2021

Applying Cognitive Measures In Counterfactual Prediction

Lori A. Mahoney
Wright State University

APPLYING COGNITIVE MEASURES IN COUNTERFACTUAL PREDICTION

A dissertation submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

by

LORI A. MAHONEY

M.S., Air Force Institute of Technology, 2004

B.S., Michigan Technological University, 2001

2021

Wright State University

WRIGHT STATE UNIVERSITY

GRADUATE SCHOOL

8 November 2021

I HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER MY SUPERVISION BY Lori A. Mahoney ENTITLED Applying Cognitive Measures in Counterfactual Prediction BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Doctor of Philosophy.

Ion Juvina, Ph.D.
Dissertation Director

Ivan Medvedev, Ph.D.
Director, Interdisciplinary Applied Science
and Mathematics Program

Barry Milligan, Ph.D.
Vice Provost for Academic Affairs
Dean of the Graduate School

Committee on Final Examination:

Ion Juvina, Ph.D.

Joseph W. Houpt, Ph.D.

Valerie L. Shalin, Ph.D.

Zheng Xu, Ph.D.

ABSTRACT

Mahoney, Lori A. Ph.D. Interdisciplinary Applied Science and Mathematics Program,
Wright State University, 2021. Applying Cognitive Measures in Counterfactual
Prediction.

Counterfactual reasoning can be used in task-switching scenarios, such as design and planning tasks, to learn from past behavior, predict future performance, and customize interventions leading to enhanced performance. Previous research has focused on external factors and personality traits; there is a lack of research exploring how the decision-making process relates to both task-switching and counterfactual predictions. The purpose of this dissertation is to describe and explain individual differences in task-switching strategy and cognitive processes using machine learning techniques and linear ballistic accumulator (LBA) models, respectively, and apply those results in counterfactual models to predict behavior. Applying machine learning techniques to real-world task-switching data identifies a pattern of individual strategies that predicts out-of-sample clustering better than random assignment and identifies the most important factors contributing to the strategies. Comparing parameter estimates from several different LBA models, on both simulated and real data, indicates that a model based on information foraging theory that assumes all tasks are evaluated simultaneously and holistically best explains task-switching behavior. The resulting parameter values provide evidence that people have a switch-avoidance tendency, as reported in previous research, but also show how this tendency varies by participant.

Including parameters that describe individual strategies and cognitive mechanisms in counterfactual prediction models provides little benefit over a baseline intercept-only model to predict a holdout dataset about real-world task switching behavior and performance, which may be due to the complexity and noise in the data. The methods developed in this research provide new opportunities to model and understand cognitive processes for decision-making strategies based on information foraging theory, which has not been considered previously. The results from this research can be applied to future task-switching scenarios as well as other decision-making tasks, both in a laboratory setting as well as the real-world, and have implications for understanding how these decisions are made.

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to take this opportunity to extend my gratitude to my advisors, Dr. Joseph Houpt and Dr. Ion Juvina, for their mentorship and for sharing their wisdom with me throughout my graduate studies. They have challenged me to be a more thoughtful and confident researcher. My remaining dissertation committee, Dr. Valerie Shalin and Dr. Zheng Xu, also challenged me to view different perspectives and think more critically about my research. I greatly appreciate all of their insight and support in developing this dissertation.

I would also like to thank the Defense Advanced Research Projects Agency for providing the data used in this research, the Ohio Supercomputing Center for the computational resources used in the analysis, and the National Geospatial-Intelligence Agency for funding my gratitude studies.

I am grateful to my fellow graduate students in the Mathematical Modeling of Human Performance Lab and the ASTECCA Lab for sharing their knowledge, constructive feedback, and friendship with me.

Finally, I would like to thank my family and friends, especially my husband, whose continual understanding, support, and love made this dissertation possible.

1. INTRODUCTION AND PURPOSE

This manuscript is structured as follows. Chapter 1 discusses the motivation and background for this research. Chapter 2 describes the existing dataset that provided additional motivation for the specific modeling approaches used in this research. Chapter 3 summarizes the overall methodological approach used for this research and then provides details of each of the modeling approaches. Chapters 4, 5 and 6 discuss the results from each of the modeling approaches while chapter 7 discusses the meaning of these results and potential changes to the approaches used, in the context of the research questions and hypotheses. Chapter 8 describes general limitations of the research and proposed future work. Finally, chapter 9 summarizes the conclusions drawn from this research.

1.1. Motivation

Forecasting decisions predict the probability of a future event occurring, where the future state is evolving as the answer is being formulated, and are commonly applied in medical, meteorological, business, and geopolitical domains. Within intelligence and security analysis, there is a need to anticipate an adversary’s doctrine, principles, and/or intent in order to predict political, economic, military, and security implications and to ensure certain objects, technologies, and capabilities remain uncompromised (Trump, 2017). Much of this analysis requires understanding and predicting patterns of human behavior using the often-incomplete available information. Unlike other decision-making problems, like probabilistic inference problems, where the answer exists somewhere, but may take time and (cognitive) resources to find, the true answer does not yet exist for forecasting problems (Juvina et al., 2020a). Analysts need to answer both probabilistic inference and forecasting questions, especially because the answer to a forecasting question can depend on using responses from an inference problem. Individual forecasts, even by professionals in the field, are frequently not better than simple models or even chance (Tetlock, 2005). While the accuracy of individual forecast scores is improved by placing the best forecasters together on a team, a practice known as 'superforecasting' (Mellers et al., 2014; Mellers, Stone, Murray et al., 2015), as well as by aggregating results (Turner et al., 2014) and using the 'wisdom of crowds' (Yi et al., 2012) to reduce systemic biases in individual forecasts, it is not always practical to use aggregation or teams of forecasters.

There are multiple different techniques an individual can use to form forecasting decisions, such as anticipatory thinking and sensemaking (Klein et al., 2007; Pirolli & Card, 2005), heuristics (Harvey, 2007; Hogarth & Makridakis, 1981), building rules- or formula-based models and extrapolation (Armstrong, 2001; Harvey, 2007), and assessing counterfactuals (Hendrickson, 2009). The method used by an analyst is partially determined by the forecasting domain, with mathematical algorithms and models used more for meteorological and business forecasts, heuristics more popular in sales forecasting (Harvey, 2007), and sensemaking used in intelligence and security forecasts. Counterfactual predictions are not widely used as a forecasting tool, but are being applied as part of third wave artificial intelligence (AI) in the geopolitical and intelligence domains (Defense Advanced Research Projects Agency, 2019; Hendrickson, 2009).

Defense Advanced Research Projects Agency's (DARPA's) Teaching AI to Leverage Overlooked Residuals (TAILOR) program focused on using counterfactual predictions to customize interventions to optimize human performance for multiple types of law enforcement and national security applications (Defense Advanced Research Projects Agency, 2019). Counterfactual predictions use 'what if' questions to consider alternate scenarios and how changes to previously observed factors would impact the outcome of interest. One question in particular concerned predicting counterfactuals for task-switching and overall task performance within a multi-team system (MTS) known as Project RED. As the participants attempted to complete the overall task (i.e., build a well on the Martian surface) they performed multiple different tasks and used multiple different tools to assist them. Several different models of the data predicted counterfactuals where the environment was the same, but the people were different (between-subjects); counterfactuals where the people were the same, but the environment was different (within-subject); and counterfactuals where both the environment and people were different. Within the TAILOR program, the between-subjects question asked "Based on the data from the mixed gender crews, what is the performance score and the probability of switching tasks for the all-female crew?," the within-subject question asked "Based on the data from zero- and one-minute communications delay, what is the performance score and the probability of switching tasks for the three-minute communications delay?," and the other question asked "Can a model constructed from the 30-day missions accurately predict the performance score and the probability of switching tasks in the final campaign 4 sessions, where time-in-habitat is two weeks longer than the analogous final sessions in the 30-day missions?"

The Crew Recommender for Effectively Switching Tasks (CREST) counterfactual model of Project RED (Mesmer-Magnus et al., 2020) considered features taken directly from the collected data (e.g., task characteristics, social factors, technology affordances, situational demands, and personality traits) to predict overall task performance and the likelihood of task-switching behavior for alternate scenarios (e.g., different team composition, different time frame, etc.). This was similar to Mellers, Stone, Atanasov et al’s (2015) study of forecasting performance that investigated the effects of dispositional variables (e.g., cognitive ability, open-mindedness), situational variables (e.g., training, environment), and behavioral variables (e.g., deliberation time, belief updating) to explain variation in forecasting performance. In both studies, the factors used to define ‘individual differences,’ like personality traits and cognitive ability, were not dependent on the task completed. In Mellers, Stone, Atanasov et al’s (2015) study dispositional variables were only weakly correlated to forecasting performance. The existing CREST model included social factors and personality traits, but did not include factors that describe the cognitive process(es) individuals used, like what strategy(ies) were employed to select a response or how information was gathered. There was a need to apply individual level metrics to describe differences in human performance for the overall design and planning task as well as task-switching, to better understand the mechanisms underlying the behavior and to make out-of-sample predictions. The research completed for this dissertation filled that gap by leveraging machine learning and cognitive models to extract information about the underlying cognitive processes used by participants during the task, to better describe and explain how participants approach solving the problem, both individually and as a member of a multiple teams. The research focused on comparing multiple models to identify which process mechanisms are favored to explain the observed behaviors. The addition of team factors impacting decision-making strategies and cognitive processes, such as interpersonal ties between participants and shared mental models, for both task-switching and overall task performance, was a novel contribution of this research. In addition, this research incorporated the latent variables into counterfactual prediction models by using the measures from the machine learning and cognitive models as factors in these models.

This research addressed four specific questions related to individual differences in task switching and task performance when switching between multiple tasks. The research questions and their associated hypotheses were:

- **Q1:** What decision-making strategy does a participant use to solve the overall task and what are the most important factors contributing to this strategy?

  H1: Using machine learning to identify and categorize factors, unknown and undefined a priori, that contribute to pattern(s) of individual and team decision-making strategies better predicts out-of-sample category data than a random assignment model.

- **Q2:** Is the preference to select a task based on individual task attributes or on the overall gain provided by the task?

  H2: An information foraging (IF) theory based cognitive model containing alternative-level preferences better predicts the Project RED out-of-sample data than a cumulative prospect theory (CPT) based model containing attribute-level preferences.

- **Q3:** Are all tasks evaluated simultaneously or is a serial process used to create a subset of tasks to consider when selecting the next task?

H3: A two-stage cognitive model better predicts the Project RED out-of-sample data than a single-stage model.

- **Q4:** What effect does including the decision-making strategies and cognitive process an individual uses to select which task to perform have on predicting overall task performance and task-switching behavior?

  H4a: Including individual and team decision-making strategy as a contextual factor in a counterfactual prediction model improves prediction accuracy for out-of-sample participants over the baseline model for within-subject counterfactuals, after accounting for model complexity.

  H4b: Including cognitive process model parameters as a contextual factor in a counterfactual prediction model improves prediction accuracy for out-of-sample participants over the baseline model for between-subject and within-subject counterfactuals, after accounting for additional model complexity.

  H4c: Including both the decision-making strategies and cognitive process model parameters as contextual factors in a counterfactual prediction model improves prediction accuracy for out-of-sample participants over the baseline model for all three types of counterfactual questions, after accounting for additional model complexity.

1.2. Theory

This section provides an overview of the theories that shape the research questions, the associated hypotheses, and the methodological approach used in the analysis, to investigate potential improvements to counterfactual forecasts of task-switching behavior and performance on a design and planning task, by extracting and using residual information about participants' underlying strategies and cognitive processes used to complete the overall task. In particular, background knowledge of forecasting and task-switching, as well as information foraging theory, another theory that can be used to describe task-switching behavior, provided a foundation to shape the research problem and guided this research. Knowledge of different types of decision strategies and cognitive process models informed the analysis methods used in this research.

1.2.1. Forecasting Techniques

Forecasting can be described as an 'open world' decision-making process that includes all possible factors, whether they occurred previously or not. Pirolli and Card (2005) describe intelligence analysis and forecasting as a sensemaking process where the analyst gathers information, represents the information in a formalized schema (structured to aid analysis), develops insight by manipulating the representation, and creates knowledge or an action based on the insight. The schemas vary by analyst and by question, but are central to the *sensemaking* process. A notional model of sensemaking (Pirolli & Card, 2005) consists of a foraging loop for seeking, filtering, and extracting information (possibly into a schema) and a sensemaking loop to develop iteratively a mental model from the schema that best fits the evidence. The processes are used iteratively both bottom-up and top-down to solve a problem. Klein, Snowden, and Pin (2007) distinguish a separate internally-focused *anticipatory thinking* process that combines externally available information with internal representations (semantic and episodic memories) and capabilities to generate possible future states. Klein et al. (2007) describe three forms of *anticipatory thinking*: finding similar events and clusters of cues from the past in the current situation (pattern matching), using the trajectory of events and extrapolating trends to prepare for future events (trajectory tracking), and noticing inconsistencies and interdependencies between events (conditional). With pattern matching and trajectory tracking, we respond to a cue (an event or the trend of a series of events), while with conditional anticipatory thinking we need to see the connections between events. Geden et al. (2019) also identify three distinct, but slightly different, forms of *anticipatory thinking*: anticipating future states and identifying their indicators (prospective branching), examining a particular future state and working backwards to identify its indicators and warnings (backcasting), and identifying paths from past states to the current one (retrospective branching). Because of the large number of possible future states (in theory, an infinite number), an analyst uses the anticipatory thinking process to only consider future states that are plausible and relevant (Geden et al., 2019). Information is reorganized for sensemaking by analysts to amplify their ability to find patterns for the conceptual schemas they use in understanding the relevant information needed for analysis (Pirolli & Card, 2005).

Adaptive toolbox theory (Gigerenzer, 2008) specifies that *heuristics* are used for situations where probabilities are unknown, goals or problems are ill-defined, part of the information is ignored (frugal), and solutions are needed quickly (fast). Logic and probability provide optimal solutions whereas heuristics provide satisficing solutions. Heuristics provide a robust, tractable method to make decisions by ignoring ‘unnecessary’ information using ecologically rational criteria (decision making in real-world domains). Heuristics are constructed and selected from the adaptive toolbox using primarily reinforcement learning, but also social learning and evolutionary learning. Adaptive toolbox theory says that heuristics consist of adjustable, adaptive building blocks for new situations. A less predictable situation requires the exclusion of more information (Gigerenzer, 2008). For example, Wübben and Wangenheim (2008) showed that a one-reason heuristic can forecast future customer purchasing activity as well as complex stochastic models.

Counterfactual models can be described as a ‘closed world’ that considers alternate scenarios about a limited number of factors, those that occurred in the previously observed data or the actual situation, and how changes to those factors would impact the outcome of interest (Juvina, et al., 2020b). Humans intuitively and consistently use counterfactual reasoning to make judgments about many everyday occurrences by considering possible alternate worlds in which our counterfactual statement is true to reach our conclusion (Pearl, 2018). For example, if we know that Jane did not take any aspirin and her headache did not go away, we can also consider an alternate world where Jane took an aspirin and her headache went away. The same factors are considered – headache and aspirin – but the alternate, counterfactual world reaches a different outcome. David Lewis argued in *Counterfactuals* (Lewis, 1973) that we compare the actual world (where Jane did not take an aspirin) to the “most similar” alternate world where she did take an aspirin and conclude that the counterfactual statement “Jane’s headache would have gone away if she had taken aspirin” is true (Pearl, 2018). People typically apply counterfactual reasoning to their own choices, rather than another person’s choices or behaviors, as these alternatives are easier to imagine (Kahneman & Miller, 1986), so an individual will likely conclude that “my headache would have gone away if I had taken an aspirin” and take an aspirin the next time he or she has a headache, than to consider Jane’s headache and apply the counterfactual conclusion to their own headache. Counterfactual reasoning is also commonly applied to the results of games, especially for a loss that was the result of an error in one's own performance (Kahneman & Miller, 1986), such as missing a strike in bowling. While this thought process is completed with minimal effort for the question of Jane's or one's own headache, or for other everyday occurrences, the same counterfactual reasoning process can be deliberately applied to more complex, cognitively demanding problems, like within intelligence analysis (Hendrickson, 2009), human performance optimization (Defense Advanced Research Projects Agency, 2019), lessons learned (Intelligence Advanced Research Projects Activity, 2018), and strong AI (Pearl, 2018).

Counterfactual reasoning occurs in many domains, including philosophy, artificial intelligence (AI), and psychology, as a framework for causal inference. Psychologists attempt to explain how and why humans use counterfactual models in their mind, while AI researchers are focused on building counterfactual structural models that implement causal reasoning in robots and other AI systems. Counterfactual predictions of health and human performance outcomes in alternate scenarios can inform experimental designs; for example, if there is an intervention that is predicted to provide a significant treatment effect for subjects with certain characteristics then the experimental group needs to include people with those characteristics. Additionally, counterfactual reasoning aids analysts' forecasting by improving causal inference, substantiating post-event reporting, guiding future scenario analysis, and encouraging innovative "what-if" thinking (Hendrickson, 2009). Counterfactual statements about what actions would have led to a different outcome provide a basis for developing lessons learned about events, policy, or analysis tradecraft (Intelligence Advanced Research Projects Activity, 2018). To produce a functional outcome, such as developing lessons learned that lead to implementing new approaches that generate successful future results, requires that the counterfactual accurately identifies an antecedent cause that can be acted upon, that it facilitates the means to alter future behavior, and that a future relevant opportunity for application is recognized (Smallman & Summerville, 2018). The criteria for functional versus dysfunctional outcomes have been studied for counterfactuals applied to events within an individual's life (Kahneman & Miller, 1986, Smallman & Summerville, 2018), but not in the context of predicting future events related to groups or larger events determined by another person's choices or behaviors.

Counterfactual models are more restricted in their use than pre-factual models as counterfactuals only consider factors that occurred in previously observed data or the "actual situation," while pre-factual forecasting models can consider any factor in making the prediction and can make predictions about outcomes that have never been observed, as long as the relationship between the predictor(s) and the outcome can be modeled, using either previous observations or conjecture. For example, a person may use a pre-factual model, but not a counterfactual model, to predict that they will lose their bowling game to a new opponent because they haven't bowled in 3 years while their opponent has bowled twice a week for the last 6 months. The person has no previous observations to use in making this prediction, but can use general knowledge that practice improves performance. To make a counterfactual prediction requires having observed data. Once the person has finished the bowling game against their opponent, they can use the observations from that first game to generate counterfactuals for predicting the outcome of a second game. Counterfactual models are useful for repeating events where there is an opportunity to apply changes in the future.

