78 total metrics: Site Energy Use, Site Average Utilization, Site Average Interaction Probability, and Site Average Diversity Score.

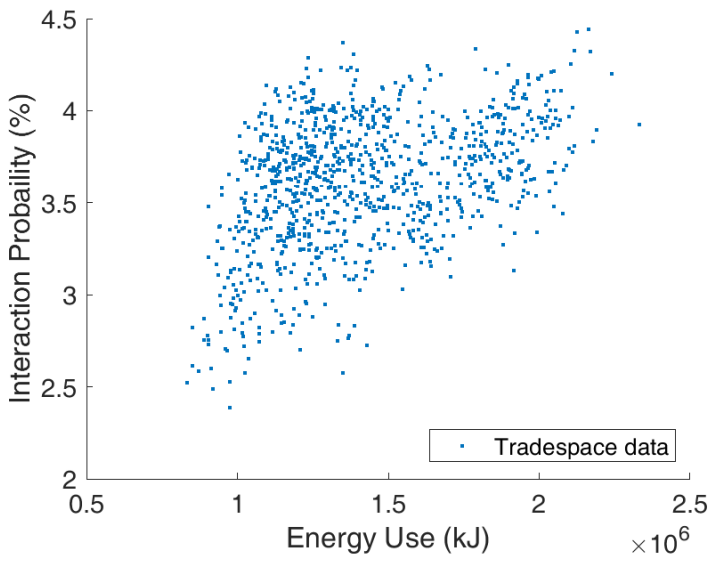
4.1.3 Overall Performance

Finally, overall performance was determined by locating teams within the Alignment vs. Awareness space (Figure 4-2).

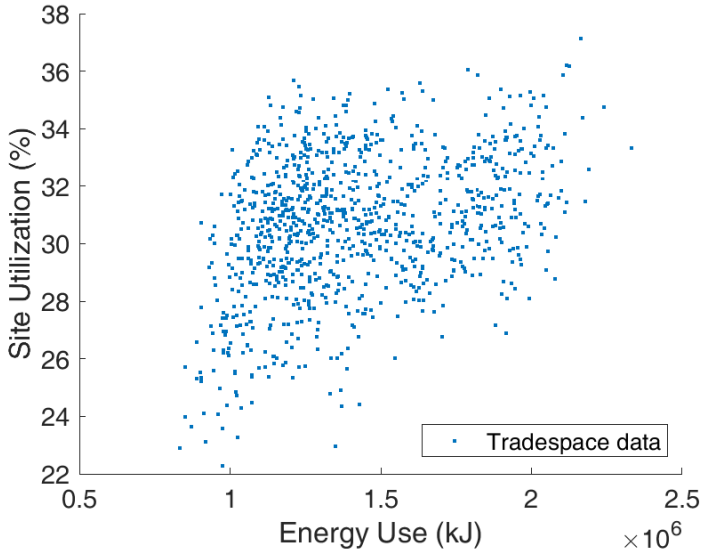
Teams were given a preliminary ranking based on their range, raw Teachman's index, and spread. Correlation were then evaluated between these ranks and the analysis metrics calculated for the team's design walk. These correlations were then used to evaluate metrics for locating teams within the Alignment vs. Awareness



(a) Energy vs. Diversity Tradespace, Enumeration.



(b) Energy vs. Interactions Tradespace, Enumeration.



(c) Energy vs. Utilization Tradespace, Enumeration.

Figure 4-3: Enumerated Tradespace Diagrams.

quadrants. Unfortunately, not enough data was collected during the experiments to statistically suggest a particularly meaningful metric for alignment or performance. Thus, for final overall performance ranking, average spread and Pareto Rank were selected as the basis of comparison for teams. Spread was selected due to its capturing of both the range of difference in survey scores and the level of agreement within a team. While Pareto Rank was selected as it is an often cited objective measure of system performance. The Pareto Rank of a team's first design vector anchors a team's design walk, while their post-exploration preferred design captures some aspect of the teams awareness of systemic effect, and the average Pareto Rank gives some indication as to where in the problem space the team explored.

To allow for plotting and comparison of teams in the Alignment vs. Awareness space, two normalized mental model alignment values and two normalized Pareto Rank values were calculated. The two mental model alignment values used were the team's pre- and post-team-activity desirability and feasibility scores average spread. While the Pareto Rank values taken were that of each team's first simulated design and their preferred design as indicated in their post-team-activity surveys. This provided starting and ending metrics for alignment and awareness, respectively. A third, normalized spread value was linearly interpolated from the pre- and post-team-activity average spread values and paired with the average Pareto Rank of the teams' simulated designs during tradespace exploration. This third value provided an intermediate representation of the team during tradespace exploration. All three pairs of these values were then normalized.

For spread normalization, the maximum raw range under the experimental design was $[0, 4.39]$. For Pareto Rank, a raw range of $[1, R]$ was used for normalization, where R was the worst Pareto Rank value from among all teams' design vectors. The use of R provided for a comparison of teams anchored between the best possible outcome that a team *could* have achieved and the worst possible outcome any team encountered. Normalization was done so as to produce a result within $[-1, 1]$. For spread a normalized value of 1 corresponded to the raw score of 0 (complete alignment). Normalization of the Pareto Ranks was done such that a normalized value of 1

corresponded to the raw score of 1. A final team ranking was then given based upon these normalized scores relative to the Alignment vs. Awareness quadrant Utopia point $(1, 1)$.

4.2 Experimental Data

4.2.1 Workshop 1 - March 18, 2021

Due to the logistical challenges and event complications discussed in Section 3.6 this workshop was treated as a prototype trial for the experimental design. The data gathered from this workshop was not analyzed as part of this work.

4.2.2 Workshop 2 - April 17, 2021

Participants

As described in Section 3.3, this workshop had nine participants that were randomly assigned into three teams of three, with one team each being assigned to treatment groups A, B, and C. Table 4.1 lists the assignment of teams to treatment groups.

Treatment Group	Team Identifier
A	Team 1
B	Team 4
C	Team 7

Table 4.1: Team to treatment group assignments. (Workshop 2)

All nine participants completed both the pre-team-activity and the post-team-activity surveys. Simulation results data was submitted by all three teams.

Given the limited number of teams per treatment group, no statistical comparisons between or within treatments can be made. In addition, for Team 4, one survey respondent mis-entered their stakeholder role in the pre-team-activity survey. In the following analysis, this datapoint was corrected based upon the workshop interactions with the participant who submitted the data and information available from the workshop event. Therefore, the data analysis for this workshop is purely illustrative.

Detection of Mental Models

The raw pre- and post-team-activity survey scores were each used to test the two null hypotheses $\mathbf{H_{0a}}$ and $\mathbf{H_{0b}}$ using a chi-square goodness of fit test (Table 4.2). The fit of the data was also visualized for each hypothesis, by plotting its probability distribution function against the probability mass distribution of the survey data (Figure 4-4).

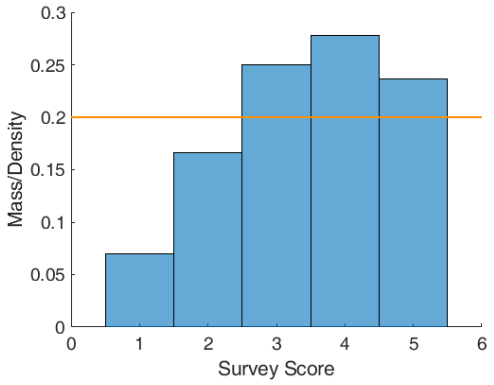
Survey Dataset	Statistic	$\mathbf{H_{0a}}$	$\mathbf{H_{0b}}$	**Alt.** — $N(3, 1.4)$
Pre-team-activity	χ^2	10.1	36.3	9.4
	df	4	4	4
	(p-value)	(0.04)	($2.5e^{-07}$)	(0.05)
Post-team-activity	χ^2	4.0	34.2	5.4
	df	4	4	4
	(p-value)	(0.41)	($6.68e^{-07}$)	(0.25)

Table 4.2: Chi-square goodness of fit test statistics for pre- and post-team-activity survey data under H_{0a}, H_{0b}, and the alternative null hypothesis of a normal distribution which cannot be rejected for pre-team-activity survey data. (Workshop 2) "df" stands for "degrees of freedom".

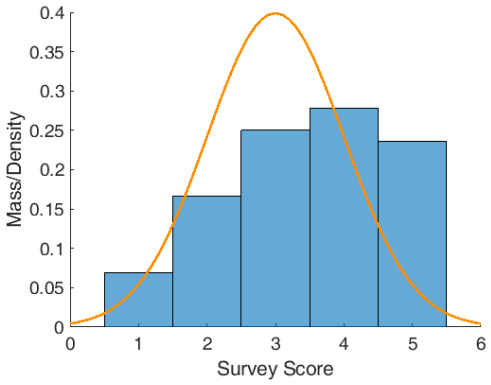
Both of the null hypotheses are rejected for the pre-team-activity data while for the post-team-activity-data only $\mathbf{H_{0b}}$ can be rejected. A sensitivity analysis was also done for alternative versions of $\mathbf{H_{0b}}$ by modifying the standard deviation of the normal distribution being used, in increments of 0.1, until the chi-square test for the pre-team-activity survey data failed to reject the alternative null hypothesis. The point at which the pre-team-activity survey data failed to reject the alternative null hypothesis occurred at a standard deviation of 1.4.

Mental Model Alignment

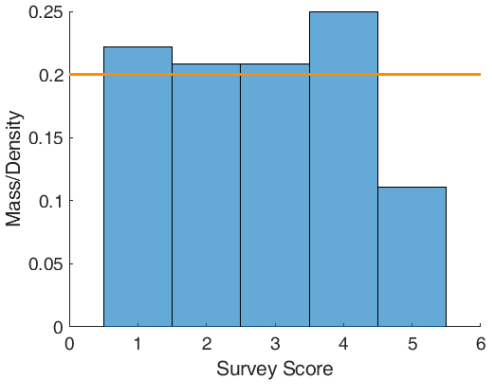
The results of mental model evaluation for all three teams are shown in Figure 4-5 with summary statistics provided in Table 4.3 and Table 4.4.



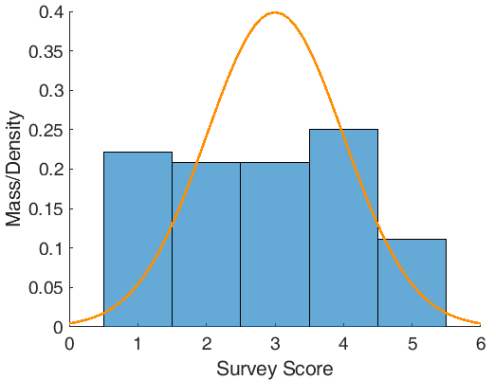
(a) Pre-team-activity survey data vs. $\mathbf{H_{0a}}$.



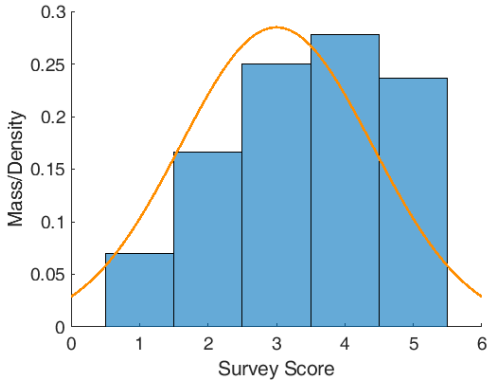
(b) Pre-team-activity survey data vs. $\mathbf{H_{0b}}$.



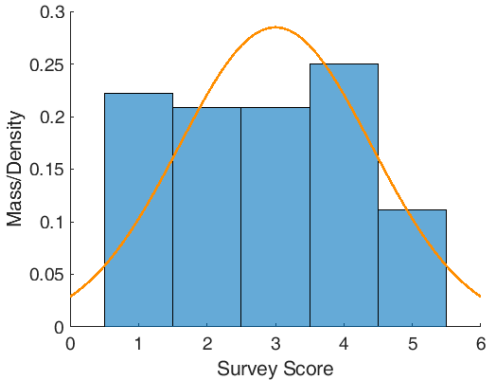
(c) Post-team-activity survey data vs. $\mathbf{H_{0a}}$.



(d) Post-team-activity survey data vs. $\mathbf{H_{0b}}$.

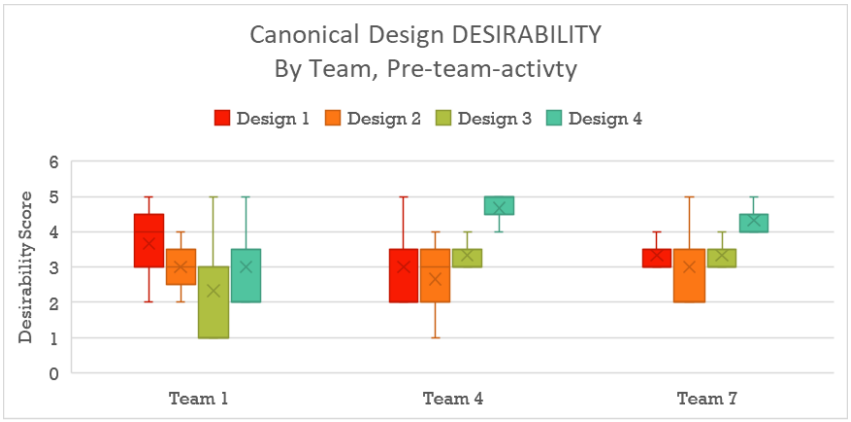


(e) Pre-team-activity survey data vs. Alt. $N(3, 1.4)$

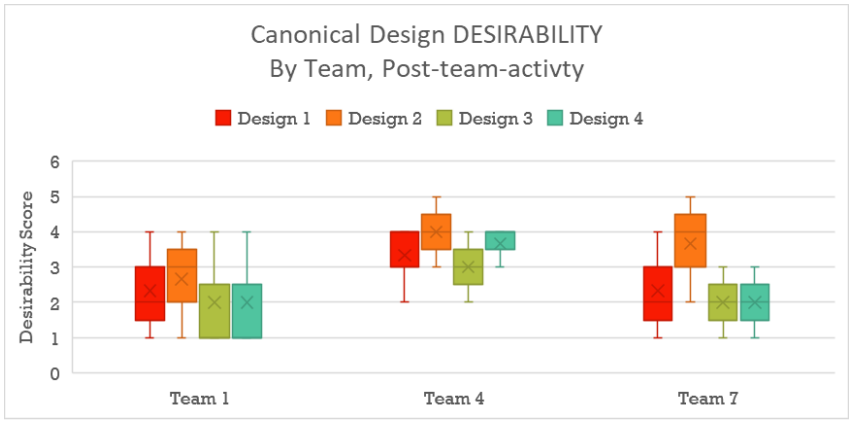


(f) Post-team-activity survey data vs. Alt. $N(3, 1.4)$

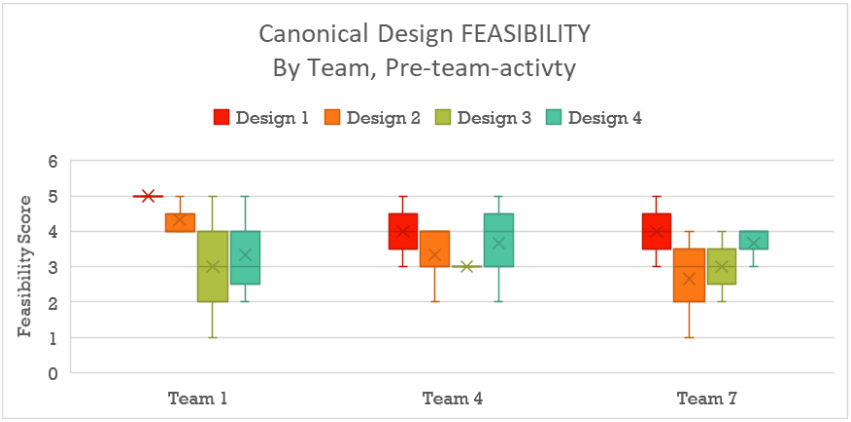
Figure 4-4: Workshop 2 survey data probability mass distribution vs. hypothesized model probability density distribution plots.



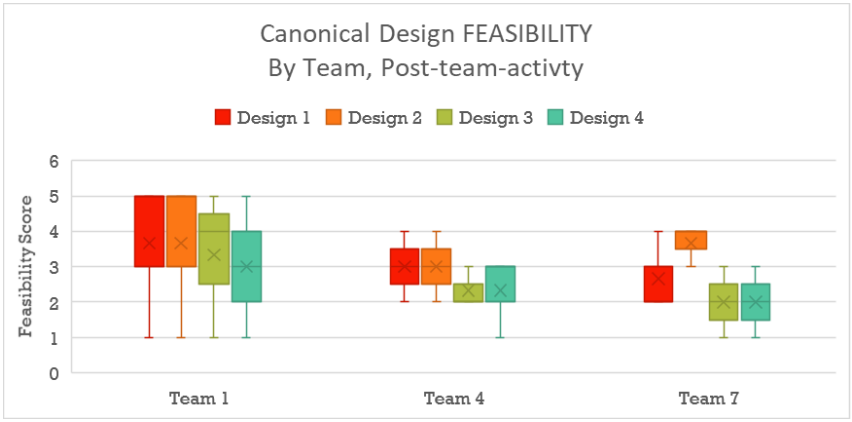
(a) Desirability, Pre-team-activity Survey.



(b) Desirability, Post-team-activity Survey.



(c) Feasibility, Pre-team-activity Survey.



(d) Feasibility, Post-team-activity Survey.

Figure 4-5: Workshop 2 canonical design evaluation survey results, by team.