1.2.2. Task Switching Theory

The research questions for this dissertation focused on counterfactual predictions of task switching behavior within a multi-team system (MTS). In this research, as well as other real-life scenarios, the larger task was completed by switching between a series of smaller subtasks, sometimes individually and sometimes as a member of a team or multiple teams. The theory of task-switching provided direction to develop cognitive and counterfactual models by supplying factors thought to be relevant to task switching behavior. Better understanding of individual differences in task switching, to be able to predict an individual's switching behavior, could help optimize performance of an all-human team as well as improve human-machine interactions, by improving the coordination of switching between tasks and overall task performance.

Wickens et al. (2013) describe two forms of multi-tasking – concurrent task performance and sequential task performance. During concurrent task performance, two tasks are performed in parallel sharing cognitive resources, such as talking on the phone while driving or reading a paper while listening to music. Sequential task performance occurs when it's not possible to perform both tasks simultaneously, such as trouble-shooting a problem and monitoring other areas or writing an email and addressing a knock at the door. There are not enough resources available to perform both tasks in parallel so the individual must focus on one task and then switch to the other. The sequential task performance process is described as one where the analyst is performing some ongoing task (OT) where an alternative task(s) (AT) is available. The analyst decides either to continue with the OT or switch to an AT. The switch decision can be voluntary (i.e., task switching) or involuntary (i.e., interruption management). Many studies of sequential task performance (e.g., Monsell, 2003; Wylie & Allport, 2000; Meiran, 1996) use simple homogenous tasks such as classifying digits and focus on the fluency and costs of switching tasks, not on the choice to make the switch (for voluntary) or to address the interruption (for involuntary) (Wickens et al., 2013). Many complex, real-world scenarios and environments require switching between different types of tasks, some independent and some coupled. There is an interest in knowing how/why and when a person switches from one task to another.

Wickens et al. (2013) completed a meta-analysis of the task-switching and interruption management literature to identify variables that influence the choice to switch tasks and the strength of that influence. They derived six influence variables from the data: switch avoidance, task inertia difficulty effect and the effects of AT difficulty, priority, salience and interest. Five of these factors are built into the Strategic Task Overload Management (STOM) model that addresses longer duration multi-tasking situations (on the order of minutes to hours) and focuses on the decision of what task to perform (Wickens et al., 2015).

The model, shown in Figure 1, defines that each task-switch is based on multiple attributes, making the problem a multi-attribute decision. Each attribute has a polarity and some have a numerical weight. The model assumes that overloaded operators (i.e., more tasks than resources available) must decide whether to continue performing the OT or switch or one of several possible ATs, making the problem a multi-alternative decision. The attractiveness of ATs varies based on their attributes and the stickiness of the OT varies based on its attributes. The STOM model uses the attribute values to determine whether to continue with the OT or to switch to another task, and if to switch, which AT to switch to.

A meta-analysis found the switch-avoidance tendency, calculated to be 60%, as an important factor in the decision to continue with the OT or switch. Four other task attributes also moderate this tendency: task interest, priority, difficulty, and the salience of the AT (as compared to the OT). In addition to the switch-avoidance tendency, an operator is more likely to stay with an engaging, high-priority task. If the operator does choose to switch, they are more likely to choose the easier, more interesting, higher-priority, and higher salience task. It is interesting to note that the STOM model assumes an overloaded operator (Wickens et al., 2015), but many task-switching scenarios (e.g., the Project RED study) are not true overload scenarios because even though the tasks require a high workload and demand multiple resources, the scenarios do not provide an opportunity to perform tasks concurrently.

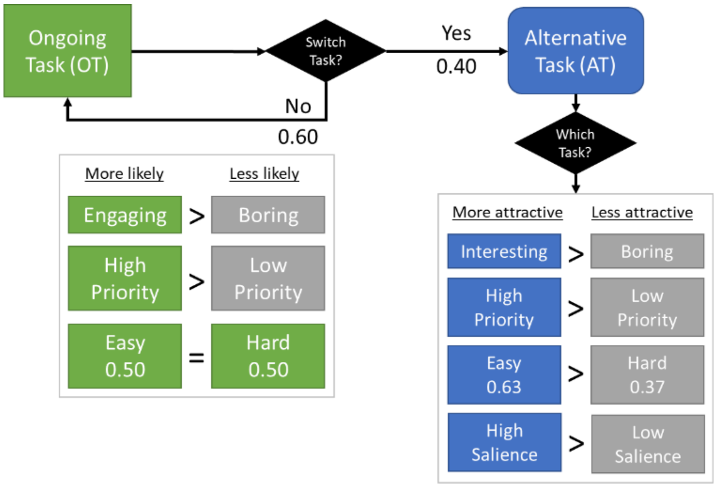


Figure 1. Strategic Task Overload Management model

1.2.3. Information Foraging Theory

Within Project RED, individuals had to combine existing knowledge with information provided by the environment to complete the overall task. Completing a larger task, composed of multiple smaller tasks, requires deciding which task to work on and when to switch to a different task. Information foraging theory explains how people gather and exploit information and provides an explanation of how people decide which task to work as well as why and when to switch.

Information foraging theory, based on optimal foraging theory from the animal foraging literature, proposes that people complete tasks either to explore and gather additional information or exploit the existing information (Pirolli & Card, 1999). The theory assumes that information foraging is embedded in the context of some other task (e.g., completing smaller tasks to find the best well design). People adapt their strategy or the structure of their environment, if possible, to maximize the amount of information gained per unit cost (e.g., time). They spend some amount of between-patch time getting to the next task (e.g., opening a file drawer or typing in a website) and some amount of within-patch time completing a task, until they decide to leave for a new one. Information patch models address how people allocate time, filter information, and complete enrichment activities in environments where information is encountered in clusters. Unlike animal foragers, information foragers can set up their environment (i.e., make frequently used sources quickly accessible) to improve the rate of information gain. Information scent models describe how people perceive the value, cost, or access path from proximal cues to navigate through a (physical or virtual) space to find a new (high-yield) patch or task. Information diet models determine how people decide how to select and pursue tasks to maximize the rate of gain of information relevant to their objective (Pirolli & Card, 1999). With all these models there is a tradeoff between exploiting the current task and exploring a new task.

Behavioral patterns (e.g., task-switching) are most commonly used to define exploitation and exploration (Mehlhorn et al., 2015); behavior that is stable over time, like remaining at a task, is interpreted as exploitative and behavior that is variable over time, like alternating between tasks, is interpreted as exploratory. Environmental, individual, and social factors all influence exploitation and exploration. Mehlhorn et al. (2015) propose that the decision to stay at a current patch or task or to leave for a new patch is not a tradeoff between exploitative and exploratory behavior, but instead a point on a continuum where the interpretation of the behavior as either exploitative or exploratory depends on the context in which it is considered. Figure 2 shows how exploration vs. exploitation can be thought of as a continuum along three dimensions: behavioral patterns, values and uncertainty related to the choice options, and outcome obtained from a choice (Mehlhorn, et al., 2015). When considered as a continuum, behavior is not seen as strictly stable or variable, but as a point somewhere between constantly switching and never switching. Instead of the behavior defining the strategy, consider exploitation as a strategy that is displayed behaviorally by remaining at a task over time and exploration as a strategy that is displayed by switching between tasks. The underlying degree of exploitation versus exploration as a strategy can also be represented in the choice outcomes and values and uncertainty in the choice options. Understanding what drives the strategy to exploit or explore the task also leads to understanding if a forager will remain at a task or switch to a new one.

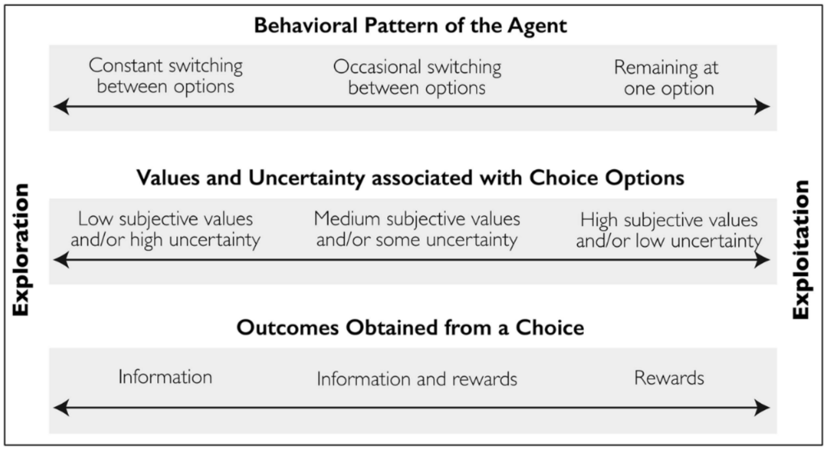

Behavioral Pattern of the Agent
Constant switching between options
Occasional switching between options
Remaining at one option
Values and Uncertainty associated with Choice Options
Low subjective values and/or high uncertainty
Medium subjective values and/or some uncertainty
High subjective values and/or low uncertainty
Outcomes Obtained from a Choice
Information
Information and rewards
Rewards
Exploration
Exploitation

Figure 2. Continuum of exploration to exploitation along three dimensions (Mehlhorn et al., 2015)

Not all foragers work as individuals; many work in teams. Project RED was set up to encourage participants to work in teams towards both individual and team goals. Foraging dynamics, such as exploration and exploitation, can be analyzed at the team level as well as the individual level. Looking again at the animal foraging literature, studies of social insects provide insight into the team dynamics of foraging, showing that task-switching is common in social insects. Empirical studies of social insects show that as conditions change, individuals decide whether or not to be active and which task to perform. These individual decisions generate the dynamics of group behavior, the number of individuals actively engaged in each task at any moment. Much of the theoretical work on social insects examines how individuals are allocated to components of a (larger) task. All share the basic idea that an individual's behavior depends partly on its assessment of its environment and partly on its interactions with other individuals (Pacala et al., 1996).

1.2.4. Decision-making Strategies

There are different possible approaches to solving a problem like the one presented as part Project RED (e.g., well design and placement) where there are multiple sources of information to use in solving the problem. A forager can exhaustively search those sources to gather all the available information or can perform a limited search to look for what are considered to be the key pieces of information needed to solve the problem. The information search strategy is tied to how the performer makes their decision. Compensatory decision-making strategies are those that seek to optimize the solution by processing all relevant information and trading off the good and bad aspects of each alternative. A compensatory strategy, by assuming unlimited time and mental resources, leads to more information seeking behavior to complete an exhaustive search. Some examples include normative theories based on mathematical models, such as Bayes' theorem and expected utility, and mental models.

Alternatively, non-compensatory strategies typically reduce information processing demands by ignoring potentially relevant problem information. They seek to find a solution that is good-enough or one that satisfices (Simon, 1956). This could be due to a trade-off between the cost of search and the benefit provided by the additional information (J. Payne et al., 1988) or because the limited information is all that is needed to solve the problem. Fast and frugal heuristics (Gigerenzer, 2008) ignore unnecessary information in an ecologically valid environment leading to limited search and less information seeking behavior. This is consistent with information foraging theory (Pirolli & Card, 1999), where people select certain types of information.

Research to evaluate decision-making strategies typically tries to determine under what scenarios (i.e., type of environment and task) a particular strategy is used (for example, take the best (TTB) vs. tally vs. weighted additive (WADD) vs. guess, Lee et al., 2019; TTB, Newell & Shanks, 2003; recognition heuristic, Oppenheimer, 2003; WADD vs. TTB, Rieskamp & Otto, 2006). In these studies, the type of task presented is selected to test the use of one or more particular pre-identified strategies. The many individual differences in strategy selection led to the development and use of models to identify and classify strategy use (Lee et al., 2019). Most studies, and associated models, only use choice data and assume that a person's strategy is fixed over the duration of the task, although Lee et al. (2019) found evidence of strategy switching using choice, search, and reporting information. As the overall task and accompanying smaller tasks in Project RED were not selected to test pre-defined strategies, the analysis methods traditionally used to determine decision-making strategy could not be applied to this dataset. Instead, this research used a machine learning approach to identify clusters of similar participants, assuming that these participants were using similar strategies in their decision-making while completing the overall task.

1.2.5. Decision-making Cognitive Process Models

Cognitive models are used for many types of decision making, such as perceptual decisions (Ben-David et al., 2014), preferential choices (Busemeyer & Townsend, 1993), and risky choices (Johnson & Busemeyer, 2010), to explain empirical results with a common set of psychological principles. These models provide the capability to apply cognitive mechanisms to better understand observed behavior, especially differences in behavior under different conditions or by different groups. Task-switching between smaller tasks to achieve a larger goal, like in the Project RED dataset, is a type of multi-alternative, multi-attribute decision. It is reasonable to use evidence accumulation models of decision making for the decision of task-switching.

Evidence accumulation models, also known as sequential sampling models, are a class of computational models that describe the decision-making process as the accumulation of evidence over time for each response option, where the response option that reaches the response boundary (i.e., threshold) first is selected. Most are implemented as noisy diffusion processes. Examples include the diffusion decision model (DDM; Ratcliff, 1978), the leaky competing accumulator model (LCA; Usher & McClelland, 2001), decision field theory (DFT; Busemeyer & Townsend, 1993), and the linear ballistic accumulator (LBA; Brown & Heathcote, 2008). Several evidence accumulation models have a multi-alternative, multi-attribute version that have been applied to preferential, risky, and perceptual choice problems, including decision field theory (MDFT; Roe et al., 2001), the linear ballistic accumulator (MLBA; Trueblood et al., 2014) and the leaky competing accumulator (MLCA; Usher & McClelland, 2004).

The parameters of the evidence accumulation models describe inhibition (or caution) and efficiency of the process as well as any bias towards an alternative, with the different models assuming different mechanisms to account for observed behaviors. MDFT, MLCA, and MLBA can each be separated into three stages describing how objective attribute values (i.e., process input) for each alternative are mapped to subjective representations, how attention is allocated across attributes, and how alternatives compete until some threshold amount of evidence accumulates to form the decision (i.e., process output) (Turner et al., 2018).

2. DATA

This research used previously collected data from a semi-controlled real-world design and planning task, known as Project RED, where participants worked individually and in teams to solve a problem. There were multiple smaller tasks they could switch between, as they wanted, to complete the overall task. COVID-19 restrictions severely limited the options for new data collection in 2020 so one advantage of using the Project RED data was that it was already available to test the hypotheses for this research. The Project RED data is described in the first section of this chapter – its composition and associated measurements as well as how it was collected. The decision strategies for completing the overall task were not controlled or measured, which was a limitation of the data. Another limitation was the small number of switches between tasks during the overall task. These limitations are mentioned here, but discussed in more detail while describing the methodological approach, in chapter 3. The remainder of this chapter describes an exploratory analysis that was completed as part of this research to identify additional residual, derived predictors to include in the analysis.

2.1. Project RED

The Project RED dataset contained performance and task-switching data for 192 participants working in teams to solve the problem of finding the best location and design, as determined by different criteria for different roles, for a new well on the Martian surface. This was a simulated task that was originally part of a study that examined team task transitions while working in space. Some of the participants were in an isolated simulated space-vehicle environment on a hypothetical mission to Mars (i.e., the Martian crew) while the others were part of the Earth-bound Mission Control Center (MCC) ground team. The participants performed multiple different tasks and used multiple different tools to assist them in completing the overall task (i.e., build a well on the Martian surface). For this research, *task* refers to these multiple different tasks available to participants to complete the *overall task* of solving the well design and placement problem. There were 15 tasks (6 individual, 6 team, and 3 multi-team) mapped to 6 tools (e.g., completing task 0 requires using tool 1) for the participants to use as they wanted for completing the overall task. Project RED did not measure, manipulate, or control what strategies participants used to solve the overall task or their associated task-switching behavior. For each task there were values specified for the four task attributes that contribute to task-switching: task difficulty, priority, interest, and the salience of the alternative task compared to the original task (Wickens et al., 2015). Task-switching data was available for each second of the overall task while task performance data was only reported at the completion of the overall task. Table 1 provides a description for each task, identifies if it is an individual, team, or multi-team task, and lists the tool(s) used to complete the task. The experimental setup required that tasks be completed sequentially.

12 people participated in each session to complete the overall task, with each person assigned to a specific role. The participants were split into four 3-member teams, where one team member was part of the Martian crew and the other two were part of the Mission Control Center (MCC) located at a university. The Martian crew inhabited NASA’s Human Exploration Research Analog (HERA), a three-story habitat that served as an analog for isolation, confinement, and remote conditions in exploration scenarios. Four of the 12 participants were part of the Martian crew and the other eight participants were part of the

MCC for each session. The Martian crew members remained the same for all the sessions within a mission, but the MCC members changed each session.

Table 1. Task descriptions and tool mapping

Task	Description	Type	Tools used
0	Building an understanding of drivers of personal performance; Developing role expertise	Individual	Info database
1	Advocating for outcomes important to role	Individual	Chat window
2	Investigate personal performance in new locations	Individual	Map interface, Decision calculator
3	Revisit/update personal outcomes in old locations	Individual	Map interface, Decision calculator
4	Sharing expertise	Individual	Chat window
5	Providing feedback on team	Individual	Popup survey
6	Understand team variable associations with different land characteristics/locations	Team	Decision calculator
7	Investigate team outcomes in new locations	Team	Map interface, Decision calculator
8	Revisit/update team outcomes in old locations	Team	Map interface, Decision calculator
9	Exchange information with teammates	Team	Chat window
10	Advocate importance of team outcomes to other teams	Team	Chat window
11	Negotiate with each other about training decisions (Only HF team)	Team	Chat window
12	Reach a final decision	Multi-team	MTS signoff
13	Exchange information about team constraints to other teams (request, provide, elaborate)	Multi-team	Chat window
14	Decide on a well location	Multi-team	Chat window

An overview of Project RED is shown in Figure 3. The study ran over two different campaigns, with campaign 3 consisting of four 30-day missions, each with three sessions, and campaign 4 consisting of five 45-day missions, each with four sessions. The Martian crew changed for each mission. In campaign 3 participants had 1800 seconds to complete the overall task whereas they had 2700 seconds in campaign 4. Session 1 was towards the beginning of the mission (i.e., mission-day 9) when the Martian crew was hypothetically still close to Earth so participants did not experience any communication delay (i.e., normal conditions). In session 2 the Martian crew was further from Earth (i.e., mission-day 16) so participants experienced a 60-second communication delay. The delay was only for communications between HERA and the MCC; Martian crew members within HERA never experienced a delay when communicating within the crew and members within the MCC never experienced a delay when communicating within mission control. For session 3 within campaign 3, the Martian crew were again close to Earth so there was no communication delay, but it was towards the end of their mission (i.e., mission-day 28) where they had experienced extended social isolation. For campaign 4, participants experienced a longer 180-second communication delay in session 3 and no communication delay in session 4. Session 4 occurred towards the end of the campaign 4 missions so, again, the Martian crew members had experienced an extended period of social isolation. Data were provided from 12 HERA crew members and 48 participants acting as mission control in campaign 3, and from 20 HERA crew members and 112 participants acting as mission control in campaign 4. Since HERA crew participants remained the same over multiple sessions the overall task was completed a total of 240 times by these 192 participants. This was referred to as the ‘not withheld’ data for this research. An additional 84 observations related to the counterfactual questions completed over 7 sessions by 20 HERA crew and 56 MCC participants were withheld (referred to in this research as the ‘withheld’ data). The ‘withheld’ data were used to predict residual parameters with the machine learning and cognitive process models, built using the ‘not withheld’ data. The results from these models were then used in the Bayesian generalized linear models to generate responses to the counterfactual questions.

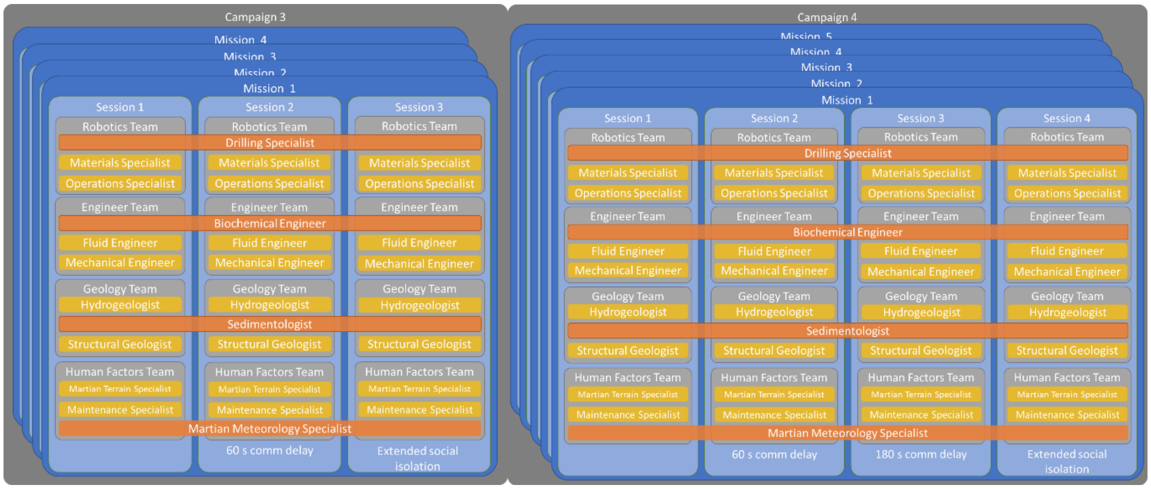


Figure 3. Overview of Project RED. Each campaign is dark grey, each mission with a campaign is dark blue, each session within a mission is light blue, and each team within a session is light grey. Participants in light orange are different for each session while participants in dark orange remain the same across all sessions within a mission.

Each of the four teams had a different primary objective for designing and placing the well, listed in Table 2. Each participant's and each team's performance were measured against their primary objective. The four teams also worked together to determine a plan for the location and design of a well to support as large a colony on Mars as possible. The set-up of the overall task gave each individual unique information so participants needed to coordinate both within and across teams in order to find a suitable location and design for the well. Task performance was evaluated differently for each team and for each of the roles. The original researchers determined the calculation of the performance scores; they were calculated using an unidentified function of the parameters that the participants chose. The participant had access to a decision calculator to preview their scores based on different parameter values, but only received their final individual, team, and multi-team system scores once at the end of the overall task based on the decisions that everyone made. Since the scales of the scores were different for different roles, as shown in Table 2, the scores were normalized to range from 0 to 1 for use in this research. This research focused on task-switching behavior as the outcome of interest because there was a larger amount of better-defined data available. It was not possible to simulate or replicate the task performance scores using the information provided with the dataset. The performance scores were only included as an outcome for the counterfactual prediction models.

Table 2. Individual role and team descriptions

Role	Location	Team	Performance Objective	Performance Score
Drilling Specialist	HERA	Robotics	Develop well construction plan that minimizes the total direct cost	0 - infinity, lower is better
Materials Specialist	MCC			
Operations Specialist	MCC			
Biochemical Engineer	HERA	Engineer	Design well to maximize total clean water output	0 - 1, higher is better
Fluid Engineer	MCC			0 - 1, higher is better
Mechanical Engineer	MCC			0 - infinity, higher is better
Hydrogeologist	MCC	Geology	Find location to maximize water available	0 – 397677, higher is better
Sedimentologist	HERA			0.035 - 1, higher is better
Structural Geologist	MCC			485.331 – 166399, higher is better
Martian Terrain Specialist	MCC	Human Factors	Minimize terrain cost	0 - infinity, lower is better
Maintenance Specialist	MCC			
Martian Meteorology Specialist	HERA			

The study was structured by Mesmer-Magnus et al. (2020) to include factors that the experimenters believed affect completing a variety of tasks in space while interacting with multiple individuals, teams, and types of tools. Data was collected for parameters that encompassed 5 different types of factors: task characteristics, social factors, technology affordances, situational constraints, and individual (personality) differences. The task characteristics were quantitative measures of Wickens' task attributes – *difficulty, interest, priority*, and *salience* – as well as the *interdependence* of the task. The interdependence measure captured if a task was performed solo (i.e., individual task), within a team (i.e., team task), or across teams (i.e., multi-team tasks). Social factors were included because individuals in space are members of multiple teams, where the members of the teams can change. For example, the MCC members within Project RED changed each session, but the Martian crew did not. *Behavioral ties* and *interpersonal ties* were measured for each session. Participants were asked several times during Project RED to answer the question "Who is a valuable source of information?" to measure behavioral ties and to answer the question "Who do you enjoy working with?" to measure interpersonal ties, where each pair of participants either are (1) or are not (0) related. *Team mental models* were also measured using pairwise comparisons of how participants motivated one another, coordinated work, managed conflict, monitored team progress, and shared information. Participants rated the extent to each pair of items were related to achieving the goals of Project RED on a scale of 1 (totally unrelated) to 7 (very strongly related). These ratings were used to calculate the Euclidean distance between each pair of participants' responses to determine the sharedness of each pair's team mental models. Two measures of technology affordance were thought to be relevant to working in space and were included in the data: *editability* and *association*. Editability is how much control the content creator has over their information over time, determined by the extent to which a tool allows users to modify or revise their content. Association refers to the extent to which a tool establishes connections among individuals or between individuals and content. Situational constraints were determined using the HERA crews' mission scenario. *Communication delays* occurred during the missions when the HERA was farther from Earth. The Martian crew was *socially isolated* on the HERA for an extended period of time for the later missions. The Big Five dimensions of personality – *conscientiousness, extraversion, agreeableness, openness, and* *neuroticism* – were measured for each session to account for characteristic differences for each participant.

2.2. Exploratory Analysis

This research included an exploratory data analysis to identify observed and derived factors that affected task-switching behavior and overall task performance, primarily focusing on task-switching behavior. Plotting the data using several different data visualization techniques helped identify patterns and provided insight into team dynamics and relationships between different factors. The exploratory analysis included line plots to capture time-series data, scatter plots to show correlations between two variables, and chord diagrams to visualize flow between entities.

The analysis and visualization of the connection of task A to task B, defined as a task pair, using a chord diagram showed that the top 10% of task pairs account for 50% of all the tasks completed. It also showed dependencies between some of the tasks, meaning that some tasks are more likely to be followed by another task, and these dependencies were consistent between campaigns 3 and 4. The chord diagram in Figure 4 shows the flow from task A to task B for the top 10% of task pairs from all the sessions in campaign 3 and campaign 4. The task dependencies quantified the number of times that one task was followed by another task. The color of the chord matches the color of task A and the thickness of the chord shows the strength of the dependency (i.e., number of times A-B occurred). There is a sector for each task, where the size of the sector was determined by the number of times that task was completed.

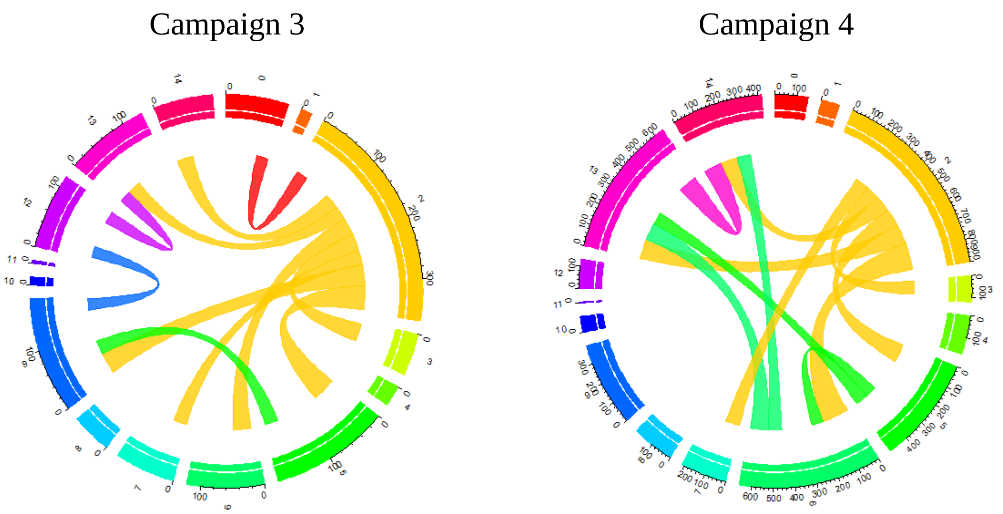


Figure 4. *Chord diagram of flow from task A to task B for top 10% of task pairs using all 'not withheld' data. Each sector shows how many times task A was completed. Each task A, and its matching chord, is colored differently. The thickness of the chord shows the number of times that a switch from task A to task B occurred.*

The connections of when participant A and participant B were concurrently completing the same team or multi-team task was defined as a functional tie. The functional ties quantified dependencies between participants as the number of seconds that each participant completed any team or multi-team task concurrently with another participant. A similar analysis and visualization of the functional ties (Figure 5) showed that the top 25% of concurrent task completion account for 50% of all the time spent by participants on concurrent tasks, meaning that the top 25% of concurrent tasks occurred for longer periods of time than the other 75% of concurrent tasks. Additionally, the plots showed that some participants spend a greater amount of time completing tasks concurrently with other participants while some participants spend very little time working on team and MTS tasks concurrently with other participants. The plots were limited to the top 25% of concurrent task completion for plot readability and to a single session as MCC participants changed each session. The color of the chord matches the color of participant A and the thickness of the chord shows the strength of the dependency (i.e., number of seconds that A and B completed the same task concurrently). The size of the sector was the number of seconds that an individual participant worked the same task concurrently as any other participant. Plots for the other sessions are available in Appendix A.

Because multiple participants were able to work on the same task concurrently the total number of seconds for each participant could be larger than the total number of seconds available to complete the overall task. One limitation of this visualization was that it matched the team tasks between all 12 participants in a session rather than limiting those matches to the smaller 3-person teams so the diagrams include participants from different teams that were concurrently working on a team task, even though they were not necessarily working in coordination with one another.

Campaign 3, Mission 2, Session 1 Campaign 4, Mission 5, Session 1

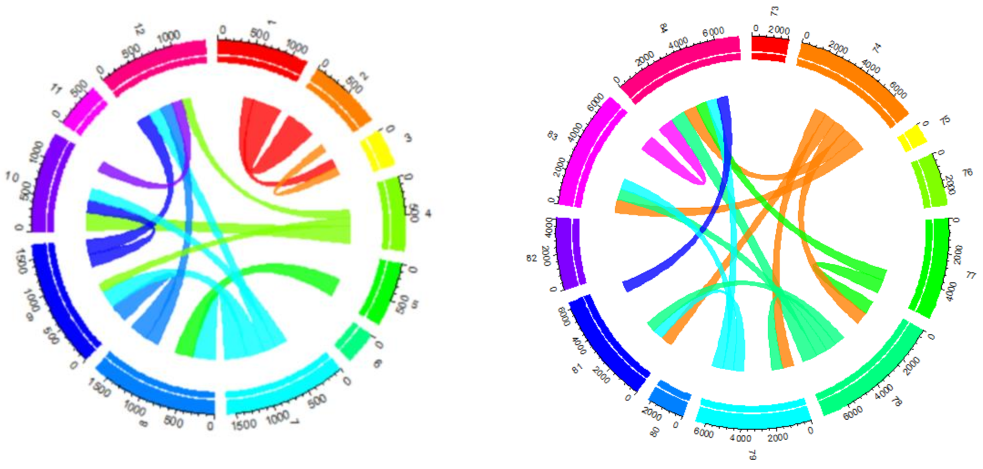
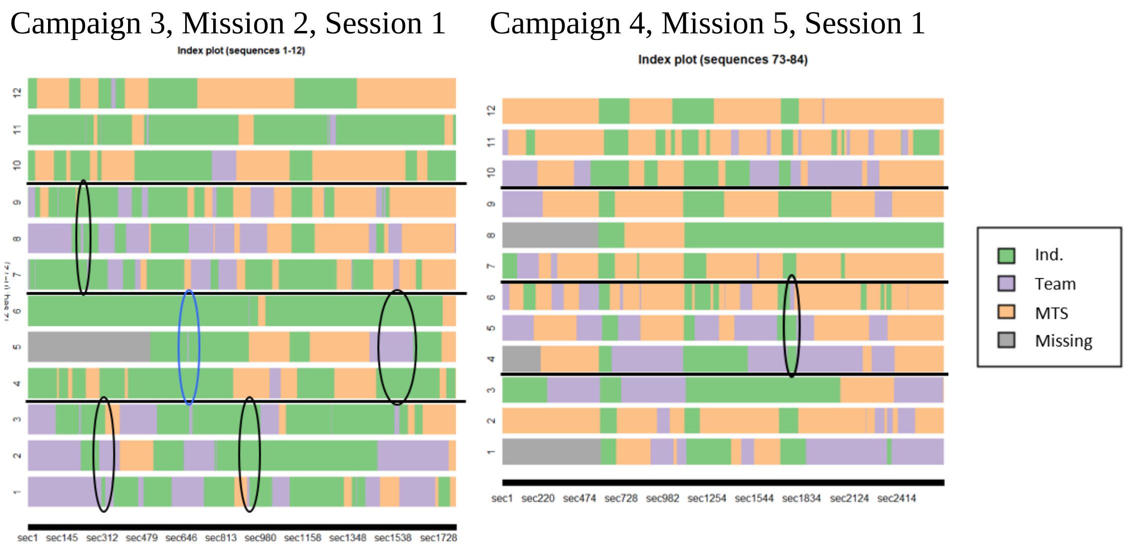


Figure 5. *Chord diagram of connections between participant A and participant B for top 25% of concurrent task completion using all ‘not withheld’ data. Each sector shows how many seconds participant A worked on the same task as any other participant in their session. Each participant A, and their matching chord, is colored differently. The thickness of the chord shows the amount of time (in seconds) that participant A and participant B completed the same team or multi-team task.*

Sequence plots visualize data to show patterns between different participants for tasks completed over time. The plots in Figure 6 show examples where only a single participant within a team completed a team task, while the other 2 team members performed individual tasks, which led to uncertainty in whether the type of task (i.e., individual, team or MTS) and the participant dependencies accurately described how participants worked together to actually perform the overall task. For example, in Figure 6a at *t*=672 sec, subject 5 performed “investigate team outcomes in new locations,” a team task, while the other two people on the team, subjects 4 and 6, performed the survey task, an individual task, illustrating that team tasks could be performed independently. There were also many times when a participant was the only member of a team completing a team task, but the other members were completing multi-team tasks. It was hard to determine in these cases if the participants were working together or not. The method of quantifying participant dependencies described above included these times as concurrent task completion, but there may be less interaction between participants and less dependency on other participants’ behavior for task-switching than quantified by the participant dependency values. Sequence diagrams for the other sessions are available in Appendix A.



Campaign 3, Mission 2, Session 1

Campaign 4, Mission 5, Session 1

Figure 6. *Sequence diagram of task type performed by each participant. The blue oval shows where subject 5 performed at team task at t=672 s while the other two people on the team, subjects 4 and 6, performed an individual task. The black ovals highlight other examples of a team task being performed by only one person on the team. The examples do not include all incidents of this occurring.*

The Project RED dataset contained many recorded variables including personality traits, tool used at each second, task completed at each second, participant role, task attributes, tool characteristics, and team dynamics information like interpersonal ties and behavioral ties. Additional predictors, including time on task, task dependencies, participant dependencies, and typical and atypical task switches were derived from these, to use in the machine learning algorithms for identifying patterns of switching and decision strategies. Time on task was calculated as the total amount of time (in seconds) that a participant spent on each task. The scatter plot in Figure 7 shows that there was a strong non-linear relationship between the average amount of time spent on all tasks and the rate of task-switching (i.e., $\frac{n_{switches}}{t_{overalltask}}$). These factors were also correlated (r = -0.73, n=240). Time on task could be thought of as another way to represent task-switching (e.g., a measure of task-staying behavior) and was excluded from the machine learning analysis and subsequent counterfactual predictions.

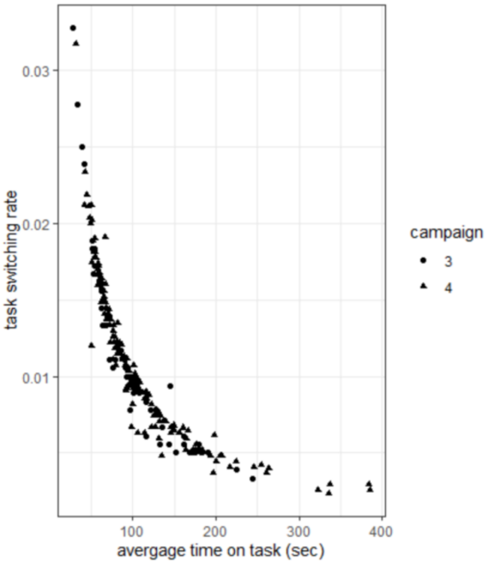


Figure 7. Relationship between a participant's average time on task and their rate of task-switching.