	Desirability					Feasibility				
	Design 1	**Design 2**	**Design 3**	**Design 4**		**Design 1**	**Design 2**	**Design 3**	**Design 4**	
Treatment Group A										
Metric					**Average**					**Average**
Median	4	3	1	2	2.5	5	4	3	3	3.75
Range	3	2	4	3	3	0	1	4	3	2
Teachman's Index	1.1	1.1	0.64	0.64	0.87	0	0.64	1.1	1.1	0.71
Spread	3.3	2.2	2.55	1.91	2.49	0	0.64	4.39	3.30	2.08
Sum	11	9	7	9	9	15	13	9	10	11.75
Treatment Group B										
Metric					**Average**					**Average**
Median	2	3	3	5	3.25	4	4	3	4	3.75
Range	3	3	1	1	2	2	2	0	3	1.75
Teachman's Index	0.64	1.1	0.64	0.64	0.75	1.1	0.64	0	1.1	0.71
Spread	1.91	3.30	0.64	0.64	1.62	2.20	1.27	0	3.30	1.69
Sum	9	8	10	14	10.25	12	10	9	11	10.5
Treatment Group C										
Metric					**Average**					**Average**
Median	3	2	3	4	3	4	3	3	4	3.5
Range	1	3	1	1	1.5	2	3	2	1	2
Teachman's Index	0.64	0.64	0.64	0.64	0.64	1.1	1.1	1.1	0.64	0.98
Spread	0.64	1.91	0.64	0.64	0.95	2.2	3.3	2.2	0.64	2.08
Sum	10	9	10	10	13	12	8	9	11	10

Table 4.3: Pre-team-activity canonical design survey statistics, by team. (Workshop 2)

	Desirability					Feasibility				
	Design 1	**Design 2**	**Design 3**	**Design 4**		**Design 1**	**Design 2**	**Design 3**	**Design 4**	
Treatment Group A										
Metric					**Average**					**Average**
Median	2	3	1	1	1.75	5	5	4	3	4.25
Range	3	3	3	3	3	4	4	4	4	4
Teachman's Index	1.1	1.1	0.64	0.64	0.87	0.64	0.64	1.1	1.1	0.87
Spread	3.3	3.3	1.91	1.91	2.6	2.55	2.55	4.39	4.39	3.47
Sum	7	8	6	6	6.75	11	11	10	9	10.25
Treatment Group B										
Metric					**Average**					**Average**
Median	4	4	3	4	3.75	3	3	2	3	2.75
Range	2	2	2	1	1.75	2	2	1	2	1.75
Teachman's Index	0.64	1.1	1.1	0.64	0.87	1.1	1.1	0.64	0.64	0.87
Spread	1.27	2.2	2.2	0.64	1.58	2.2	2.2	0.64	1.27	1.58
Sum	10	12	9	11	10.5	9	9	7	7	8
Treatment Group C										
Metric					**Average**					**Average**
Median	2	4	2	2	2.5	2	4	2	2	2.5
Range	3	3	2	2	2.5	2	1	2	2	1.75
Teachman's Index	1.1	1.1	1.1	1.1	1.1	0.64	0.64	1.1	1.1	0.87
Spread	3.3	3.3	2.2	2.2	2.75	1.27	0.64	2.2	2.2	1.58
Sum	7	11	6	6	7.5	8	11	6	6	7.75

Table 4.4: Post-team-activity canonical design survey statistics, by team. (Workshop 2)

Pre-team-activity Survey Results

For the four canonical designs, treatment group A had the largest average range of desirability and feasibility scores, 3 and 2 respectively, followed by group B's scores of 2 and 1.75 and C with average ranges of 1.5 and 2, respectively. It must be noted, however, that two members of treatment group C indicated in their survey responses that they had previous experience working together (and were, in fact, brothers) and additionally had an intimate knowledge of the Star City Martian settlement concept. Due to the small sample size, no meaningful statistical inference can be made from these values. Averaging the ranges of all desirability and feasibility scores for each team, with lower ranges being better, produced a team ranking, from best to worst, of: Team 7, Team 4, Team 1.

Treatment group A's spread was also the greatest with values of 2.49 for desirability and 2.08 for feasibility, again followed by group B with average spreads of 1.62 and 1.69, and group C having a spread 1.62 for desirability scores and 2.08 for feasibility scores. The overall lower value of the average spread, relative to the range, indicates some level of agreement between team members; this was confirmed by inspection of the individual survey responses. Averaging the spreads of each team, with lower being better, also produced a team ranking, from best to worst, of: Team 7, Team 4, Team 1.

Evaluation of the median and sum values of each team's pre-team-activity survey desirability and feasibility scores is only meaningful, for this work, in relation to the same measures post team activity. As such, they are discussed in the next section.

Post-team-activity Survey Results

In general, the change in a team's desirability and feasibility scores was mixed both for a given design and across designs. Treatment group A's median desirability score decreased for half of the canonical designs while the sum of scores for all four design decreased, with an average reduction of 2.25 points. However, although half of their median feasibility score increased, the sum of feasibility scores for three of

the four designs decreased by an average of 2.33 points and one increased by a point. Treatment group B's median desirability scores showed a small average increase of 1.33 points for three designs with two of their score sums increasing an average of 2.5 points but with the third decreasing 3 points. The one design for which the median remained unchanged had a drop of 1 point in the sum of its scores. The median and the sum of feasibility scores in group B fell an average of 1 point and 2.5 points for all four designs, respectively. For treatment group C a clear pattern was discernible, the median and sum of the desirability and feasibility scores fell for all designs except for the second design. That design showed an increase in median scores of 2 points and 1 point for desirability and feasibility, respectively, and an increase in the sum of scores of 2 points and 3 points, respectively, as well.

Post team activity, the average range of desirability for the canonical designs for treatment group A remained 3 while the spread increased, and for feasibility, group A's average range and spread both increased as well. Group B's average range and spread for desirability both decreased, while only spread decreased for their feasibility scores, range remaining unchanged. Both average range and spread increased for group C's desirability scores in contrast to their feasibility scores for which both decreased. As mental model alignment has been argued to require similarity in the assessment of both desirability and feasibility, taking the scores together and ranking the teams post team activity, the average range of scores ranked the teams, from best to worst, Team 4, Team 7, Team 1. Similarly, the average spread also gave a ranking of Team 4, Team 7, Team 1.

Pre/Post Comparison

Despite the lack of data necessary for measuring statistically significant correlations between treatment group and pre/post-team-activity mental model alignment metrics, these calculations were made regardless, as a demonstration of method. Treatment groups were encoded as a binary variable and point-biserial correlation coefficients were calculated for each pairwise combination of the groups for the team's average range, Teachman's index, and spread values for both desirability and feasibil-

ity. Correlation coefficients for pre-team-activity survey scores are shown in Table 4.5 while those for post-team-activity scores are shown in Table 4.6.

Interpretation of this data would indicate a strong influence for treatment group A on its spread having a higher mean value than treatment group B in both the pre- and post-team-activity survey scores. Treatment group A's range and spread also shows a similar but mostly moderate correlation with group C in both pre and post survey scores. Only weak correlations appear between treatment groups B and C across all three metrics in the pre-team-activity survey scores, yet a consistent moderately negative correlation appears post team activity. However, as expected due to the low sample size, the p-values indicate that the null hypothesis, that the treatment group samples were drawn from a distribution having the same mean but potentially different variances, cannot be rejected for any of these correlations.

Treatment Group Pair	Statistic	Range	Teachman's Index	Spread
A/B	ρ	0.56	0.40	0.79
	(p-value)	(0.35)	(0.54)	(0.09)
A/C	ρ	0.60	-0.07	0.58
	(p-value)	(0.31)	(0.93)	(0.33)
B/C	ρ	0.26	-0.27	0.15
	(p-value)	(0.7)	(0.69)	(0.83)

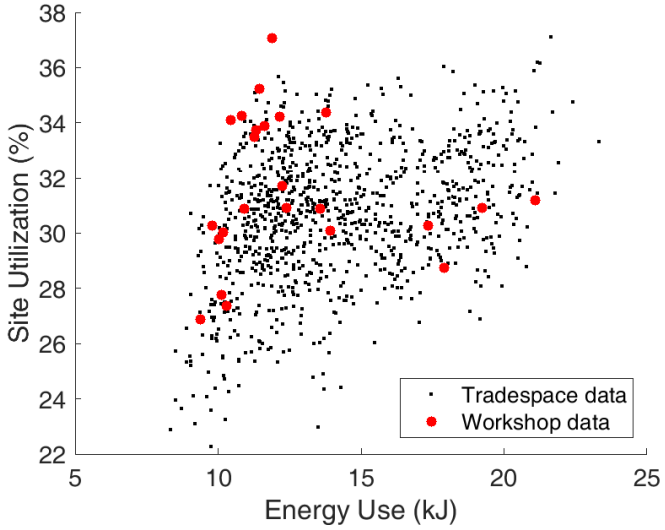
Table 4.5: Pre-team-activity survey point-biserial correlation coefficients for treatment group assignment and three metrics of mental model alignment. (Workshop 2)

Treatment Group Pair	Statistic	Range	Teachman's Index	Spread
A/B	ρ	0.80	-	0.79
	(p-value)	(0.07)		(0.08)
A/C	ρ	0.72	-0.5	0.56
	(p-value)	(0.16)	(0.43)	(0.36)
B/C	ρ	-0.5	-0.5	-0.5
	(p-value)	(0.42)	(0.42)	(0.42)

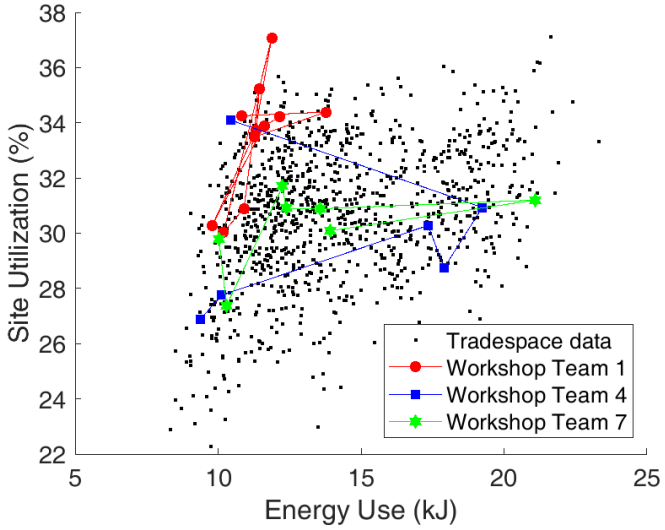
Table 4.6: Post-team-activity survey point-biserial correlation coefficients for treatment group assignment and three metrics of mental model alignment. (Workshop 2) Correlation for treatment group pairing A/B's Teachman's index values could not be calculated due to lack of variation within the data.

Tradespace Exploration

Figure 4-6a and Figure 4-6b show the overall solution space exploration and the design walks for all three teams, respectively, for one pair of system evaluation metrics — Site Energy Use vs. Site Average Utilization. The tradespace exploration and design walks for the remaining metrics used for analysis are shown in Figure 4-7 with evaluation metrics for each team's exploration shown in Table 4.7.

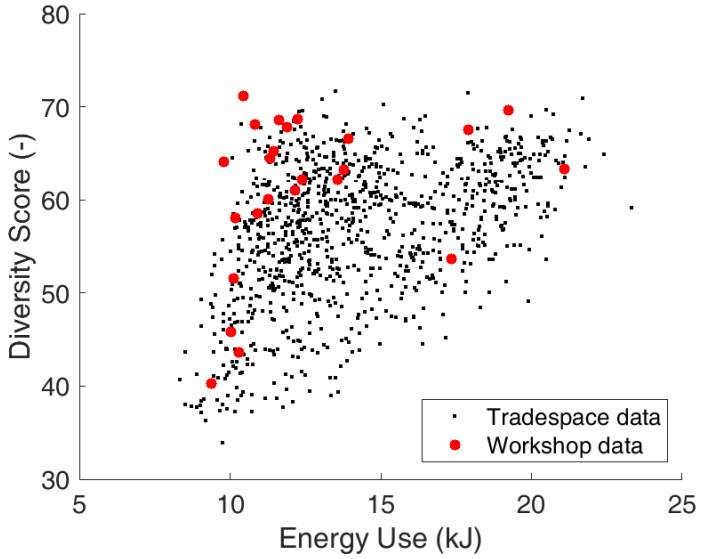


(a) Solution Space exploration.

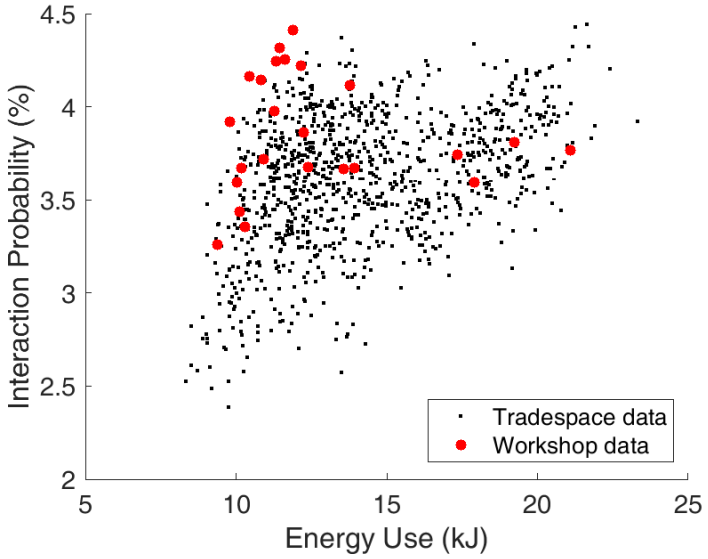


(b) Design walks for all three teams.

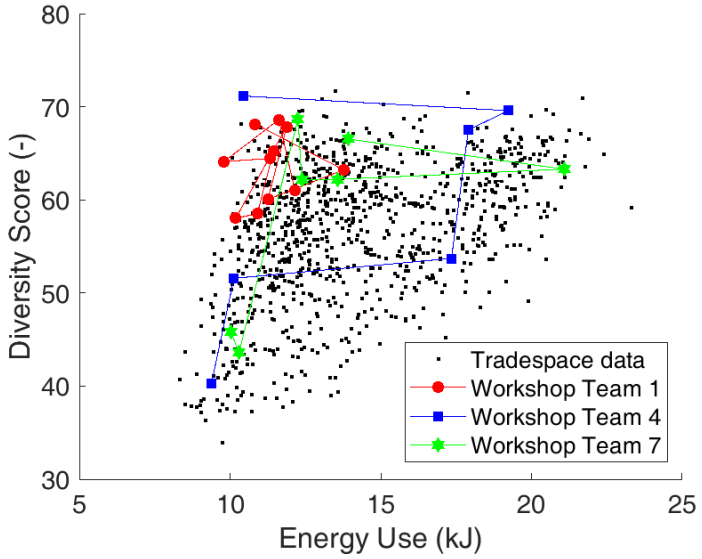
Figure 4-6: Workshop 2 Tradespace Diagrams, Site Energy Use vs. Site Average Utilization.



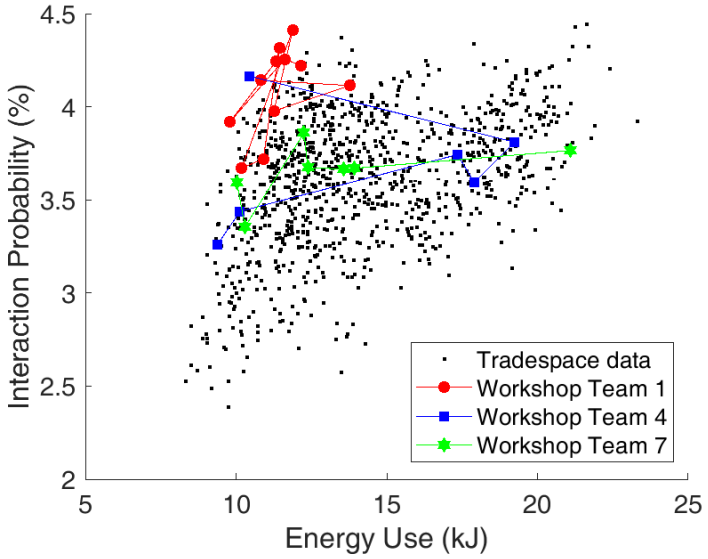
(a) Energy vs. Diversity Tradespace, Overlay.



(b) Energy vs. Interactions Tradespace, Overlay.



(c) Energy vs. Diversity Tradespace, Design Walk.



(d) Energy vs. Interactions Tradespace, Design Walk.

Figure 4-7: Workshop 2 Tradespace Diagrams.

		Solution Space Exploration Statistics							Problem Space Exploration Statistics				
		Designs		Inter-design Changes					Pareto Rank				
Treatment Group	**Team Identifier**	**Total**	**Unique**	**Min.**	**Max.**	**Avg.**	**S.D.**	**Efficiency**	**Min.**	**Max.**	**Avg.**	**S.D.**	**Preferred Design**
A	Team 1	11	8	0	10	3.4	3.29	0.37	1	5	2	1.28	1
B	Team 4	6	5	0	6	3.4	2.42	0.24	1	12	4.5	3.59	1
C	Team 7	7	5	0	6	3.33	2.43	0.34	2	8	5.14	2.29	2

Table 4.7: Tradespace exploration statistics Workshop 2. For Pareto Rank, lower is better.

Design Walk Evaluation

Visually, the design walks for each of the three treatment groups appeared distinctly different. Similarly, the tradespace exploration metrics (Table 4.7) appeared to show some differences between the treatment groups, but due to only having a single sample each, no statistical significance from these differences can be drawn. However, as a demonstration of the method of analysis, correlation coefficients were calculated, using Kendall's τ[43], between the ranking of teams using their mental model alignment metrics, previously discussed, and their tradespace exploration metrics. This was done for both the pre- and post-team-activity mental model alignment rankings with the former representing the "forward" relationship between alignment and exploration and the latter the "reverse" relationship of exploration on scores. Table 4.8 and Table 4.9 summarize the results of this analysis.

Pre team activity, the strongest, negative, correlation existed between the team rank and the average Pareto Rank of the team's exploration, followed by weaker positive correlations with number of unique design vectors, maximum and average number of inter-design vector changes, and a weaker negative correlation with the maximum Pareto Rank among a team's designs and the team's preferred design choice. Given the number of samples, the p-values show that the null hypothesis of no correlation between pre-team-activity team rank and these metrics *cannot* be rejected.

Looking at the post team activity correlations shows the strongest correlations existed between team rank and the total number of designs explored, the standard deviation of design vector changes, and the entropy efficiency of the design vectors, and the strongest negative correlation with the the maximum Pareto Rank among a team's designs and the standard deviation between Pareto Ranks. Weaker correlations existed between team rank and the number of unique designs explored and the maximum number of inter-design vector changes made by a team. Again, due to the number of samples, these values hold no meaningful statistical significance and the null hypothesis of no correlation between post-team-activity team rank and these metrics *cannot* be rejected.

		Solution Space Exploration Statistics							Problem Space Exploration Statistics				
		Designs		Inter-design Changes					Pareto Rank				
Ranking Method	**Statistic**	**Total**	**Unique**	**Min.**	**Max.**	**Avg.**	**S.D.**	**Efficiency**	**Min.**	**Max.**	**Avg.**	**S.D.**	**Preferred Design**
Range/Spread	τ	0.33	0.82	-	0.82	0.82	0.33	0.33	-0.82	0.33	-1	-0.33	-0.82
	(p-value)	(1)	(0.67)		(0.67)	(0.67)	(1)	(1)	(0.67)	(1)	(0.33)	(1)	(0.67)

Table 4.8: Pre-team-activity survey Kendall's tau correlation coefficients for tradespace exploration statistics with teams ranked by average range or spread of diversity and feasibility scores. (Workshop 2) Range and spread resulted in the same ranking of teams, therefore coefficients are reported only once. No correlation possible for the minimum number of inter-design changes metric.