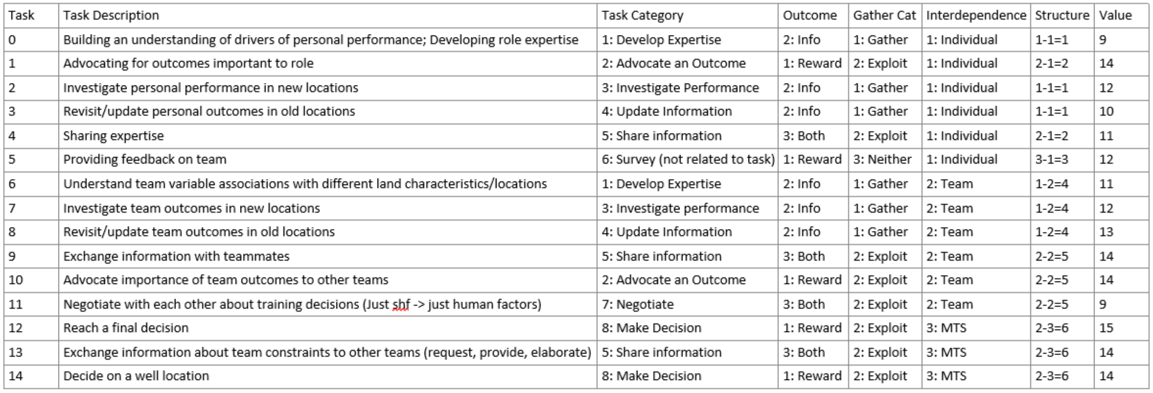
This research developed a measure of the typicality of switching from task A to task B for all participants as the proportion of switches from task A to task B out of the total number of switches. This weighted how typical each switch was across all participants, with a higher value indicating that a switch occurred more frequently. The proportions for every switch pair were then summed for the switches completed by each participant providing a measure of how much that participant completed the most popular switches. A higher value indicated that a participant completed more of the popular switches.

This research also calculated another measure of the atypicality of switching from task A to task B as the inverse of the number of switches from task A to task B (i.e., $^{1}/_{switch_{A-B}}$), with a higher value indicating that the switch occurred less frequently.

These values were then summed for the switches completed by each participant providing a measure of how much that participant completed the least popular switches. This measure was correlated to the measure of typical switches with a Pearson's r of 0.21 (n=240).

Several derived predictors used information foraging theory to inform the predictor names and appropriate values to assign to these predictors. Someone who explores will try out more different tasks leading to a derived predictor that assigned each participant as an explorer or not depending on how much they completed atypical switches. The top 25% of atypical switchers were labeled as an explorer (1) while the bottom 75% were labeled as not-explorers (0). Also, each of the 15 tasks were categorized into 1 of 8 categories: develop expertise, advocate for an outcome, investigate performance, update information, share information, complete the survey, negotiate, or make a decision, as shown in Table 3. These categories informed the characterization of each task as a gathering task, an exploitation task, or neither. Tasks that developed expertise, investigated performance, or updated information were characterized as gather tasks while tasks that advocated for an outcome, shared information, negotiated, or made a decision were characterized as exploitation tasks. The task to complete the survey was characterized as neither since it was not necessary for solving the overall task. The outcome provided by each task was also characterized as either information, a reward, or both. Tasks that developed expertise, investigated performance, or updated information provided information while advocating for an outcome, making a decision, or completing the survey provided a reward and sharing information or negotiating provided both information and a reward. The task structure was characterized using the combination of the gather/exploit category and the type (or interdependence) of the tasks. Individual gather tasks were labeled as structure 1, individual exploit tasks as structure 2, individual neither tasks as structure 3, team gather tasks as structure 4, team exploit tasks as structure 5, and MTS exploit tasks as structure 6. Finally, the value of each task was determined as the sum of the individual attribute values for each task. This factor was included in both the machine learning algorithms and the information theory based cognitive process models.

Table 3. Summary of derived predictors



Task	Task Description	Task Category	Outcome	Gather Cat	Interdependence	Structure	Value
0	Building an understanding of drivers of personal performance; Developing role expertise	1: Develop Expertise	2: Info	1: Gather	1: Individual	1-1=1	9
1	Advocating for outcomes important to role	2: Advocate an Outcome	1: Reward	2: Exploit	1: Individual	2-1=2	14
2	Investigate personal performance in new locations	3: Investigate Performance	2: Info	1: Gather	1: Individual	1-1=1	12
3	Revisit/update personal outcomes in old locations	4: Update Information	2: Info	1: Gather	1: Individual	1-1=1	10
4	Sharing expertise	5: Share information	3: Both	2: Exploit	1: Individual	2-1=2	11
5	Providing feedback on team	6: Survey (not related to task)	1: Reward	3: Neither	1: Individual	3-1=3	12
6	Understand team variable associations with different land characteristics/locations	1: Develop Expertise	2: Info	1: Gather	2: Team	1-2=4	11
7	Investigate team outcomes in new locations	3: Investigate performance	2: Info	1: Gather	2: Team	1-2=4	12
8	Revisit/update team outcomes in old locations	4: Update Information	2: Info	1: Gather	2: Team	1-2=4	13
9	Exchange information with teammates	5: Share information	3: Both	2: Exploit	2: Team	2-2=5	14
10	Advocate importance of team outcomes to other teams	2: Advocate an Outcome	1: Reward	2: Exploit	2: Team	2-2=5	14
11	Negotiate with each other about training decisions (Just shf -> just human factors)	7: Negotiate	3: Both	2: Exploit	2: Team	2-2=5	9
12	Reach a final decision	8: Make Decision	1: Reward	2: Exploit	3: MTS	2-3=6	15
13	Exchange information about team constraints to other teams (request, provide, elaborate)	5: Share information	3: Both	2: Exploit	3: MTS	2-3=6	14
14	Decide on a well location	8: Make Decision	1: Reward	2: Exploit	3: MTS	2-3=6	14

A correlation matrix of all measured and derived predictors (Table 36, in Appendix A) showed which variables have the strongest measure of linear association (i.e., highest Pearson r values) when compared to the ID of the participant completing the overall task and their switch rate. A relatively small number of parameters, shown in bold, have values greater than 0.20 or less than -0.20, which are generally considered to indicate a moderate or strong correlation between the variables. The four task attributes (i.e., salience, interest, priority, and difficulty), time on task, measure of typical switches, whether a participant is an explorer, neuroticism, and the interpersonal ties to participants assigned to role 7 were the most correlated to participant switch rate. The full set of predictors used in the decision strategies analysis contained 94 variables – all the variables in Table 36 except the personality factors, time on task, and the variables estimated using the cognitive process models.

3. METHODOLOGICAL APPROACH

This research used multiple modeling approaches to address the proposed research questions. A machine learning approach was used to identify the decision strategy used to complete the overall task and the most important factors contributing to that strategy to address the first research question, a cognitive modeling approach was used describe and explain the cognitive mechanisms used in completing the overall task and the associated task-switching behavior to address the second and third research questions, and a Bayesian generalized linear modeling (GLM) approach was used to generate counterfactual predictions of task-switching rates and overall task performance scores to address the final research question. These approaches fit together to form an overall analysis pipeline shown in Figure 8. There were multiple options for how each of the modeling approaches could be completed; the decisions leading to the methodologies included in this research were based on theories or evidence from previous research along with the constraints and objectives of this research. A primary focus of the research was to identify models that best describe and explain decision strategy and cognitive mechanisms for task switching through model comparison, and to generate models that provide the best out-of-sample predictions of decision strategy and counterfactual predictions of task-switching rates and overall performance scores. This chapter discusses, separately for each modeling approach, the methodology used to test the hypotheses, including the reasons and justification for using these methods.

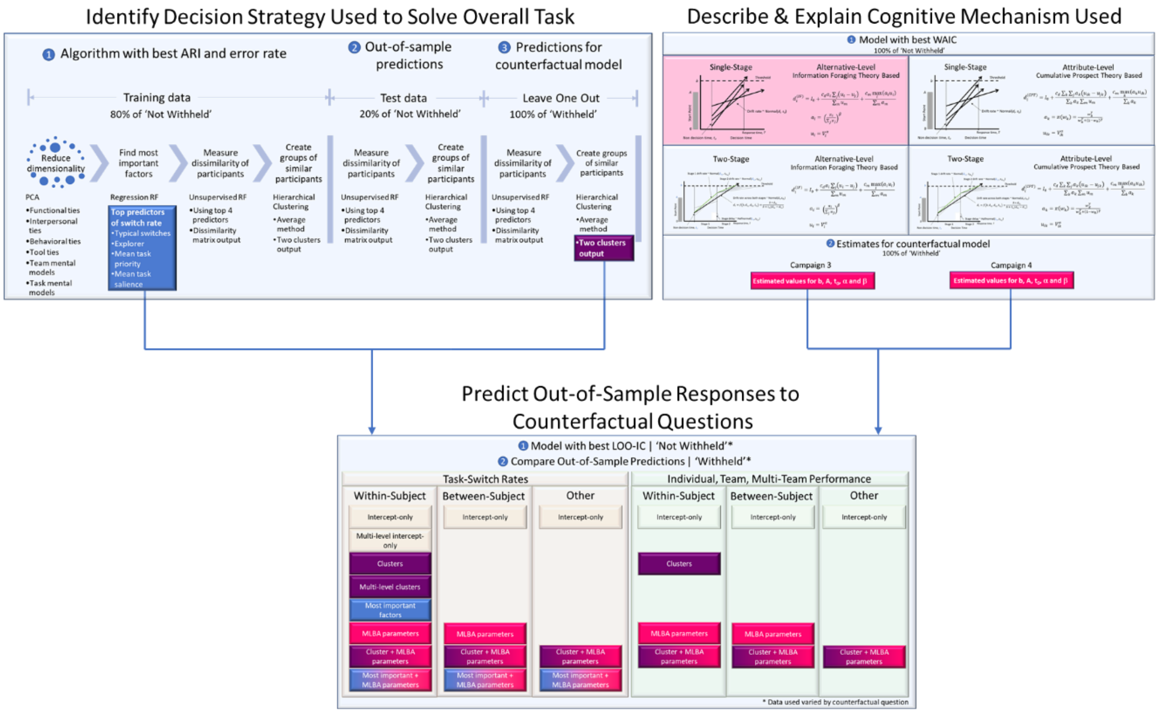

Identify Decision Strategy Used to Solve Overall Task
1 Algorithm with best ARI and error rate
2 Out-of-sample predictions
3 Predictions for counterfactual model
Training data
80% of 'Not Withheld'
Test data
20% of 'Not Withheld'
Leave One Out
100% of 'Withheld'
Reduce dimensionality
Find most important factors
Measure dissimilarity of participants
Create groups of similar participants
Measure dissimilarity of participants
Create groups of similar participants
Measure dissimilarity of participants
Create groups of similar participants
PCA
•Functional ties
•Interpersonal ties
•Behavioral ties
•Tool ties
•Team mental models
•Task mental models
Regression RF
Top predictors of switch rate
•Typical switches
•Explorer
•Mean task priority
•Mean task salience
Unsupervised RF
•Using top 4 predictors
•Dissimilarity matrix output
Hierarchical Clustering
•Average method
•Two clusters output
Unsupervised RF
•Using top 4 predictors
•Dissimilarity matrix output
Hierarchical Clustering
•Average method
•Two clusters output
Unsupervised RF
•Using top 4 predictors
•Dissimilarity matrix output
Hierarchical Clustering
•Average method
•Two clusters output
Describe & Explain Cognitive Mechanism Used
1 Model with best WAIC
100% of 'Not Withheld'
Single-Stage
Alternative-Level
Information Foraging Theory Based
Single-Stage
Attribute-Level
Cumulative Prospect Theory Based
Two-Stage
Alternative-Level
Information Foraging Theory Based
Two-Stage
Attribute-Level
Cumulative Prospect Theory Based
Threshold
Start Point
Non decision time, t_0
Decision time
Response time, T
2 Estimates for counterfactual model
100% of 'Withheld'
Campaign 3
Estimated values for b, A, t_0, α and β
Campaign 4
Estimated values for b, A, t_0, α and β
Predict Out-of-Sample Responses to Counterfactual Questions
1 Model with best LOO-IC | 'Not Withheld'*
2 Compare Out-of-Sample Predictions | 'Withheld'*
Task-Switch Rates
Individual, Team, Multi-Team Performance
Within-Subject
Between-Subject
Other
Within-Subject
Between-Subject
Other
Intercept-only
Multi-level intercept-only
Clusters
Multi-level clusters
Most important factors
MLBA parameters
Cluster + MLBA parameters
Most important + MLBA parameters
* Data used varied by counterfactual question

Figure 8. Overall Data Analysis Pipeline. Results from the best decision strategies model and the best cognitive process model feed into the counterfactual prediction models.

3.1. Decision strategies

Machine learning techniques are commonly used to reduce the dimensionality of large data, to predict out-of-sample responses when labeled data is available, and to identify similarities between unlabeled data observations. This research used multiple machine learning techniques, leveraging the strengths of each method, to determine the factors important to participants' decision-making strategies and to group participants that use similar strategies. The clusters determined with the machine learning algorithm were compared to randomly assigning participants to a group to test the hypothesis that the machine learning results better predict out-of-sample data than a random assignment model.

This research assumed that the decision-making strategy used by the individual to solve the overall task related to their task-switching behavior. Task-switching behavior, which can be displayed in varying degrees, was a result of an individual's decision-making strategy for information foraging. From an information foraging perspective switching tasks was associated with exploration whereas remaining at one option or task was associated with exploitation. This research assumed that people that used a similar strategy clustered together and exhibited similar task-switching rates. The data used in this analysis did not include any measure of decision strategy, supporting the use of unsupervised learning techniques to determine the participants' strategies. However, the data did contain a measure of the switch rate, and based on the assumption that task-switching behavior relates to the decision strategy, a supervised approach was used to identify factors important to the strategies based on the task-switching behavior.

The models used the combined 'not withheld' data from both campaigns where data from 80% of the overall task completions (n=192) was randomly assigned as the training dataset and the remaining 20% (n=48) was used to test the models. Three different sets of factors were used: one that included all the predictors (as described in section 2.2) without any principal components analysis (PCA) reductions, one that reduced the number of predictors using PCA, and one that reduced the number of predictors using PCA but also included the personality factors. The full set predictors increased the possibility of finding similarities between the participants while the reduced set of predictors decreased the possibility of overfitting. The results from each set of predictors were compared using in-sample and out-of-sample predictions to determine the best set of predictors to identify clusters of similar participants.

All the datasets included both individual and team factors that fit into several different categories: demographic information, task characteristics, social factors, technology affordances, situational constraints, personality measures, and derived predictors (as described in section 2.2). The only demographic information available was the participant's gender. The task characteristics were Wickens' task-switching attributes, described in section 1.2, with values assigned for each task by the original researchers. Task interdependence was not included in the machine learning modeling since it was directly related to the task structure. The social factors were the behavioral and interpersonal ties and shared mental models of the teams and the tasks, described in section 2.1, which were measured during the task and provided with the dataset. Each participant had 12 values for each of these measures, one for every other participant in their session and a null value for themselves. The technology affordances included editability, association, persistence, and visibility, as described in section 1.2. Situational constraints were the length of communication delay experienced in different sessions by all participants; the social isolation experienced by the HERA participants over the course of multiple sessions within a mission, which was captured using the session number; the participant's role; the mission number; and the campaign number. The personality factors included agreeableness, conscientiousness, extraversion, and neuroticism. Byrne et al. (2015) show that neuroticism and agreeableness negatively affected decision-making under pressure and multi-tasking research suggests that conscientiousness (Messmer-Magnus et al., 2020) and extraversion (Sanderson, 2012) predict an individual's level of comfort with multi-tasking and their motivation to switch among tasks. However, these measures were collected differently in an unknown manner (i.e., the values were on different scales) where the campaign 3 values ranged from 0 to 1, but the campaign 4 values ranged from 2.5 to 4. Additionally, the data were missing measurements for all the MCC participants in campaign 4. This research attempted to standardize and normalize the values between the two campaigns, and to impute the missing campaign 4 values, in order to include personality measures in the machine learning algorithms, but the results in section 3.1.1 show the attempt was unsuccessful. The derived factors were described in section 2.2; they included the explorer categorical variable, mean value of gather or exploit tasks, mean outcome value, mean task category, mean task structure, typical switches, and vertical switches (measured as a percentage of all switches).

3.1.1. Data Reduction

Principal components analysis (PCA) uses a linear transformation to create a new representation of the data, which yields a set of linearly uncorrelated orthogonal axes (i.e., the principal components). The first principal component is the direction that captures the largest variance in the data. The second principal component also finds the maximum variance in the data; it is completely uncorrelated to the first principal component, yielding a direction that is orthogonal to the first component. This process repeats based on the number of dimensions, where each next principal component is the direction orthogonal to the prior components with the most variance (Shlens, 2014).

This research ran PCA on six different subsets of the data to reduce the number of parameters describing relationships between each of the 12 participants – behavioral ties, participant task matches (i.e., participant dependencies), interpersonal ties, task mental models, team mental models and participant tool matches. Every participant had 12 values for each subset, to describe their relationship with the other 11 participants in the session and a value for themselves (mostly this is zero). The results from the model built using the PCA reduced dataset were compared to results using the full dataset to determine the best dataset to identify clusters of similar participants.

The behavioral ties, interpersonal ties, task mental models, and team mental models, described in section 2.1, were determined by the researchers that collected the Project RED data and provided with the dataset. Participant task matches were described in section 2.2 and visualized in Figure 5. Participant tool matches were calculated the same way, but using the number of seconds that each participant within a session used the same tool as another participant. The matches were limited to when participants were performing team or multi-team tasks and were using the chat box, map interface, decision calculator, or multi-team signoff box. A separate PCA (centered and scaled) was run for each set of 12 parameters (e.g., PCA on behavioral ties was run separately from PCA on interpersonal ties) with the resulting first principal component including enough variance for each set of parameters, as shown in section 4.1, that only the PC was used in the machine learning techniques. The PCA reduced the number of predictors describing relationships between the participants from 72 to 6.

Classical multi-dimensional scaling (cMDS), also known as principal coordinates analysis, is another useful technique to reduce the dimensionality of data and visualize patterns. This research used the *cmdscale* function in the *stats* package (R Core Team, 2021) to reduce the multi-dimensional set of dissimilarities between participants to a 2-dimensional set of points where the distances between the points were approximately equal to the dissimilarities. The cMDS reduced dissimilarities were input into Partitioning

Around Medoids clustering algorithm (see section 3.1.4) to visualize the clusters on a scatter plot and to compare to the other clustering outputs.