		Solution Space Exploration Statistics							Problem Space Exploration Statistics				
		Designs		Inter-design Changes					Pareto Rank				
Ranking Method	**Statistic**	**Total**	**Unique**	**Min.**	**Max.**	**Avg.**	**S.D.**	**Efficiency**	**Min.**	**Max.**	**Avg.**	**S.D.**	**Preferred Design**
Range/Spread	τ	1	0.82	-	0.82	0	1	1	0	-1	-0.33	-1	0
	(p-value)	(0.33)	(0.67)		(0.67)	(1)	(0.33)	(0.33)	(1)	(0.33)	(1)	(0.33)	(1)

Table 4.9: Post-team-activity survey Kendall's tau correlation coefficients for tradespace exploration statistics with teams ranked by average range or spread of diversity and feasibility scores. (Workshop 2) Range and spread resulted in the same ranking of teams, therefore coefficients are reported only once. No correlation possible for the minimum number of inter-design changes metric.

Alignment vs. Awareness Ranking

Using the teams' survey scores for desirability and feasibility, each team's average spread value was calculated both for pre and post team activity. The possible range for spread scores was $[0, 4.39]$, given teams of three and desirability/feasibility scores ranging from $[1, 5]$. A team's average spread value, given its possible range, was then normalized to the range of $[-1, 1]$, with the raw spread score of 0 (complete alignment) corresponding to 1 when normalized. The Pareto Rank value for a team's first simulated design vector and preferred design, as indicated in their post-team-activity survey, were also normalized to $[-1, 1]$ based upon a raw range of $[1, R]$ with R being the worst Pareto Rank among all teams' design walks. When normalized, a raw Pareto Rank score of R corresponded to -1.

Additionally, a linear interpolation was made between the raw average spread values calculated from the pre-team-activity and post-team-activity scores. This was used, along with the average Pareto Rank of a team's design vectors from their tradespace exploration (Table 4.7), to produce an intermediate point for a team within the Alignment vs. Awareness space. This intermediate point represents a proxy for a team's average position during tradespace exploration.

Given these three sets of normalized scores (Table 4.10), the teams' paths within the Alignment vs. Awareness space (Figure 4-8) were plotted and the teams ranked based on their final position relative to the Utopia point $(1, 1)$ (Table 4.11).

All three teams started with a design vector having the same Pareto Rank (2) and therefore the same starting awareness value in the normalized space. (Design vectors for Team 1 and Team 7 were the same baseline design demonstrated in the workshop walk-through while Team 4's first design vector simulated was not.) Team 1 and Team 7 both *reduced* alignment of the team, based on the chosen metric of spread, over the course of their exploration. While Team 4 and Team 7 showed a drop in team awareness during exploration but improved beyond, or at least recovered to, their starting level by the end of the workshop, respectively. Only Team 4 increased both awareness and alignment overall, given the metrics chosen to represent the space.

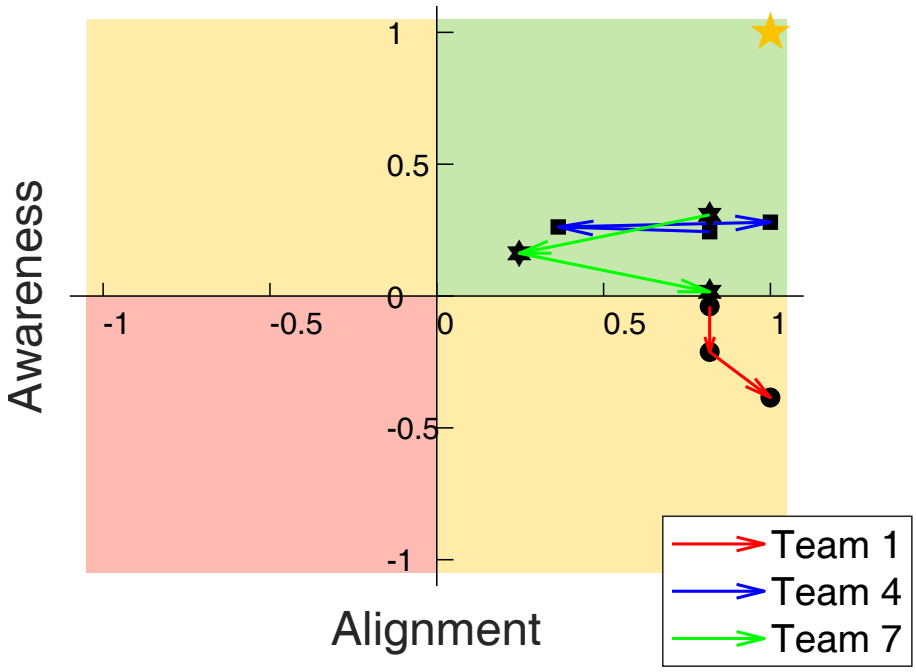


Figure 4-8: Plot of Workshop 2 Alignment vs. Awareness paths. Each team's path is based on its normalized alignment and awareness metrics. The yellow star represents the Utopia point against which teams were ranked.

Treatment Group	Team Identifier	Pre Team Activity				Exploration, Iterpolated				Post Team Activity			
		Average Spread	Starting Pareto Rank	Normalized Spread	Normalized Pareto Rank	Average Spread	Average Pareto Rank	Normalized Spread	Normalized Pareto Rank	Average Spread	Average Pareto Rank	Normalized Spread	Normalized Pareto Rank
A	Team 1	2.28	2	-0.039	0.818	2.66	2	-0.212	0.818	3.04	1	-0.385	1
B	Team 4	1.66	2	0.244	0.818	1.62	4.5	0.262	0.364	1.58	1	0.28	1
C	Team 7	1.52	2	0.308	0.818	1.84	5.14	0.162	0.247	2.16	2	0.016	0.818

Table 4.10: Raw and normalized Alignment vs. Awareness scores. (Workshop 2) For Pareto Rank, lower is better. For Workshop 2, the worst (largest) Pareto Rank of any teams' design vectors was 12.

		Awareness vs. Alignment
Treatment Group	**Team Identifier**	**Final Rank**
A	Team 1	3
B	Team 4	1
C	Team 7	2

Table 4.11: Teams' final ranking based on Alignment vs. Awareness scores. (Workshop 2)

4.2.3 Workshop 3 - April 23/24, 2021

Participants

As described in Section 3.6, this workshop had 28 participants that were randomly assigned into teams of three, with four teams being assigned to treatment group A, two teams of three to group B, and two teams being assigned to treatment group C. In treatment group B, two teams, Team 5 and Team 8, ended up as teams of two randomly assigned participants due to the total number individuals in the workshop. One team, Team 5 (Group B), had no participant capable of installing and running the workshop software and therefore, being a team of two, was combined with Team 8, also being a team of two, resulting in a single team of four participants. Therefore, Group B ended up with only three teams in total, two teams of three participants and one team of four participants. Table 4.12 lists the final assignment of the resulting nine teams to treatment groups.

Treatment Group	Team Identifier	Team Size
A	Team 1	3
	Team 2	3
	Team 3	3
	Team 4	3
B	Team 5	0
	Team 6	3
	Team 7	3
	Team 8	4
C	Team 9	3
	Team 10	3

Table 4.12: Team to treatment group assignments. (Workshop 3)

Of the 28 participants, only 18 participants ($\sim$ 64%) completed the pre-team-activity survey. However, these responses came unevenly across seven of the teams (Table 4.13) resulting in only four teams having all participants complete the survey, one in each treatment group. Only 10 participants ($\sim$ 36%) completed the post-team-activity survey with zero teams having all participants complete the survey. Simulation results data was submitted by eight of the nine teams, with Team 9's data

being lost due to unrecoverable technical difficulties in its submission.

Given the lack of post-team-activity survey responses, no longitudinal statistical comparisons within or between treatments can be made. Therefore, the data analysis for this workshop is purely illustrative and includes only pre-team-activity survey response data.

Treatment Group	Team Identifier	Survey Response Rates	
		Pre-team-activity	Post-team-activity
A	Team 1	100%	67%
	Team 2	67%	0%
	Team 3	100%	33%
	Team 4	0%	67%
B	Team 6	67%	33%
	Team 7	100%	67%
	Team 8	50%	50%
C	Team 9	100%	0%
	Team 10	0%	0%

Table 4.13: Team survey response rates. (Workshop 3) Team 8 had four participants, all other teams shown had three.

Detection of Mental Models

Again, in the same manner used in Workshop 2, the raw pre- and post-team-activity survey scores for Workshop 3 were used to test the two null hypotheses $\mathbf{H_{0a}}$ and $\mathbf{H_{0b}}$ (Table 4.14). The fit of the data was also visualized for each hypothesis, by plotting its probability distribution function against the probability mass distribution of the survey data (Figure 4-9).

Both of the null hypotheses are rejected for the pre-team-activity data while for the post-team-activity-data only $\mathbf{H_{0a}}$ can be rejected (note, for Workshop 2 it was only $\mathbf{H_{0b}}$ that could be rejected for the post-team-activity data).

A sensitivity analysis for the pre-team-activity survey data was also done for alternative versions of $\mathbf{H_{0b}}$ by modifying the standard deviation of the normal distribution being used, in increments of 0.1, until the chi-square test failed to reject the alternative null hypothesis. However, the p-values of the tests never fell to a 5% significance level and began to rise once the standard deviation passed 1.6. Thus, for the purpose

Survey Dataset	Statistic	H_{0a}	H_{0b}	Alt. — $N(3, 1.6)$
Pre-team-activity	χ^2	30.2	54.1	27.9
	df	4	4	4
	(p-value)	($4.4e^{-06}$)	($5.0e^{-11}$)	($1.28e^{-05}$)
Post-team-activity	χ^2	31	3.34	18.1
	df	4	4	4
	(p-value)	($3.1e^{-06}$)	(0.50)	(0.001)

Table 4.14: Chi-square goodness of fit test statistics for pre- and post-team-activity survey data under H_{0a}, H_{0b}, and the alternative null hypothesis of a normal distribution which minimizes the chi-square test's p-value for pre-team-activity survey data. (Workshop 3) "df" stands for "degrees of freedom".

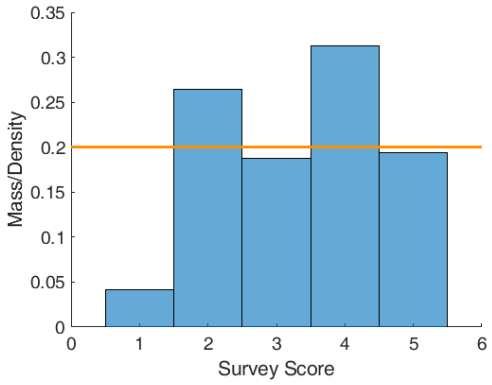
of the sensitivity analysis, the standard deviation value of 1.6 was taken to be the value which minimizes the chi-square test's p-value for the alternative null hypotheses tested.

Mental Model Alignment

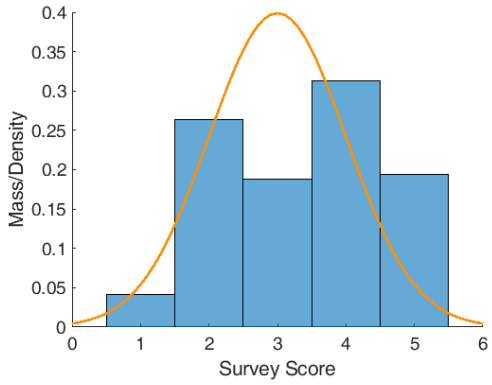
Pre-team-activity Survey Results

Following the same analysis approach used for Workshop 2, the teams were ranked based upon the average range and average spread of their aggregate desirability and feasibility scores. This gave a best-to-worst ranking of Team 1 and Team 7 (tied), Team 9, and Team 3 based upon the average range (with a lower value being better), and a best-to-worst ranking of Team 1, Team 7, Team 3, Team 9 based upon the average spread (with lower being better). In both cases, the two teams assigned to treatment group A were separated by a team from one, or both, of the other treatment groups. Point-biserial correlation coefficients were again calculated for each pairwise combination of treatment groups for each team's average range, average Teachman's index, and average spread values for both desirability and feasibility (Table 4.15).

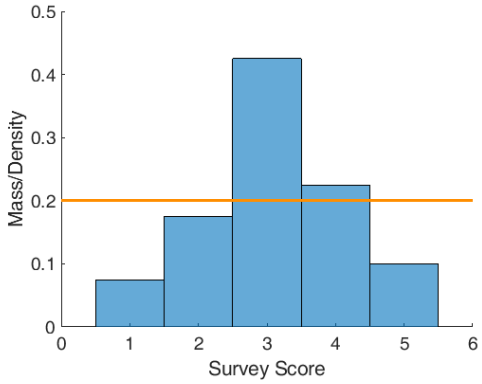
Interpretation of this data would have indicated a weak positive correlation based upon range, but weak negative correlations based upon Teachman's index and spread for the group A vs. group B comparison. Group A showed a range of weak, strong, and moderate negative correlations for all three metrics, respectively, when compared



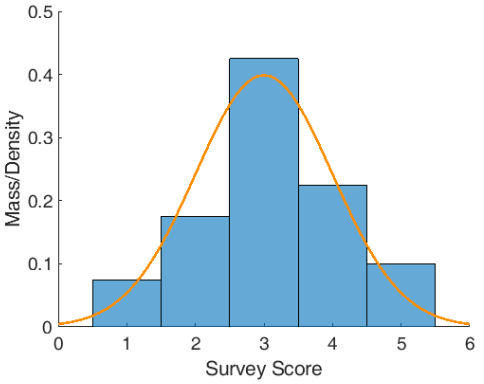
(a) Pre-team-activity survey data vs. $\mathbf{H_{0a}}$.



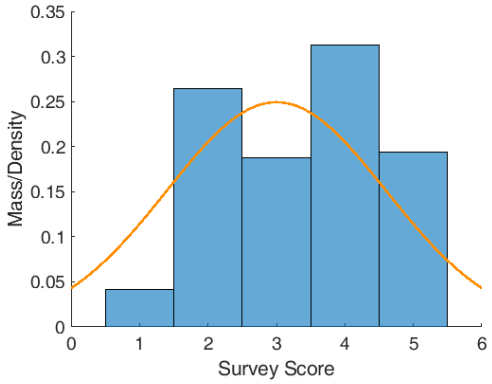
(b) Pre-team-activity survey data vs. $\mathbf{H_{0b}}$.



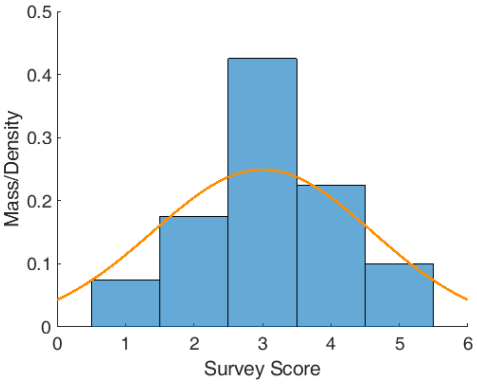
(c) Post-team-activity survey data vs. $\mathbf{H_{0a}}$.



(d) Post-team-activity survey data vs. $\mathbf{H_{0b}}$.



(e) Pre-team-activity survey data vs. Alt. $N(3, 1.6)$.



(f) Post-team-activity survey data vs. Alt. $N(3, 1.6)$.

Figure 4-9: Workshop 3 survey data probability mass distribution vs. hypothesized model probability density distribution plots.

Treatment Group Pair	Statistic	Range	Teachman's Index	Spread
A/B	ρ	0.31	-0.17	-0.07
	(p-value)	(0.51)	(0.73)	(0.89)
A/C	ρ	-0.23	-0.71	-0.59
	(p-value)	(0.63)	(0.07)	(0.17)
B/C	ρ	-0.61	-0.61	-0.42
	(p-value)	(0.3)	(0.3)	(0.51)

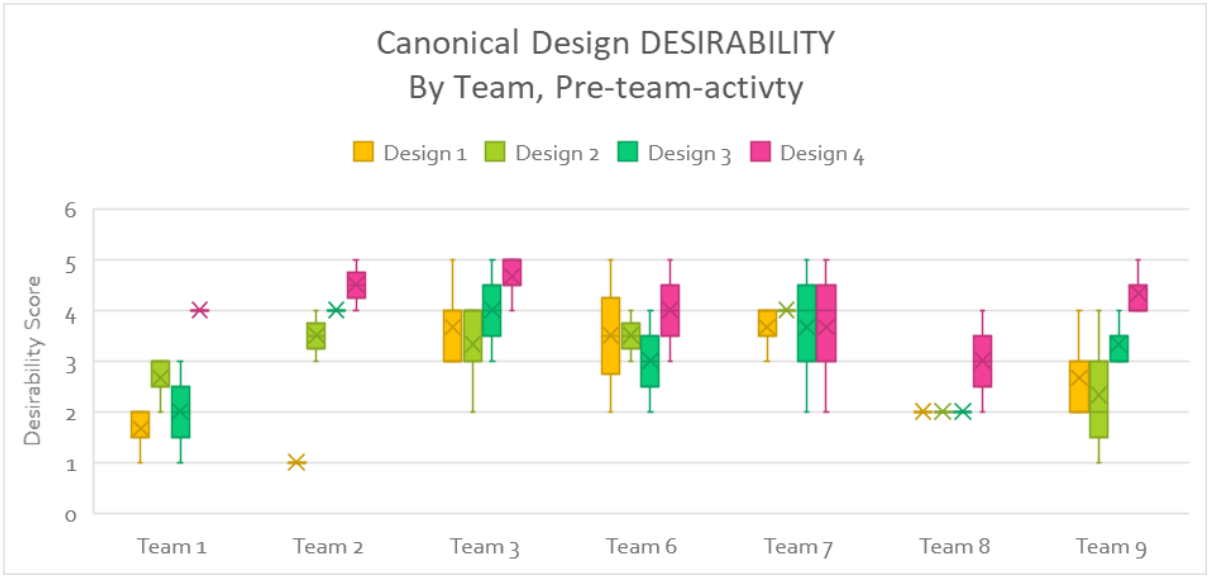
Table 4.15: Pre-team-activity survey point-biserial correlation coefficients for treatment group assignment and three metrics of mental model alignment. (Workshop 3)

to the control, group C. Similarly, group B showed a consistent negative, but in this case consistently moderate, correlation across all three metrics when compared to group C. Once again however, the p-values indicate that the null hypothesis, that the treatment group samples were drawn from a distribution having the same mean but potentially different variances, *cannot* be rejected. This is as expected due to the limited sample size of each treatment group.