3.1.2. Determine Important Factors

As mentioned above, a supervised approach that uses the labeled switch rate for each time a participant completes the overall task was used to identify factors that were important to the strategies based on the task-switching behavior. While there are many possible supervised techniques available, this research used a random forest (RF). Some advantages of a random forest were that it can take both categorical and numerical inputs, it was robust to missing data, and it can handle outliers. Another advantage of a random forest was that the RF predictors create a dissimilarity measure between labeled observations (e.g., subjects with known switch rates) as part of their construction. One drawback of the method was that the decision tree output was difficult to interpret, especially with a large number of independent variables, since the method was primarily a prediction algorithm. While it was possible to identify which independent variables contributed most to a subject’s switch rate, the results could not be used to identify which independent variables were common to subjects with similar switch rates.

Random forest can be used for classification or regression, where the RF predictor is an ensemble of individual decision trees. There is little or no correlation between the individual trees, which allows the forest to perform better than any individual tree. Each tree is constructed using a random subset of all the inputs, or independent variables, where the size of the subset is defined by the modeler (Breiman, 2001). Because the outcomes for the Project RED data were continuous, this research used random forest regression models, where each tree predicted an output (i.e., a switch rate for each subject) based on its randomly selected independent variables and the forest picked the average of the outputs of all trees.

This research used the *randomForest* R package (Liaw & Wiener, 2002) to run the regression random forest on the PCA reduced dataset (i.e., no personality measures) to generate predicted switch rates for in-sample and out-of-sample data, identify the most important predictors of switching, and measure the dissimilarity between observations (i.e., each completion of the overall task). The PCA reduced dataset was used to decrease the runtime as well as reduce noise and the chance of overfitting the data. Data reduction was completed on subsets of the data so the PCA reduced set data still contained a parameter for each subset (e.g., one parameter for behavioral ties instead of 12) and the difference in identifying the most important predictors should be negligible. During the construction of the RF, the training data were run down each individual tree and if two observations ended in the same terminal node, the similarity between the two observations increased by one. At the end of construction, the similarity between an observation and itself was set to one and the similarities between all observations were made symmetric and divided by the total number of trees, resulting in a symmetric, positive definite matrix with values between [0, 1]. The RF dissimilarity is $\sqrt{1 - SIM_{i,j}}$ and was input into classical multidimensional scaling and clustering algorithms (Shi & Horvath, 2006) to measure the accuracy of out-of-sample predictions against the random assignment model. The clustering algorithms are described in section 3.1.4.

3.1.3. Dissimilarity Measure

Many clustering algorithms require a distance measure to cluster on, which the RF dissimilarity matrix provided a measure of how far apart two observations were from one another and was used to identify clusters of subjects with similar patterns of task-switching behavior. The RF dissimilarity is a distance measure created as part of the regression algorithm, described above, that can also be generated from an unsupervised version of the RF algorithm using unlabeled data. For this research, the unsupervised random forest (uRF) dissimilarity matrix was constructed using the *UnsupRF* R package (Ngufor, 2019) which allowed the number of forests and trees used to be defined by the modeler. This research used either 50 or 100 forests in the models and between 2000 and 4000 trees for each model. Using more trees and forests improved the uRF outputs, but led to longer runtimes. Each tree was grown using a randomly selected subset that contained about one third of the variables.

The unsupervised random forest, as implemented in *UnsupRF*, used the unlabeled observed data along with additional synthetic data that were generated by taking a random sample from each variable of the observed data, either with (i.e., *empirical*) or without (i.e., *permute*) replacement. Both data were labeled with an artificial class as either class 1 (observed) or class 2 (synthetic). The uRF predictor was constructed as a classifier to differentiate the observed from the synthetic data and created a dissimilarity matrix for the unlabeled observations (Shi & Horvath, 2006). This research generated the synthetic data using *empirical* sampling, which created synthetic data by randomly sampling from the empirical marginal distributions of the variables.

Unsupervised RF dissimilarity has been applied successfully in genetics research as a distance measure for clustering in several applications (Shi & Horvath, 2006) where the resulting clusters are interpretable, providing support to using the method in this research. In this research, the unsupervised RF method generated a dissimilarity measure between participants to use in classical multi-dimensional scaling and several different clustering algorithms, as described in section 3.1.4. The clustering algorithms identify patterns that are undefined by the output variable.

3.1.4. Clustering algorithms

Because the Project RED data did not include any measure of decision strategy, an unsupervised learning method was needed to determine the participants' strategies. Cluster analysis aims to divide a set of objects into two or more clusters such that similar objects are in the same cluster and dissimilar objects are in different clusters. The uRF or RF dissimilarity matrix provided a distance measure between each of the participants to input into two different clustering algorithms: Partitioning Around Medoids (PAM), also referred to as k-medoids (Kaufman and Rousseeuw, 2005), and hierarchical. The results from these algorithms were assessed using metrics described in section 3.1.5 and compared to identify the algorithm that produced the best in-sample predictions. The clusters were also used to predict out-of-sample outputs for the test dataset by using the medoid of the cluster for all the clustering methods. The medoid was calculated using the *mediod* function in the *UnsupRF* R package (Ngufor, 2019).

The k-medoids clustering method is similar to k-means; k-medoids clusters to the nearest medoid while k-means clusters to the nearest centroid, or mean. Both k-means and k-medoids partition the dataset into groups and assign points to a cluster by minimizing the distance between that point and the center of that cluster. Unlike k-means clustering, k-medoids chooses actual data points (medoids) and not a representation of the data (means) as centers. This allowed for greater interpretability of the cluster centers than using k-means. An additional benefit was that k-medoids can be used with arbitrary dissimilarity measures, while k-means generally requires Euclidean distance, which allowed the use of the RF or uRF dissimilarity measures as an input. K-medoids is also more robust to noise and outliers than k-means.

The medoid of a cluster is the object whose average dissimilarity to all other objects in the cluster is minimal; it is the most centrally located point in the cluster. Partitioning Around Medoids (PAM) attempts to minimize the total distance (*D*) between objects within each cluster. The number (*k*) of medoids found is equal to the number of clusters desired. Once the medoids are found, the data are classified into the cluster of the nearest medoid. The algorithm has a build and a swap phase. The build phase finds a representative set of *k* objects. The first object selected has the shortest distance to all other objects; it is in the center. An additional *k-1* objects are selected one at a time to decrease *D* as much as possible at each iteration. The swap phase considers possible alternatives to the build-phase *k* objects in an iterative manner. The algorithm searches the unselected objects for the one that will lower the objective function the most if is exchanged with one of the previously selected *k* objects. The swap is made and the algorithm continues to iterate until there are no exchanges found that will lower the objective function. This research used the *pam* function in the *cluster* R package (Maechler et al., 2021) for the PAM clustering results. PAM clustering was run on the cMDS reduced uRF dissimilarities as well as the raw uRF or RF dissimilarity measures.

Hierarchical clustering creates clusters in a hierarchical tree-like structure and can be implemented as an agglomerative or divisive algorithm (Kaufman and Rousseeuw, 2005). Agglomerative is a ‘bottom up’ approach where all datapoints are isolated as separate groupings initially and then iteratively merge together on the basis of similarity until there is one cluster. There are multiple ways to determine the similarity between groupings. Divisive clustering is a ‘top down’ approach where a single initial cluster is divided based on differences between the datapoints, and is not commonly used. This research used the *hclust* function in the *stats* R package (R Core Team, 2021) which implements the agglomerative approach. At each stage, the function recomputed the distances between clusters by the Lance-Williams dissimilarity update formula according to the particular clustering method being used. The *hclust* function included multiple clustering methods (i.e., ways to determine the similarity between the groups): two methods of Ward clustering, single linkage, complete linkage, average, median, centroid, and mcquitty. This research tried each method and used the one that gave the lowest error rates and highest Adjusted Rand Index.

3.1.5. Model Comparison

Because the participants’ decision strategies for completing the overall task were unknown, the ‘true’ clustering was based on the assumption that the decision-making strategy used by the individual to solve the overall task related to their task-switching behavior, as observed by their task-switching rate. The ‘true’ clustering was determined by breaking the data into *k+1* quantiles and assigning them to *k* clusters of monotonically increasing switch rates. This was a simplification of the strategies used and likely biased the algorithm to favor models containing factors that were predictors of switch rate over other models. Another possible option was to not compare the predicted clusters to any true value since the true value was unknown and instead compare the predicted cluster to a randomly generated cluster value. This would eliminate bias from using another factor (e.g., switch rate) to generate an artificial true value. Based on the assumption that decision-making strategy was related to task-switching behavior, this research used the artificially generated true value to evaluate out-of-sample prediction accuracy to test the first hypothesis of this research.

Three different metrics were used to compare the results from the clustering of participants using multiple different machine learning algorithms. A visual assessment was used, where the assigned cluster for each participant using each of the clustering methods was plotted, using a box plot, against the switch rate and visually compared to a plot of the 'true' clustering. The other two metrics, Adjusted Rand Index (ARI) and classification error, are common quantitative measures of the difference between two clusters. The ARI compares two vectors of class labels (i.e., the predicted and true clusters) and has a range of values from 0 to 1, where zero is the expected value for random clustering and 1 is the expected value for perfect agreement between the two vectors. It is a widely used metric for validating cluster performance. This research used the *adjustedRandIndex* function in the *mclust* R package (Scrucca et al., 2016) to calculate the ARI. The classification error is the error rate between the cluster labels predicted by the algorithm and the true clustering. This research used the *classError* function in the *mclust* R package (Scrucca et al., 2016) to calculate classification errors.

3.2. Cognitive mechanisms

Cognitive process models, specifically multi-attribute linear ballistic accumulator (MLBA) models, a type of evidence accumulation model, were used to address the research questions about how a participant decides to select a new task or remain at their current task. Four versions of the model were compared to test whether participants considered each attribute of the tasks or the tasks as a whole as well as whether they considered all options simultaneously, in a single-stage, or used a down-select process by considering first the task type (i.e., individual, team, or MTS) and then the individual task from within that type as a two-stage process. One benefit of evidence accumulation models was that they contain psychologically relevant parameters and provided insight into explaining observed behaviors. A challenge of using evidence accumulation models with the Project RED data was the small number of trials (i.e., switches) for each participant. This was mitigated by creating a separate simulated task-switching dataset, with a large number of trials for each participant, to use for comparing the fit metrics for each of the models. Only parameter estimates from the real data were used in the counterfactual prediction models.

While several types of multi-alternative, multi-attribute choice models (e.g., MDFT, MLBA, and MLCA) produce acceptably good fits of preferential choice and perceptual stimuli, especially when using a hierarchical version to account for individual variability, with some variation in the results depending on model assumptions and data structure (Turner, et al., 2018), this research leveraged and expanded the MLBA model, which is more computationally tractable than other multi-alternative, multi-attribute models, to develop a cognitive model to describe, explain, and predict individual differences in task-switching performance. The large number of response options (i.e., 15 possible tasks to select) in the Project RED data led to computational runtimes of several hours to over a day for each time parameters were estimated, even using the less complex MLBA model with high performance computing resources (Ohio Supercomputing Center, 1987). Unlike existing versions of the MLBA model that focus on deliberate decisions (e.g., risky choice, perceptual tasks), task-switching for Project RED was not presented as a deliberate decision, but instead was part of the process to complete a larger task of trying to determine the best design and location on the Martian surface to drill a new well. From an information foraging perspective, participants may deliberately switch to a new task to explore for additional information, but they were not cued to do so.

In the linear ballistic accumulator (LBA) model (Brown & Heathcote, 2008), evidence accumulates at a constant drift for all possible responses until one accumulator reaches the response threshold, as shown in Figure 9. An accumulator for each possible response is stochastically created with a starting point in the range of uniform distribution $\mathbf{U}[0,A]$ and a mean drift rate, d_i, with normally distributed noise, s_d. The race to the threshold, b, which must be greater than A, is linear and deterministic. The model also includes non-decision time, t_0, to account for time to encode the stimulus and time to produce a response. Once an accumulator reaches the threshold evidence for the alternative response(s) is discarded.

Conceptually, the starting point range, A, is interpreted as variability in the initial evidence across the trials, representing bias in the response; the response threshold, b, denotes the evidence required to make a decision, representing inhibition in responding; and mean drift rate, d_i, is the speed of information processing and represents efficiency of the response. For task-switching data, the threshold parameter measured inhibition of leaving the current ongoing task (i.e., task stickiness), the starting point parameter measured bias towards selecting any alternative task (i.e., the switch-avoidance tendency), the drift rates measured the attractiveness of each alternative task, and the non-decision time measured time spent not considering an alternative task (e.g., engaging in the current task, encoding, and processing the next task choice). The drift rate measured the efficiency of information processing and the parameters within the drift rate equation specified how much the attention weight and the task value contribute to that efficiency.

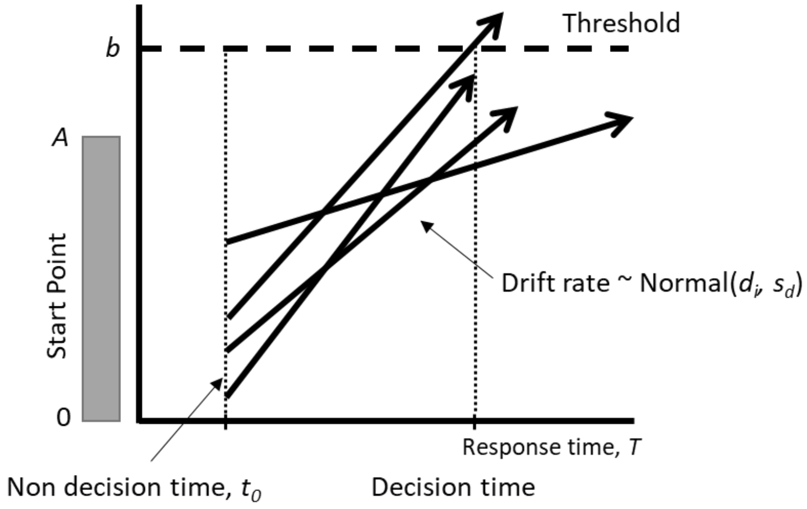


Figure 9. Linear Ballistic Accumulator model. The x-axis is time and the y-axis is the amount of evidence. Each arrow is a different possible response option.

The MLBA model (Trueblood et al., 2014) consists of a front-end pre-processing stage that determines individual pre-decision preferences (i.e., drift rates) and a back-end selection process to account for random variation of responses. Once the drift rate is determined the selection process proceeds using the LBA model (Brown & Heathcote, 2008). The MLBA model defines the drift rate for each option in terms of valuation, V_{ij}, that represent a comparison between alternatives *i* and *j* of the subjective valuation of attributes across alternatives; valuation is determined by the pairwise difference of subjective values, u_{ik} - u_{jk}, for the k^{th} attribute is multiplied by the weight of attention, a_k, given to a comparison. The weight, or amount, of attention given to a particular comparison depends on the similarity of the attribute values for each option. Cohen et al’s (2017) adaption of the MLBA model, for risky choices and perceptual stimuli with more than two attributes, defines the subjective value and weighting function using cumulative prospect theory (Tversky & Kahneman, 1992).

Four versions of the MLBA model were used to address how a person decides to remain at a task or switch to a new task and test the associated hypotheses, and Table 4 shows a breakdown of which cognitive process mechanism each of the 4 models support related to the research questions. The four versions of the model used a drift rate equation based either on cumulative prospect theory (CPT) or information foraging (IF) theory and assumed that participants either considered all options simultaneously, in a single-stage, or down-selected by considering first the task type (i.e., individual, team, or MTS) and then the individual task from within that type as a two-stage process. Cohen’s (2017) CPT version of MLBA was applied to task-switching data by assuming that less task switching corresponds to being risk averse and more task switching corresponds to being risk seeking. This research also developed an alternate MLBA version adapted to use information foraging theory to determine the drift rates.

Table 4. Summary of cognitive process mechanism that each model version supports

	Attribute-level (Q2)	Alternative-level (Q2)
All options simultaneously (Q3)	Single-stage CPT	Single-stage IF
Down-select options (Q3)	Two-stage CPT	Two-stage IF

The second research question examined whether the preference to select a particular task was based on considering each individual attribute of that task separately or if the attributes were considered as a whole, as a sum of the individual attributes. The associated hypothesis favored the model where tasks were considered as a whole and was tested by comparing two different multi-attribute LBA models; both assumed that all the response options were considered simultaneously (i.e., single-stage). The first was taken from a previous study of risky choice and perceptual stimuli (Cohen et al., 2017) and the second was developed as part of this dissertation research. The first model defined the attention weight and the subjective value for each attribute within the drift rate (i.e., the rate of accumulation of evidence) in terms of the principles from cumulative prospect theory (Tversky & Kahneman, 1992). The second model defined the attention weight and subjective value within the drift rate at the task level, while still considering the value of each attribute as a contribution to the value of the whole task, in terms of the principles from information foraging theory (Pirolli & Card, 1999)

The third research question examined whether all tasks were evaluated simultaneously or if a serial process was used to create a subset of tasks to consider when selecting the next task. The associated hypothesis favored the two-stage serial model and was tested by comparing models with either a single-stage (i.e., all tasks evaluated simultaneously) or two-stages (i.e., serial down-select process used). Two additional MLBA models, both structured so that the decision-making process occurred in two stages, were developed to address this question, where one used the drift rate equation based on cumulative prospect theory and the other used the drift rate equation based on information foraging theory. These models assumed that the first stage decision was based on a heuristic to reduce the number of response options for the second stage. The heuristic modeled in this research was whether the final decision should be an individual, team, or multi-team decision, which limits the number of possible second stage responses. Other possible heuristics were considered, such as task typicality or popularity, but were not included in this research due to logistical constraints. In the second stage the subject then decided the specific task to complete from the reduced list of possibilities. In addition to the threshold, starting point parameter and non-decision time, the two-stage model included two drift rates, one for each stage, and a stage delay to account for the time spent in stage one. The stage delay was the inferred stage one response time.