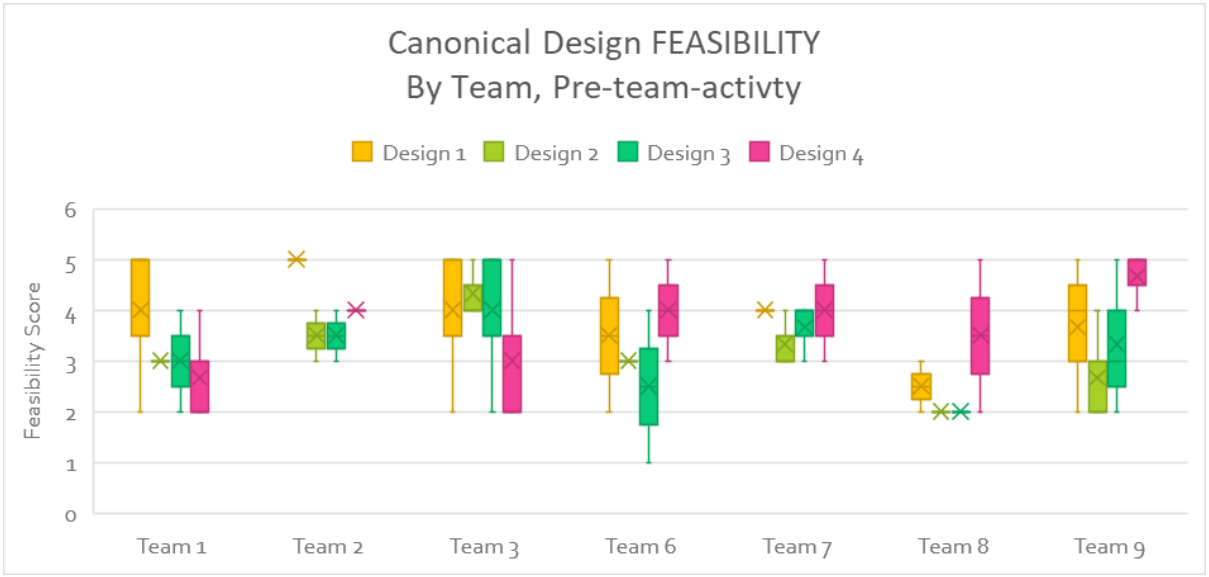
As median and sum scores of desirability and feasibility are only meaningful, for this work, in relation to the same measures post team activity, and no team had all team members complete the post-team-activity survey, these scores were not evaluated for Workshop 3.

Pre-team-activity Survey Results

As no team had all team members complete the post-team-activity survey, no evaluation of post-team-activity survey scores was done for Workshop 3.



(a) Desirability, Pre-team-activity Survey.



(b) Feasibility, Pre-team-activity Survey.

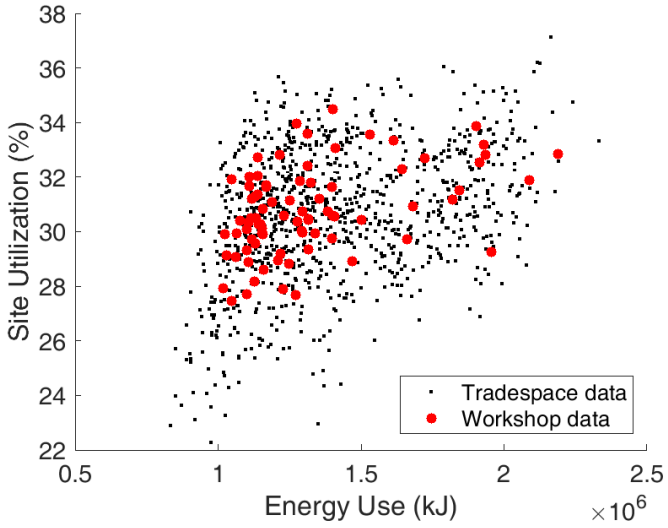
Figure 4-10: Workshop 3 canonical design evaluation survey results, by team.

	Desirability					Feasibility				
	Design 1	Design 2	Design 3	Design 4		Design 1	Design 2	Design 3	Design 4	
Treatment Group A										
Team 1										
Metric					**Average**					**Average**
Median	2	3	2	4	2.75	5	3	3	2	3.25
Range	1	1	2	0	1	3	0	2	2	1.75
Teachman's Index	0.64	0.64	1.1	0	0.59	0.64	0	1.1	0.64	0.59
Spread	0.64	0.64	2.2	0	0.87	1.91	0	2.2	1.27	1.34
Sum	5	8	6	12	7.75	12	9	9	8	9.5
Team 3										
Metric					**Average**					**Average**
Median	3	4	4	5	4	5	4	5	2	4
Range	2	2	2	1	1.75	3	1	3	3	2.5
Teachman's Index	0.64	0.64	1.1	0.64	0.75	0.64	0.64	0.64	0.64	0.64
Spread	1.27	1.27	2.20	0.64	1.34	1.91	0.64	1.91	1.91	1.59
Sum	11	10	12	14	11.75	12	13	12	9	11.5
Treatment Group B										
Team 7										
Metric					**Average**					**Average**
Median	4	4		4	4	4	3	4	4	3.75
Range	1	0	3	3	1.75	0	1	1	2	1
Teachman's Index	0.64	0	1.1	1.1	0.71	0	0.64	0.64	1.1	0.59
Spread	0.64	0	3.30	3.30	1.81	0	0.64	0.64	2.2	0.87
Sum	11	12	11	11	11.25	12	10	11	12	11.25
Treatment Group C										
Team 9										
Metric					**Average**					**Average**
Median	2	2	3	4	2.75	4	2	3	5	3.5
Range	2	3	1	1	1.75	3	2	3	1	2.25
Teachman's Index	0.64	1.1	0.64	0.64	0.75	1.1	0.64	1.1	0.64	0.87
Spread	1.27	3.3	0.64	0.64	1.46	3.30	1.27	3.30	0.64	2.13
Sum	8	7	10	13	9.5	11	8	10	14	10.75

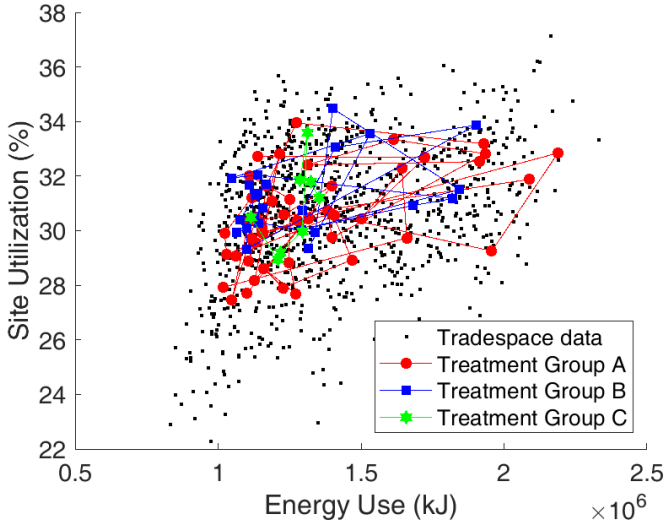
Table 4.16: Pre-team-activity canonical design survey statistics, by team. (Workshop 3)

Tradespace Exploration

Figure 4-11 shows the overall solution space exploration and design walks by treatment group for all nine teams for one pair of system evaluation metrics — Site Energy Use vs. Site Average Utilization. Similar plots for the other metrics used for analysis are shown in Figure 4-12 with the evaluation metrics for each team's exploration shown in Table 4.17. Figure 4-13, Figure 4-14, and Figure 4-15 show per-group plots.

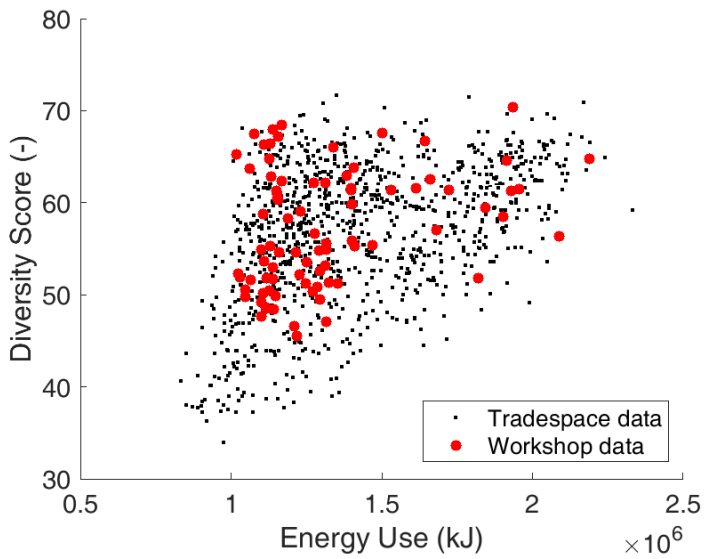


(a) Solution Space exploration.

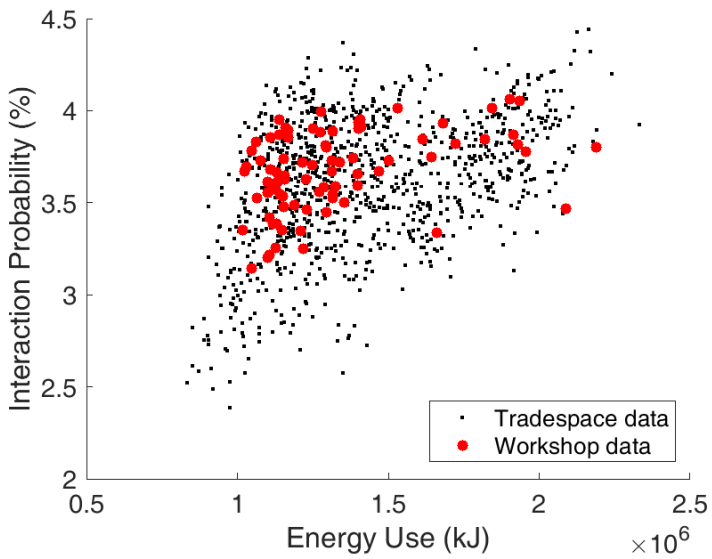


(b) Design walks for all three treatment groups.

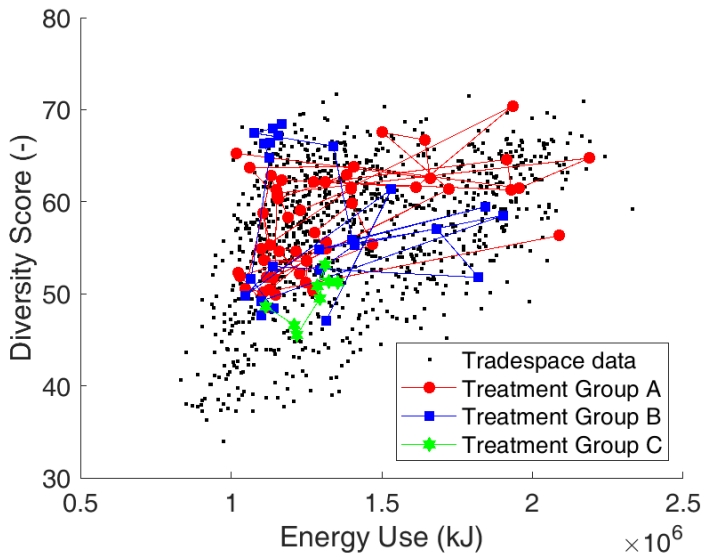
Figure 4-11: Workshop 3 Tradespace Diagrams, Site Energy Use vs. Site Average Utilization.



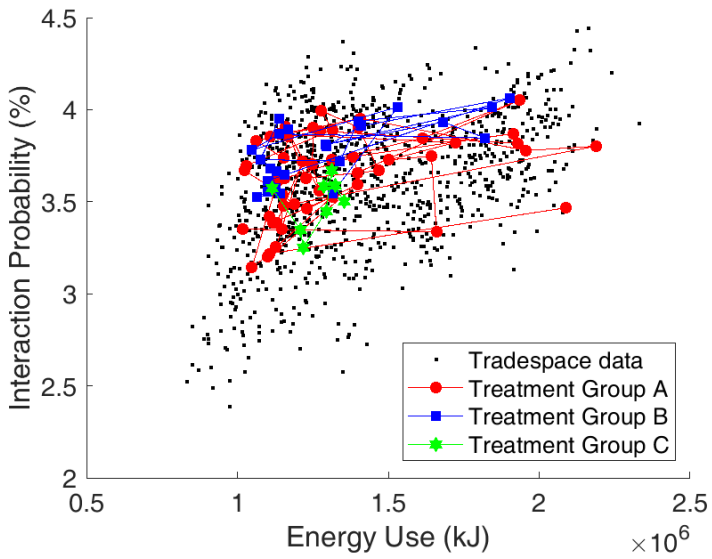
(a) Energy vs. Diversity Tradespace, Overlay.



(b) Energy vs. Interactions Tradespace, Overlay.

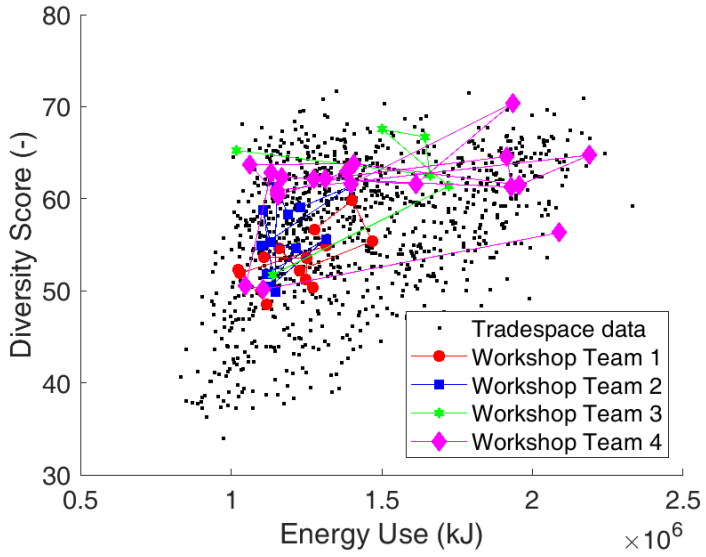


(c) Energy vs. Diversity Tradespace, Design Walk.

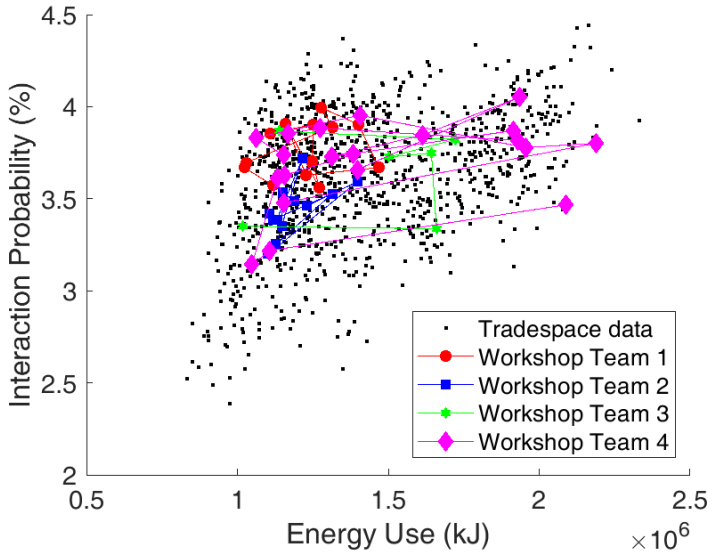


(d) Energy vs. Interactions Tradespace, Design Walk.

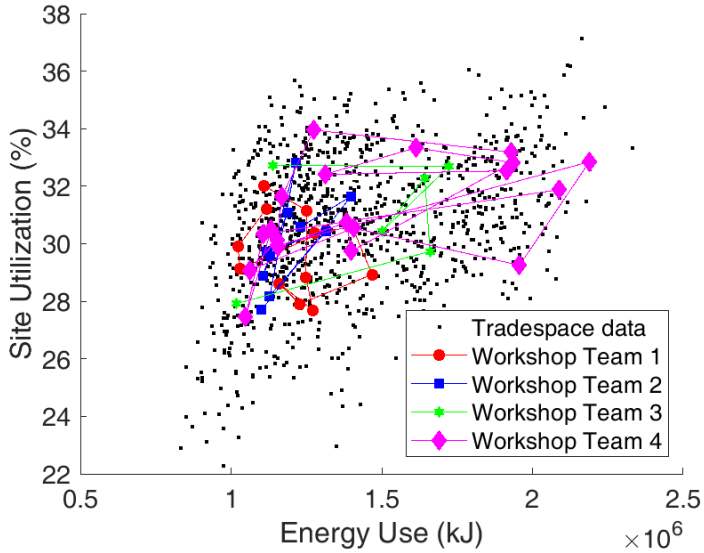
Figure 4-12: Workshop 3 Tradespace Diagrams.



(a) Energy vs. Diversity Tradespace, Treatment Group A.

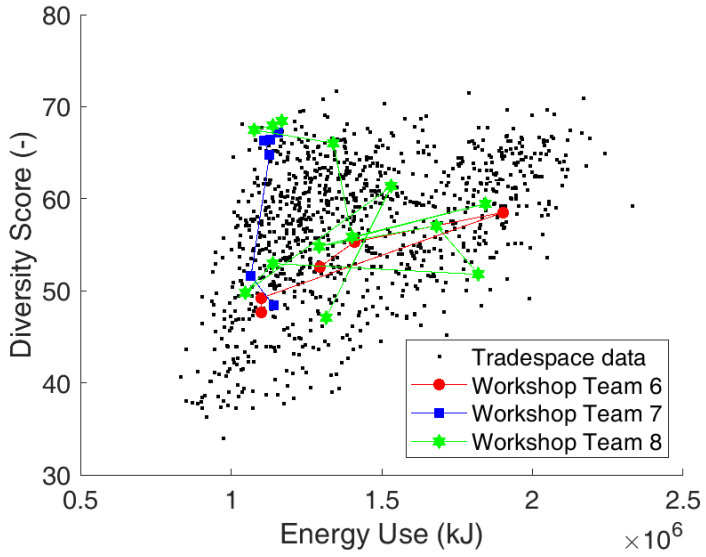


(b) Energy vs. Interactions Tradespace, Treatment Group A.

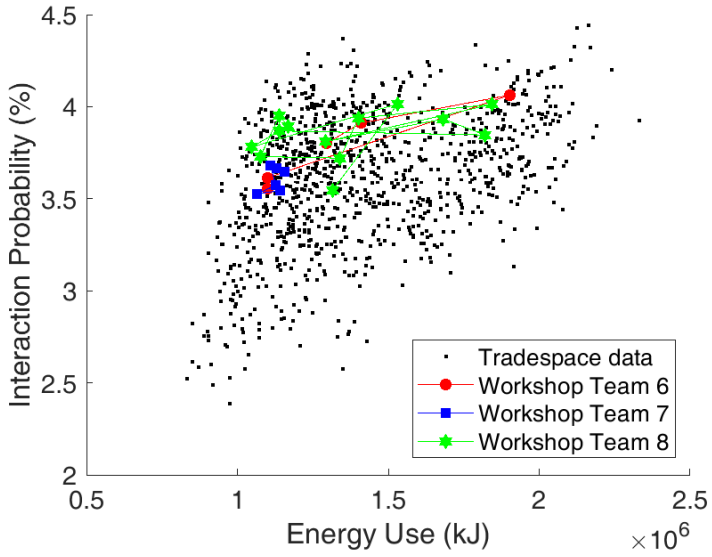


(c) Energy vs. Utilization Tradespace, Treatment Group A.

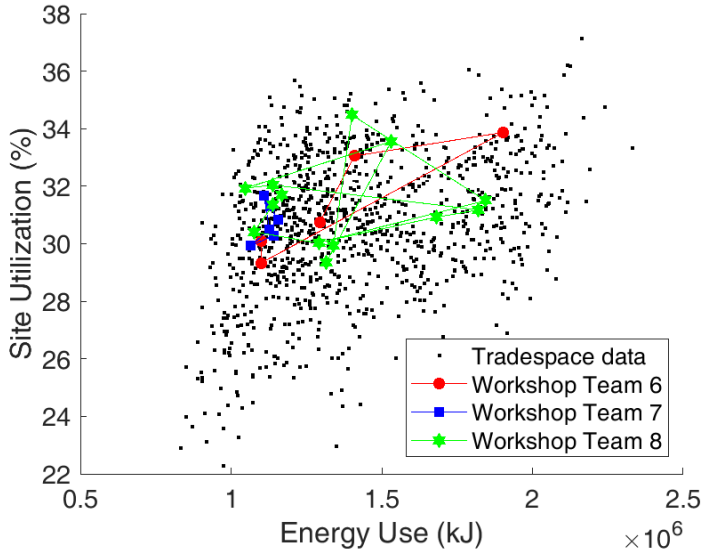
Figure 4-13: Workshop 3 Design Walk Diagrams, Treatment Group A.



(a) Energy vs. Diversity Tradespace, Treatment Group B.

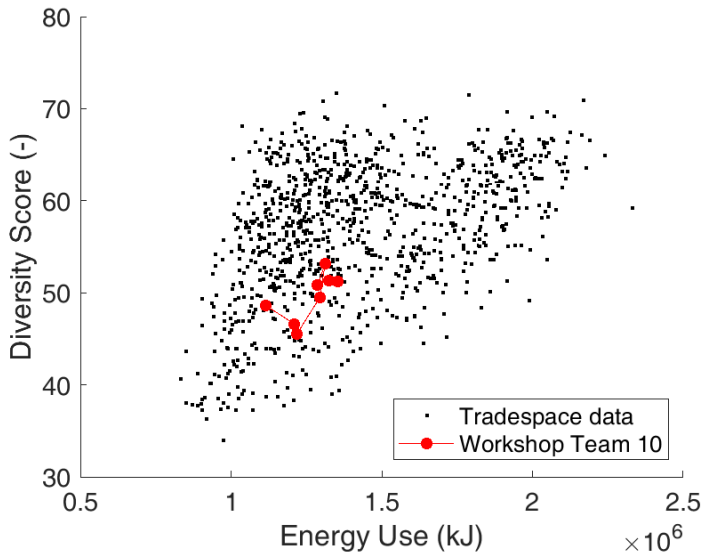


(b) Energy vs. Interactions Tradespace, Treatment Group B.

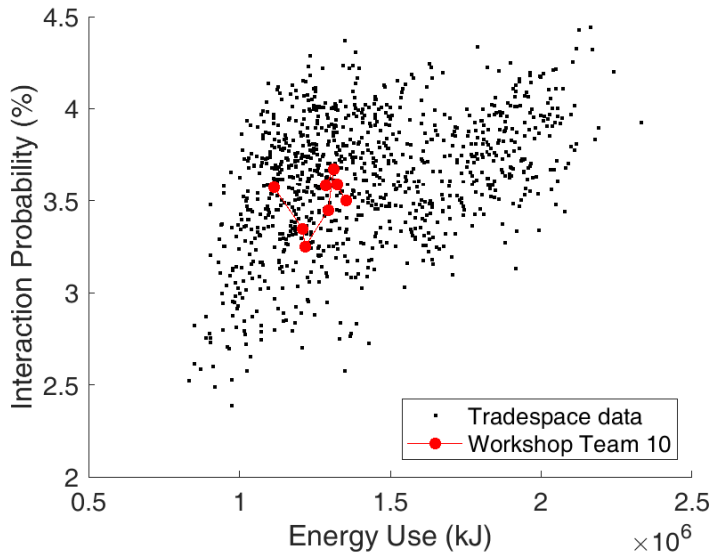


(c) Energy vs. Utilization Tradespace, Treatment Group B.

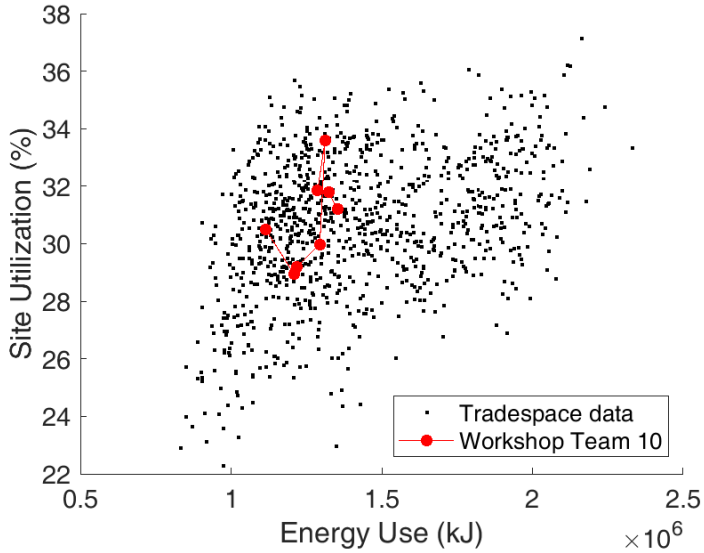
Figure 4-14: Workshop 3 Design Walk Diagrams, Treatment Group B.



(a) Energy vs. Diversity Tradespace, Treatment Group C.



(b) Energy vs. Interactions Tradespace, Treatment Group C.



(c) Energy vs. Utilization Tradespace, Treatment Group C.

Figure 4-15: Workshop 3 Design Walk Diagrams, Treatment Group C.

		Solution Space Exploration Statistics							Problem Space Exploration Statistics				
		Designs		Inter-design Changes					Pareto Rank				
Treatment Group	**Team Identifier**	**Total**	**Unique**	**Min.**	**Max.**	**Avg.**	**S.D.**	**Efficiency**	**Min.**	**Max.**	**Avg.**	**S.D.**	**Preferred Design**
A	Team 1	13	13	1	2	1.54	0.47	0.21	4	11	6.77	2.45	?
	Team 2	12	12	1	5	3.09	1.50	0.41	5	10	7.58	1.55	N/R
	Team 3	6	6	4	12	7	2.83	0.39	1	10	5.33	3.14	N/R
	Team 4	20	20	1	6	3	1.62	0.45	2	12	6.25	2.34	6
B	Team 6	5	5	1	3	1.5	0.87	0.11	5	8	6.6	1.02	N/R
	Team 7	6	6	1	5	2.8	1.83	0.21	2	8	3.83	2.12	2
	Team 8	13	13	1	7	3.5	1.66	0.36	2	11	5.31	2.95	?
C	Team 9	*	*	*	*	*	*	*	*	*	*	*	N/R
	Team 10	8	3	0	6	1	2.07	0.16	5	11	9	2.06	N/R

Table 4.17: Tradespace exploration statistics. (Workshop 3) Team 9 suffered from technical difficulty submitting their simulation results leading to data loss. For Pareto Rank, lower is better. For Preferred Design, "?" indicates a lack of agreement between team member responses in the survey data while "N/R" indicates no response data was available.

Design Walk Evaluation

The design walks between the three treatment groups appear visually distinct from one another. However, it is worth noting that treatment group A had twice as many teams (four) as treatment group B and C (two each), and that only half of treatment group C's design walk data (one team's) was available with the other half lost due to technical difficulties in its end-of-workshop submission. Therefore, the appearance of the visual distinctions had limited meaning.

Similarly, the meaning of any differences between the tradespace exploration metrics (Table 4.17) was also limited due to the small sample size. Two-sample t-tests were run pair-wise between each treatment group (Team 8 from group B, being a team of four, was excluded) for each tradespace exploration metric, in order to check for statistical significance between the groups' metric values (Table 4.18). All but one of the t-tests failed to reject the null hypotheses that the values from two different treatment groups were sampled from a distribution with the same mean but potentially different variances. The t-test between treatment group A's and B's maximum Pareto Rank metric showed a statistically meaningful difference in range, despite the small sample size.

Following the t-tests, correlation coefficients were calculated using Kendall's τ, between the ranking of teams using their mental model alignment metrics, previously discussed, and their tradespace exploration metrics (Table 4.19). As only pre-team-activity survey data had complete team responses, calculation of correlations was limited to being done only for the ranking produced from these scores. Additionally, despite having had all members complete the pre-team-activity survey, data for Team 9 was not included in the calculation of correlations, despite their responses having impacted the team rankings, due to the previously mentioned loss of their simulation results data.

Based upon the Kendall's τ correlation coefficients, many metrics appeared to have perfect or strong correlations, both positive and negative. However, as with Workshop 2, enough data was not present to make any claims of statistical significance

		Solution Space Exploration Statistics							Problem Space Exploration Statistics				
		Designs		Inter-design Changes					Pareto Rank				
Treatment Group Pair	**Statistic**	**Total**	**Unique**	**Min.**	**Max.**	**Avg.**	**S.D.**	**Efficiency**	**Min.**	**Max.**	**Avg.**	**S.D.**	**Preferred Design**
	T	1.68	1.68	0.67	0.70	0.84	0.33	2.40	-0.30	3.83	1.15	1.35	
A/B	df	4	4	4	2	4	4	4	4	4	4	4	-
	(p-value)	(0.17)	(0.17)	(0.54)	(0.52)	(0.45)	(0.76)	(0.74)	(0.78)	(0.02)	(0.32)	(0.25)	
	T	0.74	1.52	1.04	0.05	1.02	-0.43	1.73	-0.98	-0.23	-2.39	0.43	
A/C	df	3	3	3	3	3	3	3	3	3	3	3	-
	(p-value)	(0.51)	(0.23)	(0.37)	(0.96)	(0.38)	(0.70)	(0.18)	(0.40)	(0.83)	(0.10)	(0.70)	
	T	-2.87	2.87	Inf	-1.15	1.02	-0.87	0	-0.58	Inf	-1.58	-0.51	
B/C	df	1	1	1	1	1	1	1	1	1	1	1	-
	(p-value)	(0.21)	(0.21)	(0)*	(0.45)	(0.49)	(0.55)	(1)	(0.67)	(0)*	(0.36)	(0.70)	

Table 4.18: Pairwise two-sample t-test for equal means test statistics for tradespace exploration metrics. (Workshop 3) "df" stands for "degrees of freedom". t-tests were not run for Preferred Design as data from post-team-activity surveys was not present or incomplete for all teams. *As group C has only a single sample, some inconsistencies occur with t-test statistics that are not meaningful.

with all p-values well above a 0.05 threshold. No demonstrative interpretation, nor any comparison with Workshop 2 data, was made.

Alignment vs. Awareness Ranking

Due to the lack of post-team-activity survey data from Workshop 3, post-team-activity alignment scores and Pareto Rank for preferred designs could not be calculated. Thus neither could any intermediate point within the Alignment vs. Awareness space as was done in Workshop 2. However, each team's average spread, based on pre-team-activity survey desirability and feasibility scores, was calculated and normalized along with the normalized Pareto Rank of each team's first simulated design vector. These values (Table 4.17) were used to plot the teams' starting positions within the Alignment vs. Awareness space.

No final ranking of teams based on their position within the Alignment vs. Awareness space was made.

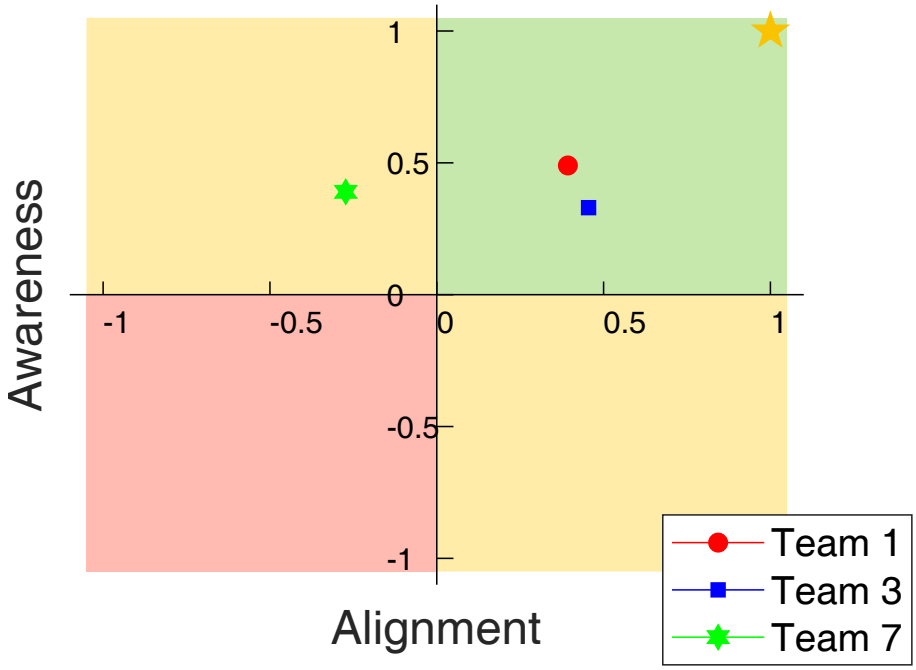


Figure 4-16: Plot of Workshop 3 Alignment vs. Awareness starting points. Each team's position is based on its normalized alignment and awareness metrics. The yellow star represents the Utopia point $(1, 1)$ against which teams would have been ranked.

		Solution Space Exploration Statistics							Problem Space Exploration Statistics				
		Designs		Inter-design Changes					Pareto Rank				
Ranking Method	**Statistic**	**Total**	**Unique**	**Min.**	**Max.**	**Avg.**	**S.D.**	**Efficiency**	**Min.**	**Max.**	**Avg.**	**S.D.**	**Preferred Design**
Range	τ	-0.5	-0.5	1	0.82	0.82	0.82	1	-0.82	0	0	0.82	-
	(p-value)	(1)	(1)	(0.67)	(0.67)	(0.67)	(0.67)	(0.67)	(0.67)	(1)	(1)	(0.67)	
Spread	τ	-0.82	-0.82	0.82	1	1	1	0.82	-1	-0.33	-0.33	0.33	-
	(p-value)	(0.67)	(0.67)	(0.67)	(0.33)	(0.33)	(0.33)	(0.67)	(0.33)	(1)	(1)	(1)	

Table 4.19: Pre-team-activity survey Kendall's tau correlation coefficients for tradespace exploration statistics with teams ranked by average range or spread of diversity and feasibility scores. (Workshop 3) Calculations based on ranking and metrics from teams 1, 3, and 7 only. Correlation to Preferred Design could not be calculated due to lack of data.

Treatment Group	Team Identifier	Pre Team Activity			
		Average Spread	Starting Pareto Rank	Normalized Spread	Normalized Pareto Rank
A	Team 1	1.11	6	0.494	0.363
	Team 3	1.47	4	0.330	0.455
B	Team 7	1.34	8	0.390	-0.273
C	Team 9	1.79	*	0.185	*

Table 4.20: Raw and normalized Alignment vs. Awareness scores. (Workshop 3) For Pareto Rank, lower is better. For Workshop 3, the worst (largest) Pareto Rank of any teams' design vectors was 12. *Simulation results data for Team 9 was lost due to technical difficulties during its submission, therefore, the Pareto Rank values for their designs could not be calculated.

4.3 Summary

This chapter presented a method for quantifying the alignment of mental models and awareness of systemic effect among a team of stakeholders exploring a systems model. The method was demonstrated for a series of experiments conducted as part of this work. In support of the argument that alignment of mental models can be assessed through subjective evaluations of design, data regarding each team member's view of the desirability and feasibility of a set of canonical designs was collected, compared against hypothesized generative models, and analyzed for both degree of disparity and level of agreement within the team. Additionally, data from each team's design walk, generated through their use of a systems model, was analyzed in order to quantify the influence of the topology of exploration, in both the solution and problem spaces, on emergent collective awareness within the team. Evidence of the influence of team alignment upon exploration of a system's tradespace and the influence of exploration upon a team's mental model alignment was then investigated. Finally, an approach for ranking teams on their alignment and awareness, as well as visualizing the evolution of these aspects of a team over time, was demonstrated.

Chapter 5

Interpretation

This chapter takes the analysis of experimental data from Chapter 4 and discusses findings that may be drawn from it, as well limitations. Limitations are further discussed in relation to the experimental design and model problem developed as part of this work, the series of teamwork workshop experiments, the workshop software underpinning those events, and the approach taken to the experimental results and analysis of data. Insights that emerged over the course of this research are then presented, followed by next steps. Finally, conclusions regarding the position of this work within the existing body of research, its contribution in the field, and future work are made.

5.1 Discussion

No claims can be made regarding the relationships between the diversity of a team's mental models and their exploration of a systems model tradespace by drawing upon the experimental results of this work, or their analysis as presented in Chapter 4. The inability to meaningfully interpret the results previously presented is exclusively due to the lack of complete and sufficient datasets, making statistically significant comparisons within and between the treatment groups impossible. However, this work does suggest that it may have successfully detected and measured the mental models of individuals on a team, role-playing as expert stakeholders.

5.1.1 Assessment of Research Questions and Hypotheses

This work set forth the following research questions and hypotheses, given a team of stakeholders evaluating designs for a sociotechnical system using a systems model —

> **Can we detect and measure the diversity of mental models within a team?**

Section 3.3 argued that desirability and feasibility are a meaningful proxy for mental models under the conditions of this work. In Chapter 4, the analysis of pre- and post-team-activity survey desirability and feasibility scores showed statistically meaningful rejection of the two following null hypotheses explored for the pre-team-activity survey data:

H_{0a} (Random Selection) : The desirability and feasibility scores are sampled from an underlying distribution that is a discrete uniform random distribution taking values in the range $[1, 5]$.

H_{0b} (Central Tendency) : The desirability and feasibility scores are sampled from an underlying discrete distribution for which the cumulative distribution function approximates that of a normal distribution of the form $N(3, 1)$.