3.2.1. Attribute-level MLBA Model

The cumulative prospect theory (CPT) version of the MLBA model (Cohen et al., 2017), given by equations 1, 2, and 3, defined the drift rate in terms of the attention weight given to each attribute for each task and the subjective value of each attribute for each task. The attention weight was defined as $a_k=\pi(w_k)$. The weighting function, π, was taken from cumulative prospect theory (Tversky & Kahneman, 1992) and w_k was the importance weight given to the k^{th} attribute. The subjective value (u_{ik}) was defined as a power function of V_{ik}, where V_{ik} was the objective value of the k^{th} attribute of the i^{th} alternative. The drift rate for each option i^{th} response option (equation 1) was the sum of the initial drift rate (I_0), the relative comparison between option i and every other j option, and the absolute value of the attributes of the i^{th} option. The initial drift rate was set to zero for the task-switching data. The second term in equation 1 multiplied the pairwise difference in subjective value, u_{ik} - u_{jk}, by the attention weight for each kth attribute. The sum of the differences was multiplied by a scaling factor, c_d, and divided by a normalization term. The scaling factor, c_d, scaled the extent that differences in subjective valuation affect the drift rate (Cohen et al., 2017; Trueblood & Dasari, 2017). The final term in equation 1 found the absolute value of option i as the maximum product of attribute weight and subjective value, multiplied by a scaling factor, c_m, and divided by a normalization term. The term increased the drift rate of an option with a high weight or subjective value, making it more likely to be selected, and was included because it was shown in Cohen et al. (2017) to improve model fits that include response time data. The scaling factor, c_m, scaled the extent that absolute value of the option affects the drift rate (Cohen et al., 2017; Trueblood & Dasari, 2017). There was a relationship between c_m and response time where response times decrease as c_m scaling factor increases, shown in Figure 10.

$$d_i^{(CPT)} = I_0 + \frac{c_d \sum_k \sum_j a_k (u_{ik} - u_{jk})}{\sum_k a_k \sum_m u_m} + \frac{c_m \max_k (a_k u_{ik})}{\sum_k a_k} \quad (1)$$

$$a_k = \pi(w_k) = \frac{w_k^\gamma}{w_k^\gamma + (1 - w_k)^\gamma} \quad (2)$$

$$u_{ik} = V_{ik}^\alpha \quad (3)$$

The CPT equation was applied to task-switching data by assuming that less task switching corresponded to being risk averse and more task switching corresponded to being risk seeking. Based on results in Cohen et al. (2017), a participant that switched tasks less should have a lower γ value while a participant that switched tasks more should have a higher value. The CPT version considered both the value and attention weight of each attribute for every response option as part of the drift rate equation. The attention weight described the importance of the attributes and the subjective value of an attribute was a power function of its objective value (Cohen et al., 2017). For this research, all 4 task attributes were weighted equally (i.e., w_k=0.25 for each attribute) as there was no evidence available to determine that any attribute was or should be weighted higher than the others by the participants. This research used the objective attribute values provided with original data to determine the subjective value of each attribute; these objective values were determined by the original researchers (Mesmer-Magnus, 2020).

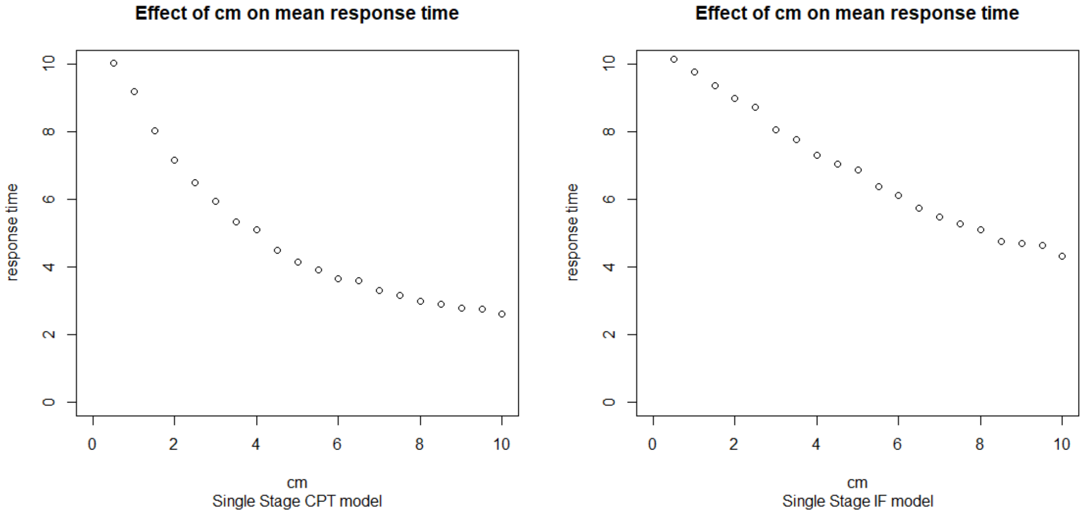


Figure 10. Relationship between the absolute scaling factor, c_m, and response time

3.2.2. Alternative-level MLBA Model

The IF version of the MLBA model (Mahoney et al., 2021) defined the drift rate in terms of the subjective value of each task as a whole, using the sum of the individual attribute values, and the attention weight given to each task. Initial IF model drift rate equations are given in equations 4, 5, and 6.

$$d_i^{(IF)} = I_0 + \frac{c_d a_i \sum_j (u_i - u_j)}{\sum_m u_m} \tag{4}$$

$$a_i = \frac{\pi_i}{\sum_j \pi_j} = \frac{\pi_i}{1-\pi_i}, 0<\pi_i<1 \tag{5}$$

$$u_i = V_i^{\alpha} \tag{6}$$

Equation 5, defining the attention weight, was replaced by equation 8 for this research. This modification of the attention weight equation added the β parameter to reduce the number of model parameters and defined the contribution of attention weight to the drift rate as a power function of Luce’s choice rule (Luce, 1977), eliminating the constraint that π must be between zero and one. The attention weight was the weight of the current option, *i*, divided by the sum of the weights of all the other, *j*, options. The weighting parameter, π, was taken from information foraging theory (Pirolli & Card, 1999) and was defined as the profitability of the response option. The profitability is the gain divided by the processing time for each option. The attention weight of option *i* depended not only on its profitability but also on the profitabilities of all the other *j* options. An initial attempt was made to include π_i as free parameters in the model, thus allowing the data to determine the value of each π parameter, but this produced a non-identifiable model, as shown in section 5.1.2. Instead, each π was calculated from the data being fit by the model (either real or simulated) by setting the gain of a response option to the number of times that option was selected for all participants and setting the processing time of a response option to the mean time on task across all participants for that task. The subjective value (u_i) was defined as a power function of V_i, where V_i was the sum of the individual attribute values of the i^{th} alternative. The attribute objective values provided with original data were used to determine the subjective value of each response option.

The IF model considered both the value and attention weight of the task as a whole for every response option as part of the drift rate equation. The drift rate equation, equation 7, took the same form as the CPT version of the model where the i^{th} response option was the sum of the initial drift rate (I_0), the relative comparison between option *i* and every other *j* option, and the absolute value of the i^{th} option. Again, the initial drift rate was set to zero for the task-switching data. The second term in equation 7 multiplied the sum of the pairwise differences in subjective value, u_i - u_j, by the attention weight for each *i*th option and a scaling factor, c_d, and divided this quantity by a normalization term. The final term in equation 7 found the absolute value of option *i* as the product of attribute weight and subjective value, multiplied by a scaling factor, c_m, and divided by a normalization term. The term increased the drift rate of an option with a high weight or subjective value, making it more likely to be selected, and was included because it was shown in Cohen et al. (2017) to improve model fits that include response time data.

$$d_i^{(IF)} = I_0 + \frac{c_d a_i \sum_j (u_i - u_j)}{\sum_m u_m} + \frac{c_m a_i u_i}{\sum_m a_m} \tag{7}$$

$$a_i = \left(\frac{\pi_i}{\sum_j \pi_j}\right)^{\beta} \tag{8}$$

$$u_i = V_i^{\alpha} \tag{6}$$

3.2.3. Single-Stage MLBA Model

The single-stage model used the structure of the traditional MLBA model, described in section 3.2. Five parameters from each version (i.e., CPT and IF) of the single-stage MLBA model were assumed to be related to task-switching behavior and were allowed to vary when fitting the models to both simulated and real data. For both versions, the threshold (*b*) and starting point parameter (*A*) measured a participant's aversion to switching; a higher sum of the starting point parameter and threshold indicated that the person required more evidence to leave the current task and may lead to a lower switch rate. A higher starting point parameter, regardless of the threshold value, also indicated a bias or a preference to remain at the ongoing task. Also, for both versions, the non-decision time (t_0) measured the time that the participant was engaged in non-decision behavior. The data showed that mean time on task had an inverse relationship to task switch rate (Figure 7) so as non-decision time increased the switch rate should decrease. For the CPT version of the single-stage model, the α and γ parameters in the drift rate equation related to different aspects of information processing efficiency. The γ parameter related to the weight of attention given to each task attribute and the α parameter defined the contribution of the value of each task attribute to the drift rate. In combination, both parameters affect the drift rate for each task. A higher drift rate indicated more efficient information processing for a particular task and may lead to higher switch rates, especially if many of the frequently selected tasks have high drift rates. Analogously, for the IF version, the β parameter defined the contribution of the profitability of a task to the drift rate while the α parameter did the same for the value of each task. Again, together both parameters affected the drift rate, which may impact the switch rates.

As shown in section 3.2.1, the absolute scaling factor, c_m, in the drift rate equation was inversely related to response time. The c_m was assumed to be a small value since response times are relatively long. It was set as a constant value for estimating the parameters of the simulated data, but was initially allowed to vary for both versions of the single-stage model, along with the relative scaling factor, c_d, and the other five parameters, when estimating parameters for the real data. Doing this caused problems with the parameter estimation, creating non-varying log likelihood values, so a subsequent iteration of the models assumed a constant c_d and c_m using the group level parameter estimate for each from the initial run. This issue could be caused by an error in the likelihood function, but an investigation of the code did not determine the issue. The cause of the problem is currently unknown. However, using constant values for c_d and c_m, as performed in the model investigation and parameter recovery, produced reliable parameter estimates.

3.2.4. Two-Stage MLBA Model

Two-stage versions of the MLBA models were developed to test whether participants used a serial process to create a subset of tasks to consider when selecting the next task. These versions of the model were compared to the single-stage versions to provide evidence for or against the hypothesis that a two-stage model was used. This research explored two different structures of the two-stage MLBA model since a two-stage structure of the MLBA did not exist already: an initial structure with nine parameters and a revised structure with eight parameters.

The initial structure for the two-stage model included a single non-decision time, a single starting point maximum value, two threshold values, two drift rates, and a stage delay parameter, for a total of nine parameters. The weight of attention (γ or β) and power of the objective value (α) parameters within each of the drift rate equations was allowed to vary; the scaling factors, c_d and c_m, are not. The stage delay parameter was assumed to be the same as the response time for stage one, as there was no measured value of the time spent in stage one. The initial two-stage model intended to use the threshold for the first stage as the exact starting point of the second stage, as shown in Figure 11a, but the initial implementation resulted in setting the threshold from the first stage as the maximum value of the starting point of the second stage, as shown in Figure 11b. This led to unintended additional variation in the model; the model recovered the choice distribution, but the posteriors predicted longer response times and the model was nonidentifiable (i.e., did not recover parameter values used to generate data). When generating the posterior predictive data, the starting point for stage two was a random point between zero and the threshold of the first stage, but the drift rates for each stage did not account for this since they were based on the observed data and actual response times, so the amount of time to reach the stage two threshold was longer than it should be. The model, as coded, uses the analytic solution of the LBA probability distribution function, as implemented in the *dLBA* function in the *rtdists* package (Singmann et al., 2020) to generate samples and likelihood values for each stage separately.

To implement the intended structure, where the stage one threshold is the exact starting point of the second stage, requires using approximation methods (e.g., Probability Density Approximation (PDA; Holmes, 2015)) to estimate the likelihood values. This research attempted to implement the PDA to estimate the likelihood values for use by the Particle Metropolis within Gibbs (PMwG) sampling methods to generate posterior parameter estimates, but the attempts did not result in a usable likelihood function. PDA simulates a model thousands of times and uses kernel density estimation to produce a synthetic likelihood function to use for Bayesian parameter estimation. Holmes (2015) used PDA combined with Differential Evolution Markov Chain Monte Carlo (DE-MCMC; Turner et al., 2013) to estimate parameters for a piecewise LBA; however, the PMwG method also uses thousands of samples for each iteration and attempting to combine PDA and PMwG resulted in a level of complexity in the parameter estimation that did not efficiently compute the parameter estimates. Additional work is needed to refactor the code to implement the structure in Figure 11b and is proposed as future work.

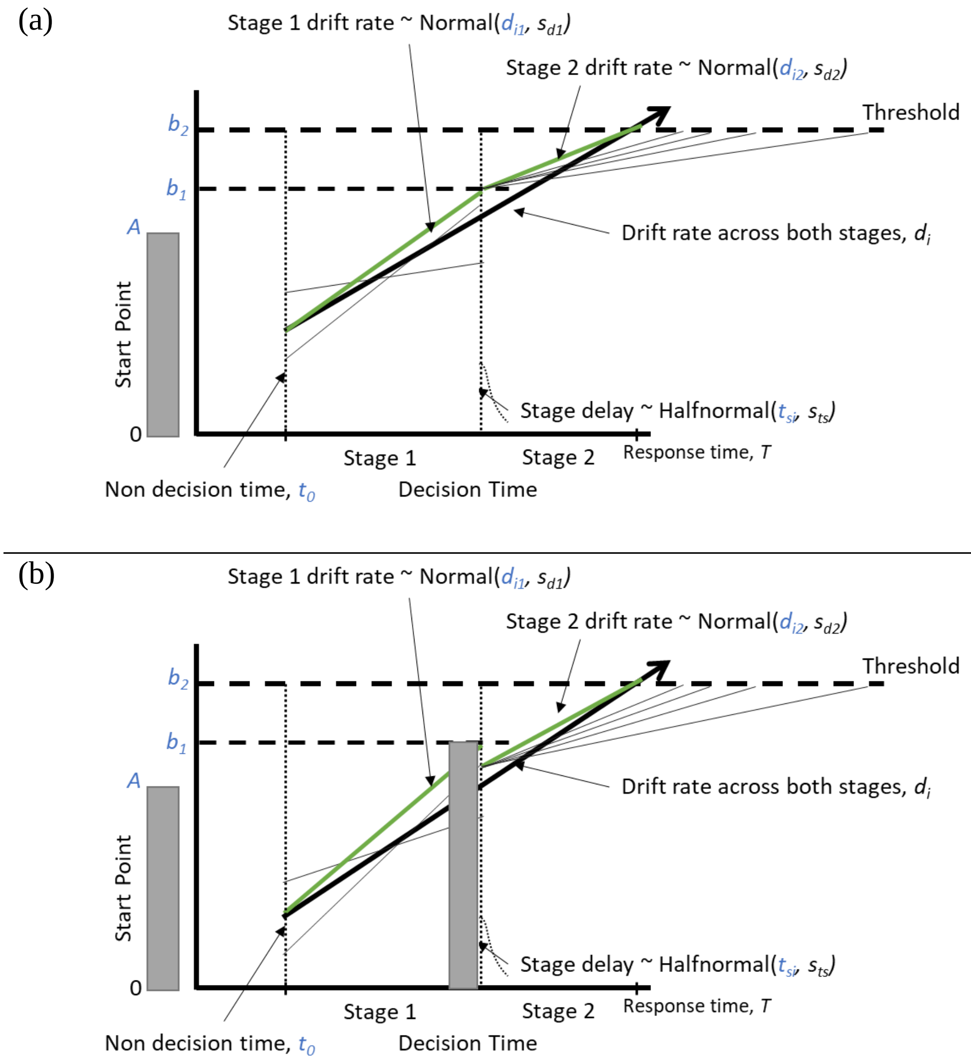

(a)
Stage 1 drift rate ~ Normal(d_{i1}, s_{d1})
Stage 2 drift rate ~ Normal(d_{i2}, s_{d2})
Threshold
b_2
b_1
A
Start Point
0
Drift rate across both stages, d_i
Stage delay ~ Halfnormal(t_{si}, s_{ts})
Stage 1
Stage 2
Response time, T
Non decision time, t_0
Decision Time
(b)
Stage 1 drift rate ~ Normal(d_{i1}, s_{d1})
Stage 2 drift rate ~ Normal(d_{i2}, s_{d2})
Threshold
b_2
b_1
A
Start Point
0
Drift rate across both stages, d_i
Stage delay ~ Halfnormal(t_{si}, s_{ts})
Stage 1
Stage 2
Response time, T
Non decision time, t_0
Decision Time

Figure 11. (a) Intended structure of the initial implementation of the two-stage model structure (b) Actual structure of the initial implementation of the two-stage model

The revised implementation of the two-stage model, shown in Figure 12, instead used a single threshold value for both stages. All other parameters were the same as the initial implementation, resulting in a total of eight model parameters. The drift rate function included the first and second stage drift rate as well as the stage delay parameter so the only response time needed was the measured value, *T*. This allowed the model to generate a single likelihood value for both stages together, using the *dLBA* function. The drift rate equation was obtained by using the relationships between LBA parameters. The general slope of the drift rate equation was $\frac{\Delta x}{\Delta y}$, where Δx was *T*-t_0 and Δy was the *b*-*A*, or $d = \frac{b-A}{T-t_0}$. This can be written out for each stage as

$$d_1 = \frac{b_1 - A}{t_s - t_0} \tag{9}$$

$$d_2 = \frac{b - b_1}{T - t_s} \Leftrightarrow T = \frac{b - b_1}{d_2} + t_s \tag{10}$$

where b_1 is the first stage threshold value.

Solving Equation 9 for b_1 gave $b_1 = d_1(t_s - t_0) + A$. This was substituted into Equation 10 resulting and Equation 10 was solved for *b*.

$$b = d_2(T - t_s) + \ b_1 = d_2(T - t_s) + \ d_1(t_s - t_0) \tag{11}$$

Equation 11 was substituted into the drift rate equation for *b*, Equation 10 was substituted for *T*, and *A* and t_0 were set to zero to simplify the calculations. The drift rate across both stages is then given by Equation 12.

$$d = \frac{d_2 * b}{b - t_s(d_2 - d_1)} \tag{12}$$

This implementation recovered both the choice and response time distributions, but only recovered some of the parameter values when all eight parameters are allowed to vary. The model became identifiable by reducing the number of parameters that vary to only the threshold (*b*), maximum value of the range of starting points (*A*), and non-decision time (t_0). Allowing more parameters to vary resulted in model parameter estimates that do not match the original values, but that still recovered the original choice and response time distributions.