These null hypotheses represented two potential models from which one could logically assume participants might generate their data if not doing so based upon a mental model of the system being evaluated. Namely, a random selection from among the five levels of the Likert item questions, and, a central tendency bias whereby a normally distributed selection occurs centered upon the middle level of the Likert item questions[44]. However, neither of these two null hypotheses held for the pre-team-activity survey data. Rather, for the latter null hypothesis, a normal distribution with a larger standard deviation was required in order to fail to reject the null hypothesis. This indicates a greater deviation away from the middle level of the Likert item questions, thus conflicting with the central tendency bias. It is argued that this can be attributed to a meaningful preference by the participants, informed by their mental model of the system.

Interestingly, for the post-team-activity survey data, the first null hypothesis cannot be rejected in the case of Workshop 2, and the second null hypothesis cannot be rejected for Workshop 3. Unfortunately, due to the lack of intentional experimental design for adequate treatments and control to test this phenomena, no specific claims around this observation can be made. One might speculate, however, that some stimulus during the team activity, or attribute of the activity itself, resulted in confusion, disagreement, and/or incomplete information causing uncertainty. Perhaps this uncertainty resulted, in one case, in participants' scores appearing almost random, and in the other, participants shifting toward a safe, central evaluation requiring little commitment. In the case of Workshop 2, the more uniform distribution of scores is also evidenced in the increase in teams' spread values.

Therefore, this work suggests to have shown the possible detection and measurement of the mental models of individuals, and thus measurement of diversity of those mental models on a team, through the use of qualitative self-assessment of a system's desirability and feasibility as defined in Chapter 3. However, no further claims about the meaningfulness of the treatments, as developed in the experimental design, on the diversity of mental models can be being made.

Does diversity of stakeholder mental models within a team affect that team's performance?

Hypothesis 1. *Teams having a greater diversity of mental models of system architecture among stakeholders are more likely to be high performing teams.*

No evidence in support of this hypothesis was demonstrated nor may it be rejected due to insufficient data collection and the subsequent, demonstrative, analysis of that limited data.

Given the metrics explored for representing diversity of mental models on a team (range, Teachman's index, and spread) and the treatments selected for attempting to create distinguishable degrees of alignment between stakeholder mental models, no statistically significant difference was detected between treatment groups. However, with only a single team's worth of samples from treatment groups B and C for the

second and third workshops, this lack of difference is not likely to be meaningful. Similarly, for both the second and third workshop, the lack of statistical significance in the rank correlations of tradespace exploration statistics, preferred design in particular, is influenced by the small number samples per variable more than the data itself. With a larger set of samples, one would not only be able to test for the rank correlations based upon the chosen mental model alignment measures, but also make a meaningful direct test for difference between the treatment groups tradespace exploration statistics. For Workshop 3, a direct test for differences between treatment group tradespace exploration statistics was demonstrated and showed a statistically significant difference between the maxium Pareto Rank encountered by groups A and B. Unfortunately, while an interesting finding, without any usable post-team-activity survey data being generated by Workshop 3, no insight may be gained.

Can one detect a pattern of model exploration by a team that is indicative of performance resulting from the emergence of collective awareness of systemic effect?

Hypothesis 2. *High performing teams will exhibit a more diverse exploration of architectures within a systems model.*

This hypothesis cannot be confirmed nor can it be rejected due to the limited amount of usable data gathered during the series of experiments and the approach taken in the data's analysis.

In addition to the limitations for measuring performance based on insufficient experimental data, previously discussed, the approach to data analysis similarly made testing of this hypothesis infeasible. Specifically, the calculation and use of a notional, rather than actual, Pareto Rank meant that neither a team's real performance nor its real Pareto-Rank-derived tradespace exploration statistics were determined. Thus, a meaningful comparison between performance and exploration was not possible.

Does the emergence of collective awareness of systemic effect lead to greater post-system-evaluation alignment of a team's mental models?

Hypothesis 3. *A more diverse exploration of architectures within a systems model will result in greater alignment among stakeholder mental models.*

Given the very small sample size of completed pre- and post-team-activity assessments of stakeholder mental models, and the previously mentioned limitations from data analysis, **H3** also cannot be confirmed nor rejected.

In Chapter 4, evidence was presented that the mental models of a team using a systems model did change over the course of a team workshop. As alluded to previously, given the experimental design used, direct attribution to the use of a systems model cannot be made as no treatments or appropriate control to test for this specific phenomena were included in the experimental design. An example alternative interpretation could be that just the team discussions by stakeholders and/or the sharing of knowledge on stakeholder goals and value functions drove the shifts in mental models.

Additionally, it was unclear if there were any correlation between the general attributes of exploration, or its pattern, and the shifts in mental models. Specifically, the dataset size made correlations between mental model alignment and the broad set of tradespace exploration metrics statistically meaningless. The relationship between treatment group and directionality of mental model changes is also unclear from the data gathered and analyzed. For example, in Workshop 2, where pre- and post-survey data is available, Team 1 (treatment group A) and Team 7 (treatment group C) both exhibited a shift away from alignment as it has been defined in this work, while Team 4 (treatment group B) showed evidence of greater alignment. However, as discussed elsewhere, Team 1 had a dissenting team member which significantly increased the spread of the team's feasibility scores. This highlights the impact of the limited dataset size as, while we can detect these shifts, there are no other teams within the treatment group available for comparison.

5.1.2 Limitations

Beyond the limitations in interpreting the experimental results and their analysis, the limitations of this work across its experimental design, the model problem selected, the workshop — its structure, implementation, and execution, the workshop software, and the results and analysis of data more broadly, were identified.

Experimental Design

The single greatest limitation within the experimental design was the failure to include adequate consideration for distinguishing what attribute(s) of the stimulus (team role-play activity) influence mental model changes. Specifically, during the team activity, individual participants encountered new information regarding the model problem, the cognitive collision of their priors with their role-play persona, the establishment of a new team with diverse mental models, the use of a systems model for exploration of design alternatives, and work on the actual team goal that necessitated discussion, coordination, and negotiation. Combined, these components of the team activity represent individually testable elements of influence, the evaluation of which would have required an alternate series of experiments utilizing a subset of features from the experimental design as used in this work.

A second element of the experimental design that limited this work was the use of individual Likert items for evaluating mental models. A tradeoff exists between the scope of the pre- and post-team-activity surveys and greater insight and analytical value gained through the use of a true Likert scale. Additionally, the lack of questions focusing specifically on the dynamics, assumptions, and constraints present in a participant's mental model, limited greater analysis and comparison between individuals and over time for a given individual. This includes the inability to directly test for the composition or accuracy of an individual's mental model, or the alignment os, regardless of accuracy, and change in its components within a team, over time.

Related to the use of Likert items was the use and definition of "desirability" and "feasibility". While the definitional intent remains sound, better clarity through dif-

ferent terminology with less ambiguity and interpretability could have been achieved. As an example of the issue, a natural interpretation of the two terms chosen could have led some participants, whom may not have had a firm handle on their meaning per the experimental design, to conclude that an infeasible design cannot be desirable from an engineering perspective.

Finally, the selection of Pareto Rank for evaluation of team exploration and performance tightly couples experimental outcomes to the limitations of its use seen in both model design, and experimental implementation and execution.

Model Problem

The model problem selected by this work suffered from limitations due to the selection and definition of the stakeholder roles, the metrics of design evaluation at a team level, and the symmetry of designs within the solution space to problem space mapping.

It is unclear whether the selection and definition of the stakeholder roles provide an adequate expression of the goals and value functions of each role such that a clear, consistent signal on the expectation of system evaluation across participants could be established. This is not to say that awareness and understanding of the signal by all participants is requisite, rather, it is simply that the presence of a such a signal regarding the relationship between an evaluation metric and the individual and team goals is necessary. The variety in response of participants to such signals is an interesting topic for research in and of itself. Thus whether or not treatment response was adequately controlled is questionable as is whether the intended tensions and alignments between the stakeholder roles played out as anticipated. This is particularly true in light of the different workshop implementation decisions made between explicit and implicit communication of system evaluation criteria for individual and team goals, respectively. Additionally, even if the goals and value functions of the stakeholder roles were perceived and understood, there is a question of whether participants, in their stakeholder persona, felt as though they had agency through the systems model to engage the problem in a way that was meaningful. It may well be that the measurement, reporting, or emergent dynamics of the evaluation metrics

offered little to some stakeholder roles such that teams ended up focusing on only one or two metrics, relevant to fewer of the stakeholders, and/or where little tension existed.

Related to the limitations just mentioned is the question around whether an objective Pareto Rank can be established and used, given the way the shared team goal, in particular, and the stakeholder value functions and goals, were defined. It is unclear, as detection was not attempted, whether the perspective of each of the roles, on the various performance outcomes for a design, were, in aggregate, aligned with each other. Or similarly, whether teams were generally aligned with the method of performance outcome evaluation used in data analysis. This then somewhat calls into question the meaning of the Pareto Ranks calculated and used in this work. Manandhar et al.[8] also encountered a similar issue during their work, where different teams took different perspectives on whether their "walking time" metric was to be maximized or minimized. A descoping of the metrics used for evaluation of the designs and the explicit setting of design evaluation criteria for the team goal, would make the fundamental phenomena in question more easily detected. Indeed, this latter approach may be a particularly valuable lever, not utilized in this work, for shaping the mental model alignment of teams.

Lastly, for the model problem development, the systems model developed had on the order of 1.9 trillion different design vectors. This pushes the reasonable scale for problems worth using in a short, team workshop experiments. It represents a serious computational challenge during later analysis, not to mention the cognitive challenge for the humans participants approaching the engineering problem. This alone may be a limitation of the usefulness of the model problem in evaluating the desired phenomena, however, an additional consideration is the number of equivalent designs represented in the full solutions space. The scale of the problem is actually more approachable once one considers the symmetries represented in the systems model itself. For example, given the allocation of the entire population to one of the residence spaces on level 4 of the systems model, the resulting objective vector will be equivalent (ignoring differences due to the non-deterministic computation) to the allocation of

the entire population to any one of the other three residences on that level, assuming no other systems model changes. Similarly, other equivalent designs are possible with more complex allocations of the sub-populations "mirrored" over the lateral or longitudinal axis of the systems model. This means that there exists a subspace of the complete solution space that will map to all points in the problem space. Computation of this subspace and its use in analysis was not taken advantage of in this work. Future work should try to remain aware of such design outcomes in their model problems and either avoid them (break symmetries) or otherwise intentionally take advantage of them in some way (compute the mapping for analysis but leave the humans to contend with the full solution space problem).

Workshops

The limitations of the experimental workshops are classified into three main categories (i) structure, (ii) implementation, and (iii) execution.

Structure

Overall the structure used for the workshops worked well to support the experimental design. However, three areas of improvement were identified through the series of workshops held.

First, the pre/post survey aspect of the workshop structure leaves open the opportunity for a large investment of effort for very little usable data in return. This is compounded by the need for all members of a team to complete the post-team-activity survey in order for the data from any member of the team to be usable. This was seen during Workshop 3 where no team fully completed the post-team-activity survey. Additionally, the separation between data submission and survey response made for additional steps where teams could "drop out" of the process or where data loss from one step could result in data from the other steps not being usable.

Second, the structure of the workshop provided no mechanisms by which the roles within a treatment group could align their mental models between themselves. While this is not necessarily a limitation per se, it does make some comparisons more difficult

or meaningless. For example, it is not possible to look at the absolute changes between all stakeholders of a given role, either within, or across, treatment groups, over time. Similarly, as each team in treatment group A had a separate discussion before exploration of the systems model, the collision of the stakeholder roles, interpreted in slightly different ways by different individuals, likely resulted in teams having different perspectives and goals going into the systems model exploration. While realistic, it does add noise to the detection of differences caused by the treatments themselves. The use of a separate discussion period during which all stakeholders of the same role gathered to discuss and align on their understanding of the role was considered, but ultimately not included as part of this work.

Third, the use of a systems model walk-through was absolutely necessary but ultimately insufficient for the complexity of the software and the engineering problem the teams were given. The use of an alternative toy problem, using the same interface, and perhaps the same set of decisions and evaluation metrics (or some subset thereof) would have been a much more robust approach. This would have allowed teams to become familiar with the software and its use, ask specific questions regarding its use and behavior, and ultimately lead to a more rapid establishment of tradespace exploration within the main team activity, all without the risk of teams learning or establishing an awareness of systemic effect in the experimental systems model before hand.

Implementation

The most significant limitation to the selected implementation of the workshop structure was the duration of time allotted to each event. The approach to using a systems model for analyzing a complex engineering problem is likely better suited to a multi-day workshop rather than an all-in-one, single shot, one-and-a-half to two-and-a-half hour event. It would have been valuable to have structured the workshop such that participants had sufficient time to understand and ask questions regarding the role-play scenario, background information, and their stakeholder roles; work with the workshop software to become familiar with its interface and use; and to have

the time necessary for the team to develop a cadence for exploration, learning, and development of awareness, before having to select a preferred solution.

Another significant limitation that resulted from the workshop as implemented was the inability to openly discuss team instructions with all participants due to the inclusion of all three treatment groups in one workshop. This was particularly noticeable due to the use of Zoom where it was more difficult to repeatedly re-group workshop participants in different ways while also maintaining separation of treatments and representation of the various roles. A related limitation that stems from the same implementation choice was seen during team formation when the total number of participants was not a multiple of three, or when participants needed to be added/dropped from teams. In these cases, it became difficult to ensure that all treatment groups would have at least one team of three, consistent members for the entire workshop. Instead, choosing to run separate workshops per treatment would allow for a more easily managed event, and isolation of unusable data such that appropriate sample sizes can be made for each treatment. Obviously, this would then require additional care in making sure the treatment group in each workshop received the same experience. Although initially considered, this need for consistency was ultimately the reason the single treatment workshop approach was not taken.

The approach to team instruction was also a limiting factor in the effectiveness of workshop implementation. As Bowers[14] noted, "Although team members receive instructions and practice, the mission usually requires a great deal of situational awareness, planning, and decision making by the crews for performance to be successful." Without adequate instruction or practice, teams may struggle to establish those elements of awareness, planning, and decision making, necessary for success. This limitation came in several forms including the lack including any form of instruction or training on team processes or specialized tasks for each of the three team members, the complexity and difficulty level in the instructional documents, and the level of organization for the provided workshop materials. The latter two elements were compounded by the fact that there were multiple treatments per workshop. In particular, the treatment group C (control) teams were especially confused by the

purpose and use of the stakeholder profiles that were discussed in the general workshop background and inclusion in the workshop materials those teams were provided. Finally, the approach to instruction followed no specific team training guidelines but was provided on a more ad-hoc basis.

Logistics and Execution

Two major limitations were identified in the logistics and execution of the workshop implementations, along with a handful of more specific details that could be improved in future workshop iterations.

The first major logistical limitation encountered during the workshops was the limited amount of time that teams had for both the main team activity, and, subsequently the debrief period, during which, post-team-activity surveys were to be completed. Beyond having an extended duration workshop, an improvement in managing the facilitation of the post-stimulus survey, data collection, and verification would have lead to better experimental data availability.

The second logistical limitation was encountered during Workshop 3 but was not due to the need to present parts of the workshop instruction in Japanese, which was well handled by Dr. Bryan Moser. Instead, the limitation was caused by the interaction between the use of the Zoom web conferencing system, the workshop structure, and the number of participants in the workshop. During team formation, there were not enough trained workshop support personnel available to quickly guide the teams in each treatment group through the instructional materials. Instead, each team needed to wait while the workshop support personnel moved between them and attempted to provide consistent one-on-one instruction, as well as answer questions and provide technical support with the workshop software.

The smaller limitations in execution almost all exclusively came down to insufficiently clear instruction leading into the team activity. Areas where the intended delivery of information could have been improved include the explanation of the teams' higher order, shared goal; the criteria for evaluating success on the shared goal; whether survey responses were to be provided from the perspective of the in-

dividual, their role-play persona, or the team; and what was meant by selecting a "preferred" design. Lack of these elements lead to greater confusion among participants and teams throughout the workshop, potential disparities in understanding how performance of the team would be assessed, and inconsistency in survey responses between team members. For all of these items, better execution would have lead to more data of better quality due to an increase in total and usable survey responses, less noise within those responses, and greater focus on tradespace exploration by the teams.

Workshop Software

Very few issues were encountered in the use of the workshop software provided to the participants. All teams, with the exception of one team in Workshop 3, were able to have at least one participant install and run the software without issue. A number of limitations with the software were uncovered during the three workshop events, some of which indicate potentially useful features for future versions.

One of the known workshop software limitations which did not cause any issues during the workshop was the lack of constraints on design variables. Specifically, there were no constraints regarding the assignment of sub-populations to residences nor on the assignment of activities to spaces.

In the case of sub-population assignment, the lack of constraints took two forms. First, there were no limitations on the number of individuals assigned to a given residence. In fact, the systems model was designed such that each space could accommodate the entire population of the Star City village at once, partly to avoid issues of over allocation in the residences. Second, there had been an intent to include a user interface constraint to prevent any sub-population from not being allocated. However, after Workshop 3 it was discovered that this constraint did not function under all patterns of user interface use and, additionally, it was possible to select a design with no allocation of *any* sub-population and yet still successfully execute a simulation *and* have results that indicated the allocation of a small number of people. This was clearly a software bug. It is believed, but was not confirmed, that the latter behav-

ior is caused by residual relationships between an unallocated sub-population and its last assigned space, in combination with the variance in sub-population distributions such that an allocation of 0% of a sub-population may still result in some people from the sub-population "waking up" in the last space to which the sub-population was allocated.

The lack of constraints on assignment of activities to spaces resulted in teams being able to create designs where certain sub-populations would not have at least one of their two occupations represented within the site. Under the definitions used in this work, such a design may end up being considered desirable by some stakeholders due to its performance, but should be given a low feasibility score as the design would generally not fulfill its mission goal. However, it would have been more realistic for the software to have included a constraint such that all three of the selectable activities were represented on at least one level. This would have also reduced the solution space available and made the engineering problem more approachable for the teams.

A second, more significant, limitation of the software was related to a tradeoff made during its implementation. As the intention was to have each team with at least one participant running the software on their own personal hardware, the performance capabilities of the hardware upon which the software would be running were unknown. In conjunction with the use of multiple Monte Carlo runs per simulation execution, a balance needed to be struck between the runtime of an execution and the convergence for the results from the simulated runs. This tradeoff was made in favor of runtime, with a performance benchmark target set at no more than 30 seconds per simulation execution. Given the performance of the ABMS at the time of Workhop 2 and Workshop 3, this meant 12 and 25 simulation runs per execution, respectively. Due to this relatively small number of runs, executing a simulation for the same design variable choices would return simulation results with differing objective vector values. This behavior was observed by some teams and was exacerbated by the automatic rescaling of the tradespace plot which made event small deviations between simulations appear rather large. For the number of runs used in the workshops reported in this work, the error between simulations for the same design fell

within the range of 2-8% (Figure 5-1). It should be noted, that the runtime shown in the plot is for a revised version of the ABMS engine with significant performance improvements over the version available during the workshops, and was also run on high performance hardware.