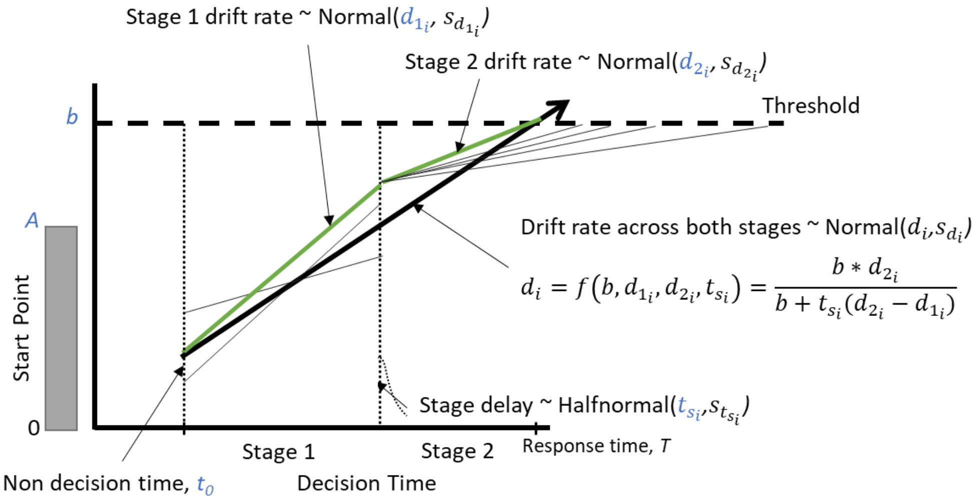

Stage 1 drift rate ~ Normal(d_{1_i}, $s_{d_{1_i}}$)
Stage 2 drift rate ~ Normal(d_{2_i}, $s_{d_{2_i}}$)
Threshold
b
A
Start Point
0
Drift rate across both stages ~ Normal(d_i,s_{d_i})
$d_i = f\left(b, d_{1_i}, d_{2_i}, t_{s_i}\right) = \frac{b * d_{2_i}}{b + t_{s_i}(d_{2_i} - d_{1_i})}$
Stage delay ~ Halfnormal(t_{s_i},$s_{t_{s_i}}$)
Stage 1
Stage 2
Response time, T
Non decision time, t_0
Decision Time

Figure 12. Structure of the revised implementation of the two-stage model

3.2.5. Model fitting and parameter estimation

The MLBA models are typically applied to structured data collected in a laboratory with hundreds or thousands of trials per subjects whereas the Project RED data used in this research was more complex, noisy, and sparse. A separate simulated dataset was generated with a larger number of trials per subject to produce parameter estimates with lower error to only use in model comparison and testing the hypotheses associated with explaining how a participant decide to switch tasks. This research implemented two different methods to estimate parameter values: maximum likelihood estimation (MLE; Farrell & Lewandowsky, 2018) and particle Metropolis within Gibbs (PMwG) sampling (Gunawan et al., 2020), a hierarchical Bayesian parameter estimation method. MLE found a single parameter value for each participant for which the observed data was most likely. This research adapted Steve Fleming's Matlab code for fitting the LBA model (available at https://github.com/smfleming/LBA), implemented to minimize the negative log likelihood, to iterate over many possible parameter values, and selected the best fit values using the *fmincon* function in Matlab. The PMwG method included random effects for subjects and gave a distribution of values for each parameter at the participant and group levels. The Project RED data had small sample sizes so only the hierarchical Bayesian method was applied to this data. Additionally, only PMwG sampling was used to estimate parameters for the two-stage models, using any of the datasets, while both methods were used for the single-stage models.

The PMwG method was selected for this research because it more efficiently generates posterior samples for models with highly correlated parameters, like the LBA model, and the samples have lower autocorrelation, than other MCMC samplers (e.g., DE-MCMC). The method also accounts for non-independence of random effects, allowing individual level parameters to be correlated in the prior, by reparameterizing them and estimating the covariance structure between parameters in a principled manner (Gunawan et al., 2020).

PMwG uses Gibbs sampling assuming a multivariate normal distribution for group-level parameters and a particle MCMC approach (Gunawan et al., 2017) to sample random effects for the subject-level parameters. The sampler starts with an initial set of parameters (θ) and random effects (α), provided by the modeler. For each iteration, the PMwG algorithm samples the group-level parameters of the MLBA model using Gibbs steps conditional on the particle (i.e., vector of random effects) from the previous iteration. A large number of new particles are generated from the current particle using the conditional MC algorithm (Gunawan et al., 2020). The particle from the previous iteration is compared to the new particles (i.e., the proposals) and PMwG selects whichever (i.e., previous or newly generated particle) best matches the data and prior as the new particle for the next iteration, by maximizing the likelihood (i.e., minimizing the negative log likelihood) that the estimated parameter values generate the data given the prior values. This continues for the required number of iterations. The conditional MC algorithm is easily parallelized, which increases computational speed, an additional benefit of the PMwG approach. While there wasn't a need for near-real-time analysis, the large number of response options in the Project RED data led to runtimes of several hours to over a day for each parameter estimation run, using supercomputing resources.

The sampler was applied in three stages: burn-in, adaptation, and sampling. The burn-in stage allowed the Markov chain to move from its initial randomly drawn value to a stable range of posterior values; burn-in samples were discarded prior to determining the estimated parameter values. Only samples from the sampling stage were used to determine estimated parameter values. The burn-in particles for each subject were sampled from a mixture of the group-level distribution and a multivariate normal distribution centered on the current particle, with a variance that is smaller than the group-level distribution. The group-level distribution provided a safety net for situations where the particles generated from the subject's random effects vector were unusual or unlikely. This led to the group-level proposal being chosen instead of the sampler taking a long time to generate a sensible vector of random effects, leading to a faster sampling time. The adaptation stage continued using the sampling algorithm from the burn-in stage until obtaining a minimum of 20 unique samples from each subject's posterior distribution. These samples provided a reasonable idea of the posterior distribution for each subject's random effects vector and were used to build an adaptive proposal distribution that makes very efficient proposals in the sampling stage. The adaptive proposal distribution was a multivariate normal distribution summarizes the unique samples in the adaptation stage and was used to generate sampling stage proposals. This distribution, for each subject, summarized both the posterior distribution of their random effects as well as the way these random effects related to the group-level parameter. This was important because it allowed the sampler to draw conditional proposals. Conditional proposals were consistent with both that subject's random effects and with the current proposal for the group-level distribution leading to generated proposals being frequently accepted, so fewer new particles were needed in the sampling stage. The adaptive proposal distribution was updated throughout the sampling stage leading to a more accurate proposal distribution. The sampling stage also included a few proposal particles from the burn-in algorithm to protect against very poor conditional proposal distributions (Newcastle Cognition Lab, 2021).

The PMwG sampling method was implemented in this research using the *pmwg* R package (Cooper et al., 2021) with a customized likelihood function for each version of the MLBA model. The likelihood function relied on *rtdists* package (Singmann et al., 2020) to analytically solve the LBA model's PDF. The functions in the *pmwg* R package allowed the modeler to set for each stage the number of iterations, the number of particles, and the width of the proposal distribution (ε). More particles were needed for more complex models to give the sampler a greater chance of accepting a new particle for each iteration. Narrower proposal distributions (i.e., smaller ε) led to higher acceptance of new particles because more new particles were closer to the current particles, but this also led to slower convergence of the posterior. The number of iterations was tied to the number of particles and value of ε. If the number of particles was lower and ε was small, more iterations helped ensure the sampler reached the posterior space, but increasing the number of particles also led to the same result. Increasing particles and iterations increased the computational run time, with increased particles increasing the run time more because they were evaluated for each subject. For this research, the number of iterations for the burn-in stage varied between 500 and 2500 and for the sampling stage ranged from 1000 to 4000. ε ranged from 0.3 to 0.7. The number of particles was mostly 1000 for the burn-in stage and 100 for the adaptation and sampling stages. The values were adjusted as needed after evaluating of the outputs of each stage, to select parameters that led to stable posterior estimates.

The sampler provided three types of samples from the posterior distribution of the model: the means for the group level parameters (θ); the vectors of random effects for each subject (individual level parameter values, α); and the group-level variance covariance matrix (Σ). This research used the resulting individual level parameter estimates for all model parameters except the drift rate scaling factors, which used the group-level parameter estimates as described in section 3.2.3.

3.2.6. Simulated Data

As mentioned previously, the provided Project RED task-switching data had a small number of trials for many of the participants. The number of trials ranged from a minimum of 5 to a maximum of 85. Even the maximum number of trials in this data was near or below the minimum of trials that is generally used to estimate model parameters for the MLBA models. For this reason, a separate simulated task-switching dataset was created to compare the fit metrics for each of the models.

Initially, a dataset was created to simulate data from 20 independent subjects with 5 available response options. Responses were not allowed to repeat the previous response. The threshold (b), starting point parameter (A), and non-decision time (t_0) were constant for all participants and the drift rates for each subject were randomly selected from a uniform distribution with a minimum value of zero and a maximum value defined in Table 5. The number of trials for each subject were randomly selected from a uniform distribution between 975 and 1025. Some limitations of this dataset were that only 5 response options were included and that only the drift rates varied for each subject; the threshold, starting point parameter, and non-decision time did not. Also, the number of trials did not vary much across subjects so there was little variation in switch rates between subjects. Figure 75 in Appendix B shows an example of the choice and response time (RT) distributions from this simulated dataset. MLE fitting of the single-stage models was used to find the best-fit parameters of this data, but the results were not used to test the hypotheses due to the limitations in the data.

Table 5. Parameter values used to create simple simulated dataset

Parameter	Value
Threshold, b	5
Starting point parameter, A	15
Option 1 drift rate mean, v_1	0.2
Option 2 drift rate mean, v_2	1.2
Option 3 drift rate mean, v_3	0.05
Option 4 drift rate mean, v_4	0.6
Option 5 drift rate mean, v_5	0.05
Non-decision time, t_0	10

The simulated dataset used to test the hypotheses, in addition to the Project RED data, included 15 task choices for 48 subjects and was generated using the simplest LBA model, where each subject was independent of the others, and the threshold, starting point, drift rates and non-decision time varied for each subject. The number of subjects was set to 48 because of computational runtime considerations of the subsequent parameter estimations. The responses were not allowed to repeat the previous response; the response time of any repeat responses was added to the previous trial and the repeat trial was removed from the dataset to model task stickiness. The parameter values for each subject were randomly selected from a uniform distribution between a minimum and maximum value for that parameter, shown in Table 6. The threshold, starting point, and non-decision time values were selected based on the parameter estimates from the Project RED data. The drift rate values were selected to produce a distribution of responses with variation in responses (in terms of counts not specific to a choice option) similar to the Project RED data. Figure 76 in Appendix B contains the histogram of actual values for each parameter. The model generated 6000 trials for each subject, with response times for each trial ranging from 1 or 2 seconds to over 600 seconds. The total time each subject spent over all the trials (i.e., switching between tasks on the simulated overall task) was found by summing the response times over all the trials; the minimum total time (t=51755 sec. or 14.4 hrs.) was used as a cutoff for all other participants to complete the overall task. As a result, the number of trials for the other participants was adjusted, with a minimum of 836 trials, a maximum of 5429 trials, and a median of 1522 trials, creating a different number of task switches and a different switch rate for each participant, since all participants had the same total time. Neither the resulting response time distribution nor the choice distribution, in terms of the median and maximum response times (median RT_{real}=52 sec. vs. median RT_{sim}=17 sec.; maximum RT_{real}=1167 sec. vs. maximum RT_{sim}=608 sec.) and the particular responses selected, were sufficiently similar to the Project RED data to use the simulated data in place of the Project RED data to explain the task-switching behavior. However, a visual assessment determined there was sufficient similarity in the shape of the response time distribution and the variation in the responses that enabled the simulated data to be used to compare differences in switching behaviors using the different models and address the research questions. Figure 13 shows an example of the choice and response time (RT) distributions for 3 subjects from the simulated data. This 48-subject simulated dataset was evaluated with the PMwG parameter estimation method for the single- and two-stage models.

Table 6. Parameter values used to create 48-subject simulated dataset

Parameter	Minimum value	Maximum value
Threshold, b	0.1	10
Starting point parameter, A	100	1000
Option 1 drift rate mean, v_1	0	0.6
Option 2 drift rate mean, v_2	-1.55	-0.55
Option 3 drift rate mean, v_3	0.85	2.45
Option 4 drift rate mean, v_4	0.5	2.8
Option 5 drift rate mean, v_5	0.03	0.07
Option 6 drift rate mean, v_6	0.6	1.1
Option 7 drift rate mean, v_7	0.01	0.25
Option 8 drift rate mean, v_8	0.15	2.75
Option 9 drift rate mean, v_9	0.5	0.9
Option 10 drift rate mean, v_{10}	-0.2	1.4
Option 11 drift rate mean, v_{11}	-2	2
Option 12 drift rate mean, v_{12}	-3	-2
Option 13 drift rate mean, v_{13}	0.25	3.45
Option 14 drift rate mean, v_{14}	0.8	1.6
Option 15 drift rate mean, v_{15}	-0.1	0.1
Non-decision time, t_0	0.5	3

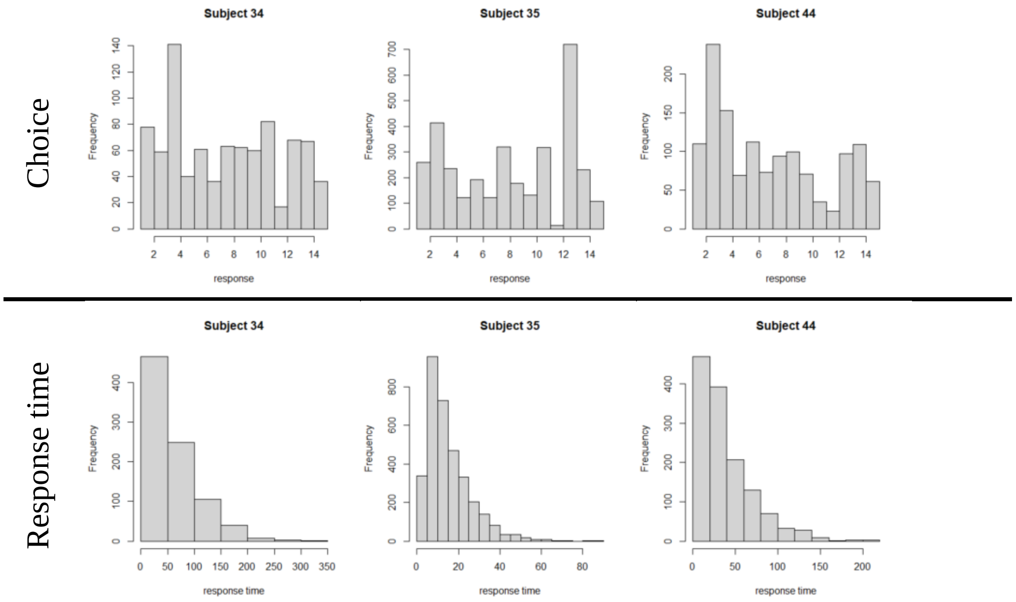


Figure 13. Example choice and response time distributions from simulated data

3.2.7. Parameter recovery and model investigation

Prior to estimating parameters on the real or simulated datasets, all versions of the model were run using known values for all the parameters to determine if the models could recover the original parameter values. This was completed using both the MLE and PMwG methods. Some model parameters were also adjusted over multiple model runs to determine the effect of changing different values on the ability of the model to recover the original parameters. Section 5.1 provides an explanation and details for this.

3.2.8. Comparing posterior distributions to original distributions

Different metrics were used to evaluate the similarity of the choice distributions and response time distributions. The first section discusses the Kullback-Leibler divergence, which was used to compare the original and posterior predicted choice distributions. Then, the second section discusses the two-sample Kolmogorov-Smirnov (K-S) test, which was used to compare the response time distributions.

3.2.8.1. Comparing choice distributions

$$D_{KL}(P||Q) = \sum_{x \in X} P(x) \log\left(\frac{P(x)}{Q(x)}\right) \tag{13}$$

The Kullback-Leibler (K-L) divergence (equation 13) measures the difference of one probability distribution from another reference probability distribution (e.g., McElreath, 2020). The formula is usually applied with known distributions (e.g., uniform, Gaussian, binomial) as the reference distribution where zero values indicate the values are outside of the true distribution. For this research, the reference distribution was the distribution of choices in the original dataset and the K-L divergence quantified how closely the posterior distribution of choices generated using the estimated model parameters matched the choice distribution of the original data, where the posterior distribution had more samples than the original data. The data used in this research had zero values within the true distribution as well as within the posterior distribution. The Project RED data had many subjects that selected some of choice options zero times and the posterior distributions also had some subjects with zero responses for some of the choice options, so the formula was modified to add 1 to each option and the total number of choices (e.g., 15) to the denominator when calculating *p* and *q*. For subjects with a large number of trials, such as the simulated data or some of the Project RED subjects, this modification had little effect on the K-L divergence value. When all other choice options have counts in the tens or hundreds, changing a zero to a one is negligible. However, for subjects with a very small number of trials (e.g., 5) this adaptation smoothed the reference distribution and led to higher K-L divergence values for those subjects. This was true for results from all the models, though, and since the range of K-L divergence over all subjects was used when comparing the models, the effect of this modification should not change the results of the comparisons.

Since the K-L divergence is a relative measure, a baseline needed to be established to compare to each model's K-L divergence (i.e., where the measured K-L divergence falls within the baseline distribution). The baseline was created using a simulated 2-subject, 1000-trial dataset for the reference distribution, which was generated by the single-stage CPT model using input parameters of b=5, A=20, α=1, γ=1.5, and t_0=10. These values generated a distribution with variation in response selection (in terms of counts not specific to a choice option) that was similar to the real data, and were also used to investigate parameter recovery, in section 5.1. The choice distribution for the baseline dataset is shown in Figure 14a. The same single-stage CPT model then simulated another 10,000 datasets using parameters randomly selected from a distribution of values, shown in Figure 15, to generate distributions of choices to compare to the original. The threshold, starting point parameter, and non-decision time were selected from lognormal distributions and the α and γ values are selected from normal distributions. The K-L divergence was calculated for each distribution of responses. The baseline was run using a K-L divergence calculation that did not account for zero responses; after removing NA values the baseline contained 8938 K-L divergence measures with a minimum value of 8.6e-5 (i.e., the best value), a maximum value of 1.26 (i.e., the worst value) and a median of 0.133. The choice distribution for the maximum K-L divergence in the baseline (Figure 14b) was very obviously different than the reference distribution, while the distribution for the minimum K-L divergence in the baseline (Figure 14c) appeared to be identical to the reference distribution. The baseline K-L divergence (Figure 16) provided a distribution of values to use for quantifying how well the posterior distributions matched the original distributions of the simulated and real data.