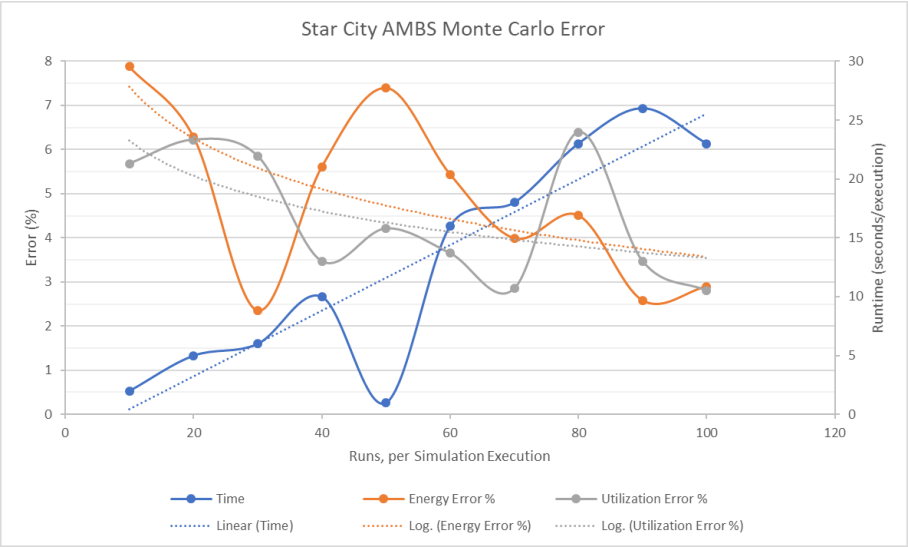


Figure 5-1: Plot of the error in select objective vector values and the ABMS runtime per simulation execution given the number of Monte Carlo simulation runs being executed. This data was generated by executing a simulation for the baseline design scenario 11 times. The first execution was taken as the reference and the sum of squares error calculated between that run's results and the results from the other 10 runs. Runtime was taken as the average of the 11 execution runtimes.

Not only was this behavior noticed by the teams, at least one member of a team (Power Budget Engineer, Team 1, Workshop 2) called into question the validity of the simulated results and their ability to infer any meaning from them for the design decisions they had made. This participant subsequently completed the post-team-activity survey with a desirability and feasibility score of 1 for all canonical designs.

Additionally, and perhaps even more importantly for the analysis presented in Chapter 4, the variability present in the objective vector values affected the perceived performance outcomes of a design as observed by a team and subsequently used in making future design decisions. During results analysis such deviations in objective vector values could result in different Pareto Rankings for a given design vector depending on whether the objective values used for the ranking were those observed by

the team, or, a set of more accurate values generated by a more robust (higher run count) simulation. This work erred on the side of ranking designs based upon the objective vector values as seen by the teams. Overcoming this through improved ABMS performance to allow for more runs and thus greater convergence of results, and/or reducing the error between results, if appropriate, would likely lead to significant improvements in overall workshop and experimental outcomes.

One of the features of the workshop software was a restriction that required users to enter their team number prior to being presented with the "game" interface. This was an intentional design choice in order to ensure that the software could capture the necessary team identification information and incorporate it into the simulation results when exported. However, an emergent result of this choice was that no team could participate in the software walk-through and simulate the baseline scenario demonstrated. This was due to the workshop structure/implementation choice of having team formation come after the walk-through. In Workshop 3, team assignments were preemptively made such that they could be shared ahead of the walk-through to avoid this limitation. However, a significant number of teams would have ended up with fewer than three participants once the actual attendance of the workshop was known, and so random team allocation was made during the event. This once again prevented teams from having assigned team numbers ahead of the walk-through and thus they were unable to follow along using the software in real time.

A further minor limitation of the software was the inability for the interactive systems model visualization to highlight multiple systems model objects. The software as implemented for this work could only highlight a single model object at a time. This manifested once the switch to a more aggregated set of design decisions in the user interface had been made, with per-level assignment of transportation mode and space activity. When clicking on such aggregate decision input, only a single, representative, model object was highlighted. Adding a multi-highlight feature would allow participants to click on such decision inputs and be presented with a visualization of all impacted objects, thus giving them a sense of spatial and scope impact regarding the decision being making.

Another limitation, from the perspective of workshop structure and implementation, was that the software used for this work was built with a custom, fixed "game" interface where the decision inputs and their expected values were more tightly coupled to the systems model than is desirable. This made it too difficult to implement a toy problem for an improved instructional experience. A valuable future enhancement to the workshop software would be to support dynamic configuration of the game interface as part of, or based upon, the loading of a given systems model.

Lastly, the version of the software used in this work allowed participants to document their design notes, thoughts, and expectations prior to simulating a particular design. One additional feature that may be of value for future workshop experiments would be to include the capability to document team discussions and notes regarding the simulation results directly in the software as well. In particular, the ability to create simple, survey like questions that can be responded to for each/any design, post simulation, could enable valuable data gathering during the team activity.

Results & Analysis

The approach taken in Chapter 4 to the analysis of data collected — a demonstration of method — was directly the result of the lack of usable data acquired over the series of team workshop experiments. The demonstration was meant to provide a grounded analysis that would have supported the investigation of this work's hypotheses given sufficient amounts of the necessary data. However, in taking this demonstrative approach, some aspects of the data analysis were simplified in order to more clearly communicate the overall intent of the method. The results of this are that, even for that data which was complete and usable, the evaluation of a team's exploration and performance was incomplete. In particular, the enumeration of only a subspace of the complete problem space from which Pareto Ranks were calculated, as well as the use of only a limited number of metrics to compute the the Pareto Rank of objective vectors, both result in a notional, rather than actual, Pareto Rank value. Thus, measurements of exploration and/or performance which rely upon Pareto Rank are not truly representative.

Additionally, the metrics used for considering team performance and exploration were necessary but likely insufficient. For the solution space metrics, the use of entropy calculations needs to be enhanced to properly support the case of constrained design variables. A robust method of quantifying exploration in the solution space should also be capable of handling the case where teams rightfully ignore variables with no systemic effect. For the problem space metrics, more consideration needs to be given to the structure of the design walk, the position of elements explored within the overall problem space, and the structure of the tradespace itself. Evaluation of exploration on the basis of the type of exploration, as a measure of the pattern of change, should also be considered by the metrics used. Patterns of exploration may include concentrated/diffuse exploration that describe the space covered, or incremental/architectural exploration the describes the step-wise change between points or clusters of points in the design walk.

Finally, the analysis of results presented in this work is limited by the use of Pareto Rank for evaluation of performance. The concept of Pareto optimality comes from the idea that an optimal solution is calculable, and this may very well deviate from the intention of human-in-the-loop model based engineering. Indeed, there may be many reasons why a team does not select their best Pareto Ranked design as their preferred design. It could be a lack of awareness of how well their best design performed on evaluation metrics that went unevaluated by the team, it could be a disconnect between the team's perception of value and the standards set by the research in calculating Pareto Ranks, or, it could be because the team is purposely choosing that design for a reason not captured within the objective vector itself. Considering other more holistic measures of team and system performance represent an area for future research to distinguish itself from this and similar work that has relied upon Pareto Rank.

5.2 Insights

Over the course of this work numerous insights were gained across multiple domains including experimental design, teamwork workshop execution, ABMS, and modeling language development. The following is an itemization of a few of the more salient insights that did not fit elsewhere.

- The use of web video conferencing software to run teamwork workshop experiments provides tremendous opportunity for expanding the reach of such research, but also creates new challenges for adequately providing instruction and oversight.

- The use of cloud based platforms for sharing workshop materials and gathering workshop data was of mixed success. The creation of a data handling and validation pipeline should be included as part of an experimental design. Design for the data — acquisition, storage, and retrieval. Make it easy, make it robust.

- The use of cloud based platforms for real time computing/simulation and/or management of workshop software should be considered over the reliance upon hardware provided by the participants. The benefits of pre-staging and testing software, and ensuring successful post-workshop data retrieval, outweigh the cost of setup.

- The use of role-playing within the teamwork workshops was deemed both successful and a valuable addition to getting participants to engage with the model problem.

- Be prepared with good materials, well organized — 2x what you think you will need will get you 50% of what you expect.

- Consider the impact of dropout rates in the anticipated availability of final data when designing and promoting teamwork workshops.

- The more obscure the bug in an ABMS, the more valuable the insight to be gained. The emergent behaviors from the interaction between model problem configuration and the assumptions upon which the ABMS was built can seem inexplicable. Have a validation plan and write it down.
- Simple rules can grow to be quite complex, fight the urge to code for behavior. Focus on defining the appropriate agent policy for the phenomena being modeled and seek to understand unexpected emergence before adapting to corner cases.
- Teamwork research is hard!

5.3 Next Steps

Several aspects of this research effort remain to be completed ahead of new efforts that may build upon it in the future.

This work had considered a number of additional metrics and analytical methods, beyond those presented, for evaluating the data collected in Workshop 2 and Workshop 3. Going forward, a per-role analysis of the effect of the treatments on the pre-team-activity scores for the canonical designs would be a valuable addition to the analysis done to date. Also, a complete reanalysis of the combined data from both workshops must be done. Additionally, the simulation results already collected provide sufficiently differentiated real-world data for testing a variety of methods for comparing design walk structure through the application of both graph theory and information space measurement. While the results of these approaches may not provide new insight into the research questions of this work, the evaluation of their effectiveness in distinguishing features of exploration present in the data would be a meaningful contribution.

Lastly, additional workshop events utilizing the experimental design and workshop structure used in this work remain to be held. With the goal of gathering sufficient data for an analysis capable of formally testing the hypotheses put forth by this

research, workshops going forward will draw upon the refinements to implementation and execution as suggested above.

5.4 Conclusion

This work builds upon the general body of cognition and teamwork research from the likes of Dyer[18], Bowers[14], Cannon-Bowers[9], Gentner[21] Maitheu[15], Rooney-Varga[32], Stout[27], Sundstrom[12] and many others. However, it builds in particular, upon the work of MIT's GTL on instrumented teamwork workshops, specifically, Fruehling[5], Manandhar[8], Pelegrin[7], and Tan[6]. The previous works of these researchers laid out the frameworks and theoretical models necessary for framing both the research questions and the experimental design presented in this work.

Several models of teamwork exist in the literature, with perhaps the most well known being that of the Input-Process-Output (I-P-O) model put forth by McGrath in 1964[17]. In this model of teamwork, McGrath sought to classify the large number factors that influence teams and teamwork, and presented them in a simplified feedback loop of inputs influencing the emergent, dynamic team processes that produce a variety of outputs which then affect the original inputs to the team. This work has followed in the spirit of that same theoretical representation of input-process-output whereby it was hypothesized that the input mental models of the team would influence the process of systems model exploration and that the outputs of such exploration would have an effect on the mental models of the team.

Given the volume of research on the various individual factors that McGrath managed to narrow down into these three basic categories, this representation of a team and teamwork is oversimplified. The intention of this research was not to directly probe the validity of the I-P-O representation, but rather to capture one aspect of teamwork that was theorized to follow such a model. In doing so, the methodology presented herein attempted to control for several factors of teamwork across the I-P-O model, while many others went unexamined.

The composition of the teams within this work was a randomly assemblage in

order to distribute variance in ability, prior knowledge, and personality, and the use of role-play sought to establish a common level of background, stimulate an affective response, and present motivation for all team members. The roles selected for role-play were chosen to impart no particular influence on the perception of group structure expectations, and the high level team goal was common among all teams. However, no attention was paid to controlling or manipulating the team structure and processes that emerged from the collision of the random distribution of team members. This includes not having a particularly well defined set of specializations for team members beyond background knowledge provided by the role-play personas. Nor was the environment in which the teams operated particularly well controlled and no particular reward stimulus was included. Additionally, the breakdown, classification, and detection of the tasks that teams may have or needed to execute in order to achieve their goal was not addressed at all within this research. Finally, assessment of the outputs that emerged from the teamwork event were extremely limited, focusing exclusively on one view of performance and one aspect of the effects induced within the team.

Despite the narrowness of the slice of teamwork phenomena covered, the contributions of this work are broad and meaningful. In its exploration of shared cognition through the use of instrumented teamwork workshops, this work has demonstrated method across mental model elicitation, the use of low fidelity simulations, the use of participant role-play in team experimentation, and contributed to the research and development of the workshop software that was used. These methods built upon previous GTL research that had developed approaches to quantifying solution space and problem space exploration and demonstrated success of application in teamwork workshop experiments, such as Tan and Moser's[6] work on team decision processes and Manandhar et al.'s[8] use of an instrumented teamwork research platform.

Through the use of qualitative, subjective evaluations, this work suggests to have shown a method of successfully eliciting a representation of stakeholders' mental models and explored the idea that diversity in this representation is influential in how teams perceive the tradespace that exists for a systems model, and therefore

the systemic awareness that can arise through such a model's use. While the limited availability of data prevented a meaningful analysis of exploration and how mental models both influenced and were influenced by the teamwork activity, the analytical methods applied set a foundation to address questions around the influence of diversity of mental models on team performance, the definition of performance as an information generating process, and the effect of tradespace and exploration structure on teams.

The many questions of this work that have been left unanswered, and the many more raised along the way, remain to be addressed as a continuation of this research effort, and contribute to the opportunities for the future work of others.

5.4.1 Future Work

Throughout this research additional questions and potential applications of the methods have been identified.

One of the most important aspects of teamwork not considered as part of this work is that of the social interactions taking place during the team activity. It has been taken as a given that such actions contribute, and are core, to the development of shared mental models during teamwork. However, the experimental design that was used meant only to detect this phenomena implicitly through its effect on the tradespace exploration and the resultant change in mental models. Quantifying the quality and quantity of the social aspect of teamwork during tradespace exploration offers an interesting opportunity for future research. One that is well aligned with the previous work of MIT's GTL[10]. Of particular interest is the extension of this work with the methods of Peng et al.[45]. Peng's work seeks to detect and visualize the variation in team behaviors during team dialogue. Inclusion of such an approach would add an interesting additional axis for analyzing the tradespace exploration of a team and assessing correlations with the changes in mental models.

Another area of future work valuable to demonstrating the usefulness of the methods used here, is in the attribution of the changes in mental models detected as part of this work. An alternative formulation of the experimental design used herein could

elucidate the degree to which the use of the systems model influenced the changes observed, as compared to the social aspects of teamwork. For example, the use of the pre/post qualitative assessment of canonical designs could be paired with treatments where stakeholders participate in only discussion about the system with no model, where there are no stakeholder roles but general information information and background on the model problem, and where there is some other arbitrary task such as reading unrelated material or playing a simple game.

A variety of broad, related research questions exist around shared mental models formed during teamwork and the change in mental models seen in the results of this work. These questions include:

- How permanent are shared mental models or the changes in an individual's mental model stimulated by teamwork?
- What do such changes mean for future teamwork efforts for the same set of stakeholders? For the stakeholders as they form new teams with others?
- To what degree are mental models that are previously aligned, or have become aligned through teamwork, disrupted by changes in team composition?
- How does disruption of teamwork influence the effects of mental model diversity on systems model use?
- How does the temporality of disruption influence the process of shared mental model development and mental model change in a team? (Does adding or removing diversity later in the process of team formation and team task have a different effect?)
- Does the duration of tradespace exploration mediate the degree of mental model change for a team and in what ways?
- How do the characteristics of current knowledge and new information interact during tradespace exploration? When and why is new information taken advantage of, ignored, or cause a team to turn back the way they came?

In addition to the questions on mental models, there are at least two outstanding questions from this work regarding the structure of the tradespace and its influence on exploration. First, how does the starting point of a team, in its knowledge, assumptions, and early designs, influence their pattern of exploration? Second, can we detect and differentiate the aspects of a teams exploration that influence their shared mental models? If there are parts of the tradespace which are constrained and cannot provide the team with new information, can we tell which teams were aware of this? Can we attribute changes we see in mental models to the meaningfulness of the areas of the tradespace that have been explored?

Finally, another area of future work lies in revisiting the approach to elicitation of mental models. Use of a more robust approach that captures the elements and relationships that compose these models could be valuable for determining what is changing and by how much it changes through teamwork.

5.5 Summary

This chapter has presented a discussion of the findings and limitations of the analysis of Chapter 4 and the experimental data upon which the findings were drawn. The additional limitations of the experimental design, model problem, the series of teamwork workshop experiments, workshop software, and the method of results analysis developed in this work were also examined. Insights that emerged over the course of this research were shared as were the next steps in this research stream. Finally, conclusions regarding the position of this work within the existing body of research, its contribution in the field, and future work were made.

May the efforts of this research be helpful to those who follow — its methods prove useful, the insights valuable, and the knowledge shared serve as a building block for the continued research of teams and teamwork, such that it leads to a positive impact for all.

Mens et Manus

The End.

Appendix A

Workshop Materials

シナリオ　Scenario

GLOBAL TEAMWORK LAB

今年は2041年で、MITの卒業生であるGeorge Lordos（SDM'16）によって設立された世界最大の航空宇宙企業「Lordos＆Lordos」で働いています。 あなたの会社は、火星の開拓を解き放つスターシップクラスのロケットを提供したスペースXによってのみライバルになります。

The year is **2041** and **you work for the world's largest aerospace company** – "Lordos & Lordos", founded by MIT alum George Lordos (SDM'16). Your company is rivaled only by SpaceX, which has provided the Starship class rockets that have unlocked the **settlement of Mars**.

火星の最初の人間居住地である「スターシティ」を建設するミッションは、今年開始される予定です。 コンセプトは技術的に検証されており、新しい技術開発は必要ありません。 あなたのチームは、集落の最初の村のサイト設計アーキテクチャを選択する任務を負っています。

The mission to build Mars's first human settlement, "Star City", is **scheduled to launch this year**. The concept has been technically verified and **no new technology development is needed**. Your team has been tasked with **choosing a site design architecture for the settlement's first village**.

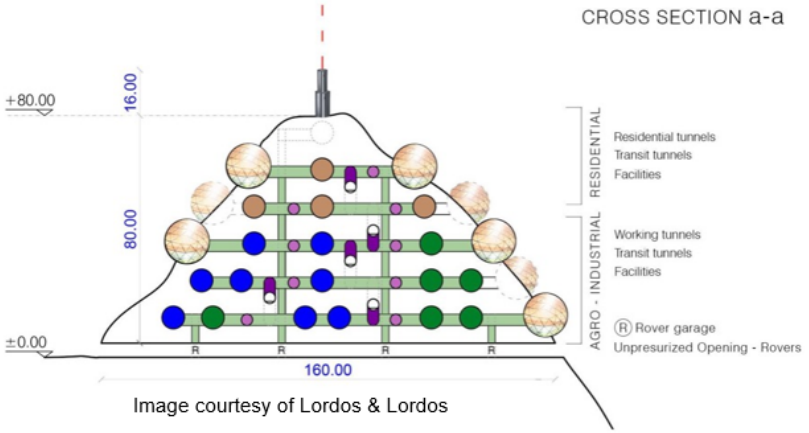


Image courtesy of Lordos & Lordos

Figure A-1: Star City Background & Scenario.

システムモデル / The systems model

GLOBAL TEAMWORK LAB

割り当て可能なスペース：教育、ヘルスケア、文化	Assignable Space: Educational, Healthcare, Cultural
住宅スペース	Residential Space
共同ドーム	Communal Dome
農業空間	Agricultural Space
製造スペース	Manufacturing Space
接続	Connection
エレベーターロビー/ジャンクション	Elevator Lobby/Junction

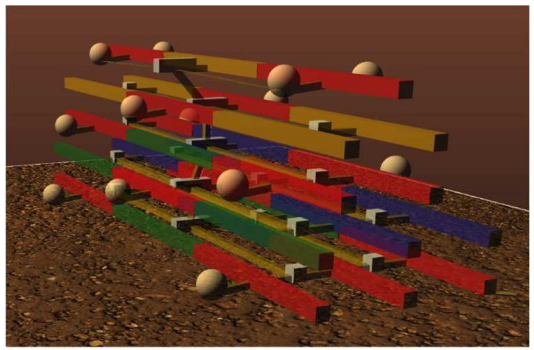


Figure A-2: Star City Site Model Summary.

The systems model - simulation

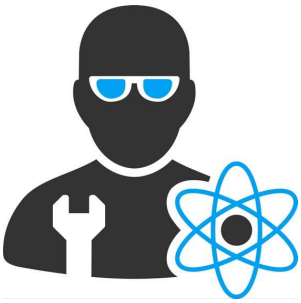
GLOBAL TEAMWORK LAB

- The model will simulate (but not visualize) the populations moving & engaging on the site
  - No hard capacity constraints, but people will avoid/move more slowly through crowded spaces
- Some variation between simulations, with the same settings, is expected

Figure A-3: Star City Simulation Summary.

Mars Star City Site Design & Simulation Workshop
Stakeholder Profile 3

Mission Power Budget Engineer



Primary Objective:

- Ensure effective utilization of constrained power resources.

Secondary Objective:

- None

Description:

As the Mission Power Budget Engineer, your role is to ensure that the ample but scarce power resources of Start City are appropriately and effectively utilized. Every kJ of energy that is spent in daily activity is a kJ of energy that is not available for supporting the expansion of Star City or critical activities that happen outside of the village (production of in-situ resources).

Martian Transit Modes

Mode	Speed (Relative)	System Energy Cost*
Walking	1x	0 kJ/use
Moving Walkway	1.53x	129.2 kJ/use
Scooter	7.34x	396 kJ/use

*1 use == movement down one connection

Your value function(s):

Energy Use:
Over 1,250 MJ has minimal value
Between 1,250-1,000 MJ is acceptable
Under 1,000 MJ is your goal

v1.2 2021/04/24

Figure A-4: Example role-playing scenario stakeholder profile.

Mars Star City Site Design & Simulation Workshop
Canonical Design

Canonical Design #3

Populations:

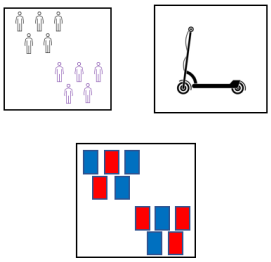
- Populations are separated and located as close to their Primary Occupation activities as possible.

Connections:

- All connection modes are Scooters only.

Functional Activities:

- Education and Cultural spaces are on the upper floors while Healthcare and Cultural are on the lower floors.



v1.1 2021/04/17

Figure A-5: Example role-playing scenario canonical design.

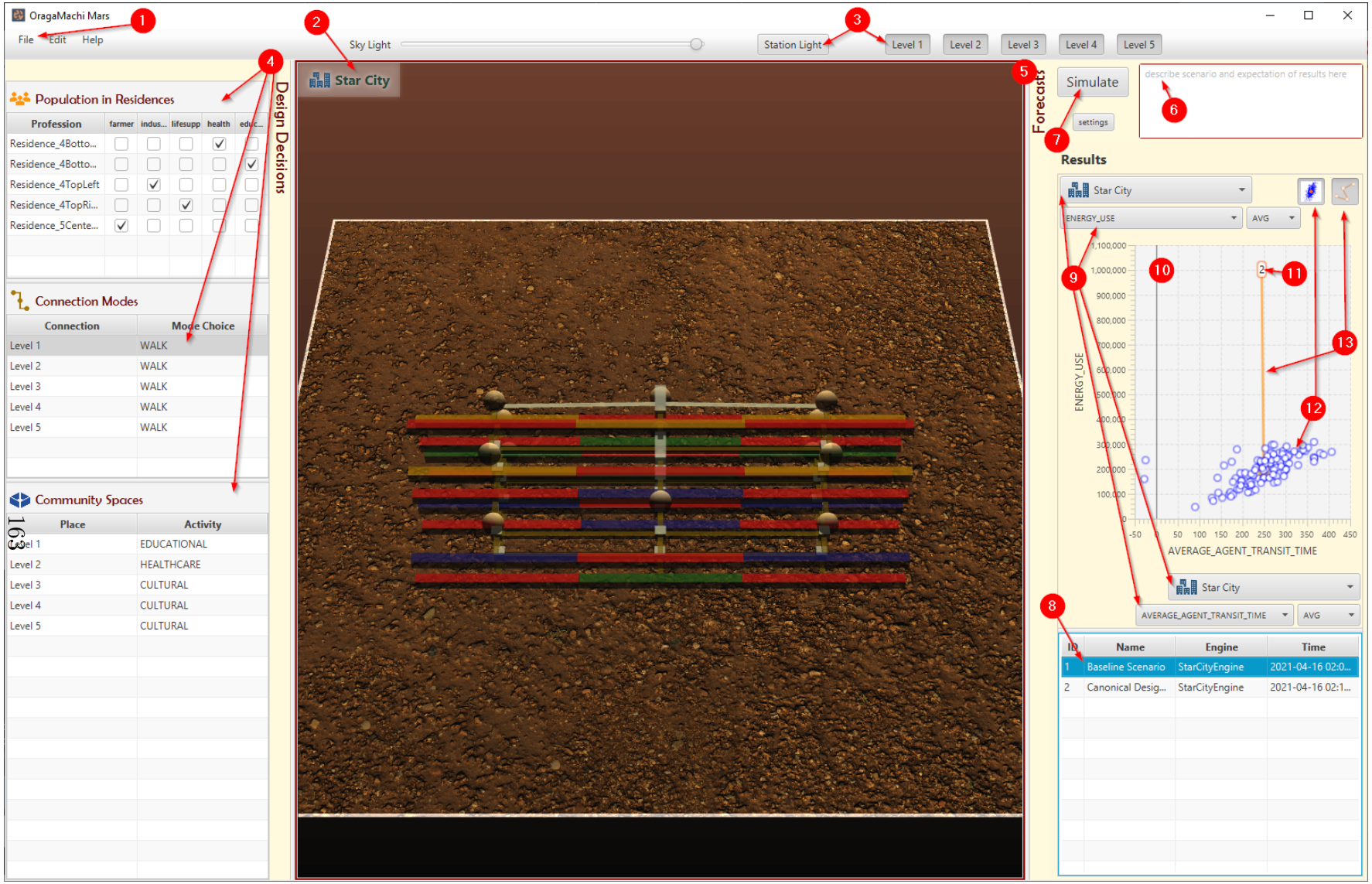


Figure A-6: Workshop software user interface, labeled. (1) menu bar, (2) systems model visualization, (3) visualization scene controls, (4) design decisions interface, (5) interactive simulation results interface, (6) simulation description input, (7) simulation execution button, (8) chronological list of simulation results, (9) tradespace plot axis selection dropdowns, (10) tradespace plot, (11) simulation result point, (12) simulation result values cloud and toggle button, and (13) simulation results design walk and toggle button.

Bibliography

[1] Steve W.J. Kozlowski and Daniel R. Ilgen. Enhancing the Effectiveness of Work Groups and Teams. *Psychological Science in the Public Interest*, 7(3):77–124, December 2006. Publisher: SAGE Publications Inc.

[2] Michael Bacharach. Foreward: Teamwork. In Natalie Gold, editor, *Teamwork : multi-disciplinary perspectives.*, pages xxi–xxv. Basingstoke, Hampshire ; New York : Palgrave Macmillan, 2005., 2005.

[3] Olivier L. de Weck, Daniel Roos, and Christopher L. Magee. *Engineering Systems: Meeting Human Needs in a Complex Technological World*. The MIT Press, October 2011.

[4] Herbert A. Simon. A Behavioral Model of Rational Choice. *Quarterly Journal of Economics*, 69(1):99–118, February 1955. Publisher: Oxford University Press / USA.

[5] Carl Fruehling and Bryan R. Moser. Analyzing Awareness, Decision, and Outcome Sequences of Project Design Groups: A Platform for Instrumentation of Workshop-Based Experiments. In Eric Bonjour, Daniel Krob, Luca Palladino, and François Stephan, editors, *Complex Systems Design & Management*, pages 179–191, Cham, 2019. Springer International Publishing.

[6] Puay Siang Tan and Bryan R. Moser. Detection of Teamwork Behavior as Meaningful Exploration of Tradespace During Project Design. In Michel Alexandre Cardin, Daniel Hastings, Peter Jackson, Daniel Krob, Pao Chuen Lui, and Gerhard Schmitt, editors, *Complex Systems Design & Management Asia*, Advances in Intelligent Systems and Computing, pages 73–87, Cham, 2019. Springer International Publishing.

[7] Lorena Pelegrin, Bryan Moser, and Vivek Sakhrani. Exposing Attention-Decision-Learning Cycles in Engineering Project Teams through Collaborative Design Experiments. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*, January 2019. Accepted: 2019-01-02T23:40:27Z http://hdl.handle.net/10125/59475.

[8] P. Manandhar, K. Rong, K. Carroll, R. de Filippi, I. Winder, J. Dieffenbach, and B. R. Moser. Sensing systemic awareness and performance of teams during

model-based site design. In *2020 IEEE 6th World Forum on Internet of Things (WF-IoT)*, pages 1–6, June 2020.

[9] Janis A. Cannon-Bowers and Clint Bowers. Team development and functioning. In Sheldon Zedeck, editor, *APA handbook of industrial and organizational psychology, Vol 1: Building and developing the organization.*, pages 597–650. American Psychological Association, Washington, 2011.

[10] Lorena Pelegrin, Bryan Moser, Shinnosuke Wanaka, Marc-Andre Chavy-Macdonald, and Ira Winder. Field Guide for Interpreting Engineering Team Behavior with Sensor Data. In Eric Bonjour, Daniel Krob, Luca Palladino, and François Stephan, editors, *Complex Systems Design & Management*, pages 203–218, Cham, 2019. Springer International Publishing.

[11] Eric Sundstrom, Michael McIntyre, Terry Halfhill, and Heather Richards. Work groups: From the Hawthorne studies to work teams of the 1990s and beyond. *Group Dynamics: Theory, Research, and Practice*, 4(1):44, 2000. Publisher: US: Educational Publishing Foundation.

[12] Eric Sundstrom, Kenneth P. De Meuse, and David Futrell. Work teams: Applications and effectiveness. *American Psychologist*, 45(2):120–133, February 1990. ISBN: 9781557980922 Num Pages: 120-133 Place: Washington, US Publisher: American Psychological Association (US).

[13] Richard A. Guzzo and Gregory P. Shea. Group performance and intergroup relations in organizations. In *Handbook of industrial and organizational psychology, Vol. 3, 2nd ed*, pages 269–313. Consulting Psychologists Press, Palo Alto, CA, US, 1992.

[14] Clint Bowers, Eduardo Salas, Carolyn Prince, and Michael Brannick. Games teams play: A method for investigating team coordination and performance. *Behavior Research Methods, Instruments, & Computers*, 24(4):503–506, December 1992.

[15] John E Mathieu, Tonia S Heffner, and Gerald F Goodwin. The Influence of Shared Mental Models on Team Process and Performance. *Journal of Applied Psychology*, page 11, 2000.

[16] Eduardo Salas, Terry L. Dickinson, Sharolyn A. Converse, and Scott I. Tannenbaum. Toward an understanding of team performance and training. In *Teams: Their training and performance.*, pages 3–29. Ablex Publishing, Westport, CT, US, 1992.

[17] Joseph Edward McGrath. *Social psychology, a brief introduction.* New York, Holt, Rinehart and Winston, 1964.

[18] Jean Dyer. Team research and team training: A state-of-the-art review. *Human factors review*, 26:285–323, 1984.

[19] Norbert Seel. Model-based learning: a synthesis of theory and research. *Educational Technology Research & Development*, 65(4):931–966, August 2017. Publisher: Springer Nature.

[20] D. E. Rumelhart, P. Smolensky, J. L. McClelland, and G. E. Hinton. PART IV: PSYCHOLOGICAL PROCESSES: CHAPTER 14: Schemata and Sequential Thought Processes in PDP Models. *Parallel Distributed Processing*, pages 7–57, July 1987.

[21] D. Gentner. Mental Models, Psychology of. In *International Encyclopedia of the Social & Behavioral Sciences*, pages 9683–9687. Elsevier, 2001.

[22] Norbert M. Seel. Model-Based Learning and Performance. In J. Michael Spector, M. David Merrill, Jan Elen, and M. J. Bishop, editors, *Handbook of Research on Educational Communications and Technology*, pages 465–484. Springer, New York, NY, 2014.

[23] John S. Carroll, John Sterman, and Alfred A. Marcus. Playing the Maintenance Game: How Mental Models Drive Organizational Decisions. In Jennifer J. Halpern and Robert N. Stern, editors, *Debating rationality: Nonrational aspects of organizational decision making*, pages 99–121. Frank W. Pierce Memorial Lectureship and Conference Series, no. 10., 1998.

[24] Barbara C. Buckley. Model-Based Learning. In Norbert M. Seel, editor, *Encyclopedia of the Sciences of Learning*, pages 2300–2303. Springer US, Boston, MA, 2012.

[25] Florence Allard-Poesi. REPRESENTATIONS AND INFLUENCE PROCESSES IN GROUPS: TOWARDS A SOCIO-COGNITIVE PERSPECTIVE ON COGNITION IN ORGANIZATION. *Scandinavian Journal of Management*, 14(4):395–420, December 1998.

[26] Eduardo Salas, Dana E. Sims, and C. Shawn Burke. Is there a "Big Five" in Teamwork? *Small Group Research*, 36(5):555–599, October 2005. Publisher: SAGE Publications Inc.

[27] Renée J. Stout, Janis A. Cannon-Bowers, Eduardo Salas, and Dana M. Milanovich. Planning, Shared Mental Models, and Coordinated Performance: An Empirical Link Is Established. *Human Factors*, 41(1):61–71, March 1999. Publisher: SAGE Publications Inc.

[28] Eric Specking, Gregory Parnell, Edward Pohl, and Randy Buchanan. Early Design Space Exploration with Model-Based System Engineering and Set-Based Design. *Systems*, 6(4):45, December 2018. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

[29] Alex D. MacCalman, Paul T. Beery, and Eugene P. Paulo. A Systems Design Exploration Approach that Illuminates Tradespaces Using Statistical Experimental Designs. *Systems Engineering*, 19(5):409–421, 2016. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sys.21352.

[30] John Sterman, Travis Franck, Thomas Fiddaman, Andrew Jones, Stephanie McCauley, Philip Rice, Elizabeth Sawin, Lori Siegel, and Juliette N. Rooney-Varga. WORLD CLIMATE: A Role-Play Simulation of Climate Negotiations. *Simulation & Gaming*, 46(3-4):348–382, June 2015. Publisher: SAGE Publications Inc.

[31] Leslie A. DeChurch and Jessica R. Mesmer-Magnus. Measuring shared team mental models: A meta-analysis. *Group Dynamics: Theory, Research, and Practice*, 14(1):1–14, 2010.

[32] J. N. Rooney-Varga, J. D. Sterman, E. Fracassi, T. Franck, F. Kapmeier, V. Kurker, E. Johnston, A. P. Jones, and K. Rath. Combining role-play with interactive simulation to motivate informed climate action: Evidence from the World Climate simulation. *PLOS ONE*, 13(8):e0202877, August 2018.

[33] Juliette N. Rooney-Varga, Florian Kapmeier, John D. Sterman, Andrew P. Jones, Michele Putko, and Kenneth Rath. The Climate Action Simulation. *Simulation & Gaming*, 51(2):114–140, April 2020.

[34] George Lordos and Alexandros Lordos. Star City: Designing a Settlement on Mars, November 2019.

[35] Mars Society Awards Prizes to Mars Colony Design Contest Winners, October 2019.

[36] Katherine M. Carroll. Agent-Based Modeling of Population Activity in Complex Terrestrial and Martian Sites. Master of Science in Aeronautics and Astronautics, MIT, Cambridge, Massachusetts, June 2021. Publisher: Massachusetts Institute of Technology.

[37] Robert Axelrod. Introduction. In *The Complexity of Cooperation: Agent-Based Models of Competition and Collaboration*, pages 3–9. Princeton University Press, 1997.

[38] K.M. Carley, D.B. Fridsma, E. Casman, A. Yahja, N. Altman, Li-Chiou Chen, B. Kaminsky, and D. Nave. BioWar: scalable agent-based model of bioattacks. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 36(2):252–265, March 2006. Conference Name: IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans.

[39] C.M. Macal and M.J. North. Tutorial on agent-based modeling and simulation. In *Proceedings of the Winter Simulation Conference, 2005.*, pages 14 pp.–, December 2005. ISSN: 1558-4305.

[40] Bryan Robert Moser. *The Design of Global Work: Simulation of Performance Including Unexpected Impacts of Coordination across Project Architecture*. Doctoral dissertation, 東京大学 (University of Tokyo), 2012.

[41] Keiji Kimura. The Effect of Introducing New Transportation Services on the Community Engagement of Elderly People and Parents. Master of Science in Engineering and Management, MIT, Cambridge, Massachusetts, June 2021. Publisher: Massachusetts Institute of Technology.

[42] Jeremy Francis Dawson. *MEASUREMENT OF WORK GROUP DIVERSITY*. PhD thesis, Aston University, 2011.

[43] Harry Khamis. Measures of Association: How to Choose? *Journal of Diagnostic Medical Sonography*, 24(3):155–162, May 2008. Publisher: SAGE Publications Inc STM.

[44] Igor Douven. A Bayesian perspective on Likert scales and central tendency. *Psychonomic Bulletin & Review*, 25(3):1203–1211, June 2018.

[45] Sixiong Peng, Takashi Amakasu, Hiroki Kawauchi, Hideyuki Horii, and Kazuo Hiekata. Development of method for visualizing behavioral states of teams. *Unpublished manuscript*, page 10, 2020.