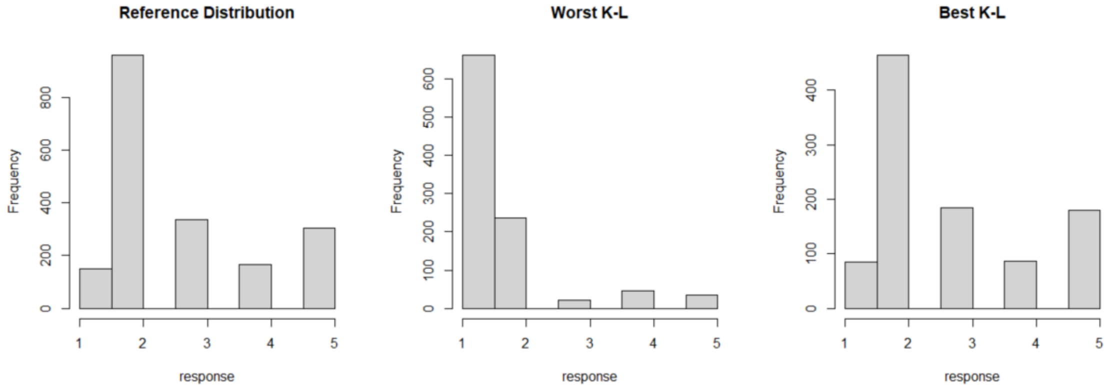


Figure 14. Choice distributions from the K-L divergence baseline for (a) the reference distribution (b) the maximum K-L divergence value in the baseline, and (c) the minimum K-L divergence value in the baseline

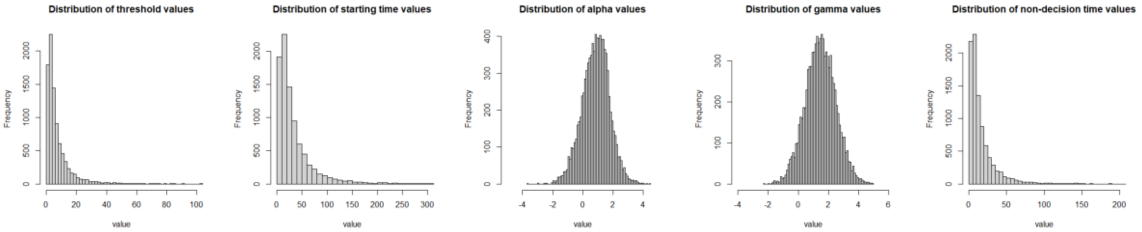


Figure 15. Distribution of values for parameters used to generate datasets for determining the K-L divergence baseline

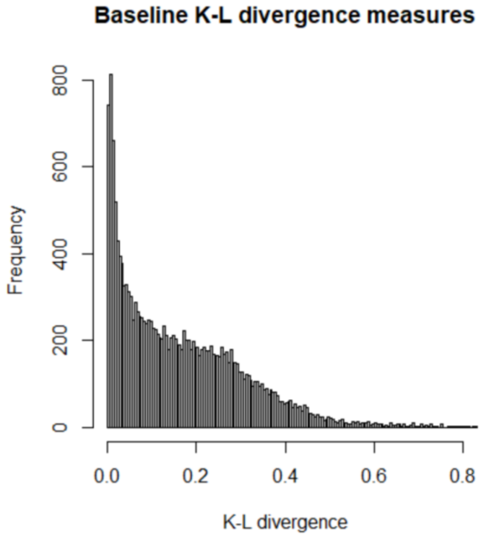


Figure 16. Distribution of baseline K-L divergence values

3.2.8.2. Comparing response time distributions

This research used two-sample Kolmogorov-Smirnov (K-S) test (equation 14) to determine if the posterior distribution of response times generated using the estimated model parameters matched the response time distribution of the original data. The two-sample K-S test tested the null hypothesis that two continuous samples come from the same distribution (Massey, 1951). For the two-sample version of the test, the test statistic (*d*) was the largest absolute deviation between the two observed cumulative step functions, irrespective of the direction of the difference. The closer *d* was to zero, the more likely that the two samples came from the same distribution. The *ks.test* function in the *stats* R package also outputs a p-value for the test statistic that has the same interpretation as other p-values. If the p-value was less than the pre-designated significance level of α=0.05, then the null hypothesis that the two samples were drawn from the same distribution was rejected.

$$d = \max|S_1(Y) - S_2(Y)| \tag{14}$$

where

d is the maximum deviation Kolmogorov statistic, $S_1(Y)$ is the observed cumulative distribution of sample 1, and $S_2(Y)$ is the observed cumulative distribution of sample 2.

3.2.9. Model Comparison

Both mean absolute error (MAE) and root mean square error (RMSE) were calculated to measure the accuracy of the best-fit or estimated parameter values. Both are standard metrics for determining the accuracy of an estimated value where a lower value is better. MAE was calculated using equation 15 and RMSE was calculated using equation 16.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|X_{obs,i} - X_{pred,i}\right| \tag{15}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}\left(X_{obs,i} - X_{pred,i}\right)^2}{n}} \tag{16}$$

Additionally, the resulting best-fit parameters from the different versions of the model found by MLE were compared using the Bayesian information criterion (BIC) and Bayes factors. BIC (Equation 17) is a type of information criteria, to construct a theoretical estimate of the relative out-of-sample K-L divergence, and is a commonly used metric for comparing models. It is an approximate method of comparison and generally gives a larger punishment to models with more parameters.

$$BIC = -2 \ln L(\hat{\theta}|y, M) + k \ln n \tag{17}$$

The first term is the deviance; it is -2 times the maximized value of the log-likelihood of the model, M. The second term accounts for model complexity using k, the number of free model parameters, and n, the number of data points on which the likelihood is based. When comparing several models, the one that produces the lowest BIC was preferred. The strength of evidence against a model with the higher BIC was defined (Kass & Raftery, 1995) using the ΔBIC, where 2-6 is positive evidence, 6-10 provides strong evidence, and greater than 10 is very strong evidence against the model with the higher BIC value.

Bayes factor (BF) is a measure the amount of evidence in favor of one model over another by calculating the ratio marginal likelihoods. The Bayes factor can also be approximated from the BIC values (Equation 18; Wagenmakers, 2007), expressed in terms of evidence in favor of Model 1 over Model 2. This research used commonly accepted guideline for the strength of evidence of favor of Model 1 (Raftery, 1995) which defined BF upper thresholds of 3, 20, and 150 for weak, moderate, and strong evidence, respectively.

$$BF = \exp\left(\frac{BIC_2 - BIC_1}{2}\right) \tag{18}$$

Watanabe-Akaike Information Criterion, also known as Widely Applicable Information Criterion, (WAIC; Equation 19) is another estimate of out-of-sample deviance that, for a large sample, converges to the cross-validation approximation (McElreath, 2020;

Watanabe, 2010). Unlike the BIC, it makes no assumptions about the shape of the posterior distribution. The increased generality of the WAIC comes from a more complicated formula. The first term is the log-pointwise-predictive-density; this is the Bayesian version for measuring the distance from the target. The second term adds a penalty proportional to the variance of the posterior predictions; it provides an overfitting penalty. It is also referred to the as the effective number of parameters, p_{WAIC}. The p_{WAIC} is the same as computing the variance in log-probabilities for each observation *i*, and then summing these variances.

$$WAIC = -2(lppd - \sum_i var_\theta \log p(y_i|\theta)) \tag{19}$$

As the label of the first term in the WAIC indicates, WAIC is a pointwise measure. Prediction is considered point-by-point in the data, leading to WAIC having an approximate standard error. This means also that the data can be split into independent observations, which creates difficulty in understanding the resulting WAIC value for data where a prediction depends on the previous prediction (e.g., time series data). For this research, while there were some response options that followed others more frequently (Figure 4), there was no strong evidence that any response depended on the previous response, making WAIC a reasonable metric to use for comparing the MLBA models. The WAIC was used to compare single-stage and two-stage MLBA models utilizing the PMwG method of parameter estimation. Like BIC, a smaller WAIC value provided evidence in favor of that model. ΔWAIC tells how different each model was from the best model. There is not a list of standard values to define the strength of evidence for ΔWAIC; instead, the standard error of the difference in the WAIC estimates was compared to the difference in the estimates. If the ΔWAIC was reasonably larger than the standard error of the differences, then the model with the lower WAIC had improved out-of-sample accuracy over the other model.

3.3. Counterfactual predictions

This research used Bayesian generalized linear models (GLMs) to make predictions about participant switch rate and task performance to address the fourth research question. The associated hypotheses were tested by comparing out-of-sample predictions from models that used the residual parameters from the decision strategy and cognitive processing modeling, which are referred to as the revised models, to a baseline intercept-only model. The baseline and revised models addressed the three counterfactual questions using different data to build each model. The between-subjects question used all of the combined campaign 3 and campaign 4 ‘not withheld’ data to build the model. The counterfactual predictions were made for the HERA ‘all-female crew’ participants in campaign 3, mission 1 for all 3 sessions. The within-subjects question used data from only the HERA participants in sessions 1 and 2 of each mission within campaign 4 to build the model. Only the HERA participants repeated the task multiple times. The counterfactual predictions were made using HERA participants’ data from session 3 of each mission within campaign 4. The other question used the campaign 3 data to build the model and the session 4 data from each mission within campaign 4 to make the counterfactual predictions.

This research established baseline intercept-only models of task-switching, individual performance, team performance, and multi-team performance for each counterfactual question using the data as described above. The performance values were normalized, as described in section 2.1. A binomial distribution was used for task-switching and a Gaussian distribution was used for performance. The baseline model had no predictors.

$$
\begin{gathered}
sc_i \sim Binomial(n_i, p_i) \\
logit(p_i) = \alpha \\
\alpha \sim Normal(0,1)
\end{gathered}
$$

$$
\begin{gathered}
perf_i \sim Normal(\mu, \sigma\ ) \\
\mu \sim Normal(0,1) \\
\sigma \sim Exponential(1)
\end{gathered}
$$

Revised models for both task-switching and performance were built for each question, using different predictors for each question to test the proposed hypotheses. The revised models used the participant's cluster assignment as a predictor to test the first hypothesis associated with within-subject counterfactual predictions. The first hypothesis focused on including decision-making strategy as a contextual factor. Including the most important factors as well as structuring the model as a multi-level model with participant ID as a hyperparameter were also explored, but were not part of the original hypothesis. For the second hypothesis, the revised models used a linear combination of the best LBA model's parameters as predictors to test including cognitive process model parameters as contextual factors for both within-subject and between-subjects counterfactual prediction. For the third hypothesis, the models used a linear combination of participant cluster and the best LBA model parameters as predictors to test including both the decision-making strategy and cognitive process model parameters as contextual factors for all three types of counterfactual questions.

The models were compared using Pareto-smoothed information sampling (PSIS) leave one out information criterion (LOO-IC) to compare the models (McElreath, 2020; Vehtari et al., 2020). For each question, the revised models were compared to the baseline models. The resulting predicted outcomes from each model were compared using RMSE and a visual assessment.

4. DECISION STRATEGIES RESULTS

The strategy used by a participant to complete the overall task was not identified or controlled as part of data collection so this research used machine learning techniques to identify clusters of similar participants that use a similar strategy. The clusters were compared to a baseline random assignment model, to test the hypothesis for the first research question, and the results showed that the learned clusters better predicted out-of-sample categories than the baseline, confirming the first hypothesis. The results included a combination of several machine learning methods, as shown in Figure 28. This chapter first discusses the results from data reduction, then describes the supervised and unsupervised machine learning outcomes using several different sets of parameters to group participants, and concludes with a summary of the best algorithm.

4.1. Data Reduction

Principal components analysis was run on six sets of 12 parameters that describe the relationships between each of the 12 participants (see section 3.1.1) using all the 'not withheld' data. The first PC (PC1) accounted for only 18% of the variance within the interpersonal ties parameter set, but for each other set of the parameters PC1 accounted for at least 25% of the variance and up to almost 50% for the tool dependencies, as shown in Figure 17. The first PC was used to reduce each of the parameter sets by a factor of 12 (from 12 values to 1). Both the full set and the reduced set of parameters were used to generate clustering results and the results were compared to select set of parameters that produces the best clustering results.

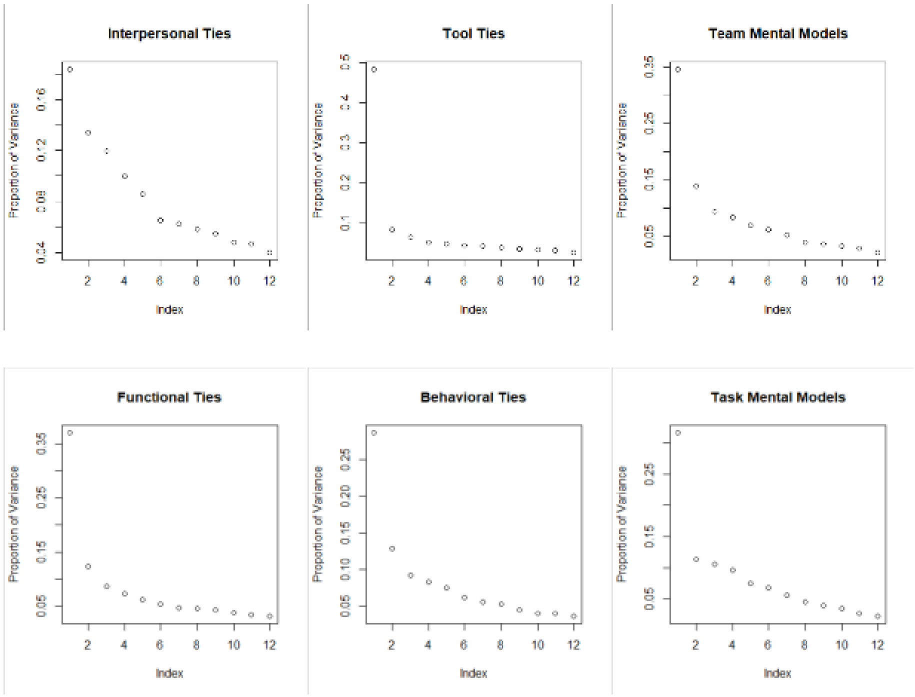


Figure 17. Proportion of variance accounted for by each PC for the six parameter sets

Personality measures, as described in section 2.1, were not included in any models used to evaluate the first hypothesis due to unknown differences in scale of the measured values as well as a large number of missing values in the campaign 4 data. A model was run that included personality variables along with the PCA reduced set of predictors to generate clusters of participants, for the training dataset (i.e., 80% of the ‘not withheld’ data), using the unsupervised random forest (uRF), as described in section 3.1.3. The uRF model identified the personality measures, in the order of agreeableness, conscientiousness, extraversion, and neuroticism, as the most important. After performing PAM clustering with the cMDS reduced uRF dissimilarity matrix, plotting the results showed two clear groupings (Figure 18). Labeling the points using the campaign grouping showed that the grouping was due to the differences in the personality measures between the two campaigns. These results supported the decision to not include personality measures in any models to identify clusters of decision-making strategies.

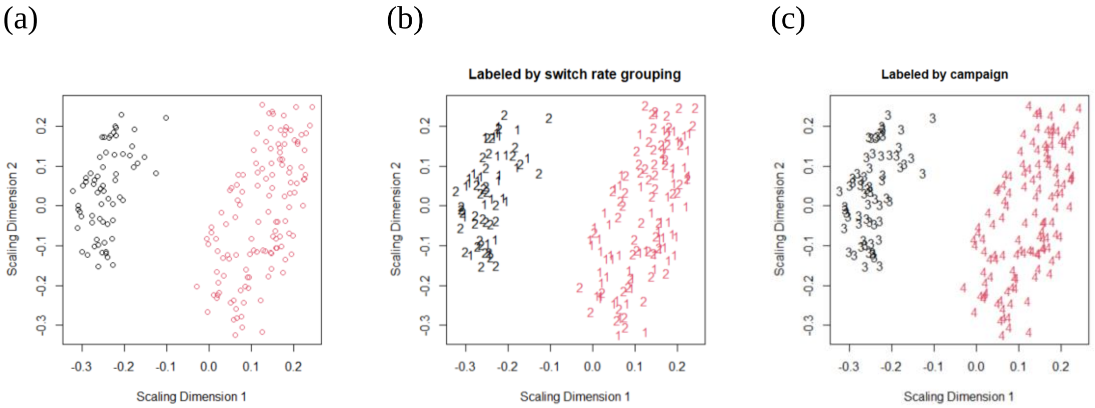


Figure 18. PAM clusters of cMDS scaling coordinates for (a) unlabeled cMDS clusters, (b) using numbers to label the switch rate 'true' clusters, and (c) using numbers to label campaign. Colors show cMDS clusters for all three plots.

4.2. Decision-making Strategies

The models in this section focused on identifying patterns of individual behavior related to the participant's task-switching rates. This research proposed that the patterns indicate differences in the strategies used by clusters of participants in completing the overall task. While the outcome of interest is an individual's task-switching, the models included predictors that measure both individual attributes and team interactions. Team task-switching information was unreliable, as shown in section 2.2. Because of this, plus because variables measuring team interactions were already included in the individual strategy modeling, team decision-making strategy was not explicitly modeled.

4.2.1. Most Important Predictors

The most important predictors of task-switching behavior were found with a regression random forest model, using 500 trees, on the training dataset using the PCA reduced set of input variables and an outcome of switch rate. This model explained 69% of the variance in the task-switching data with a mean squared error of 9.5e-06. The model provided a good fit to the data, but with 64 terminal nodes the results were hard to interpret. The importance metrics identified which parameters contributed most to the trees, but did not identify any similarities between the participants. For the supervised RF model, the most important variables were: typical task switches, explorer indicator variable, mean task priority, and mean task salience. Each variable contributed more than a 10% increase in MSE (Figure 19), and modeling each variable alone showed that typical task switches explained 40% of the variance, explorer status explained 25%, mean task priority explained 15%, and mean task salience explained 5%.

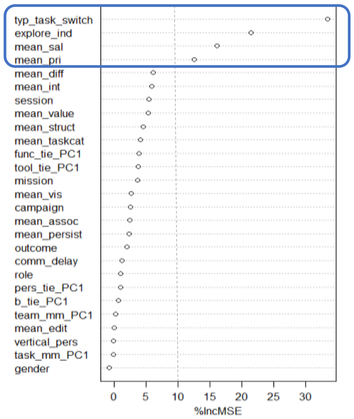


Figure 19. Most important predictors of switch rate – regression RF, PCA reduced dataset

Applying the model to the test dataset (n=48) gave a Pearson’s r correlation of 0.82 between the predicted and true switch rates (Figure 20). A permutation test cross-validation using 99 models, performed using the *rf.crossValidation* function in the *rfUtilities* package, (Evans & Murphy, 2018) gave a MAE cross-validation error variance of 8.2e-08. The regression RF provided good out-of-sample switch rate predictions, but the predictions did not explain the patterns of behavior.

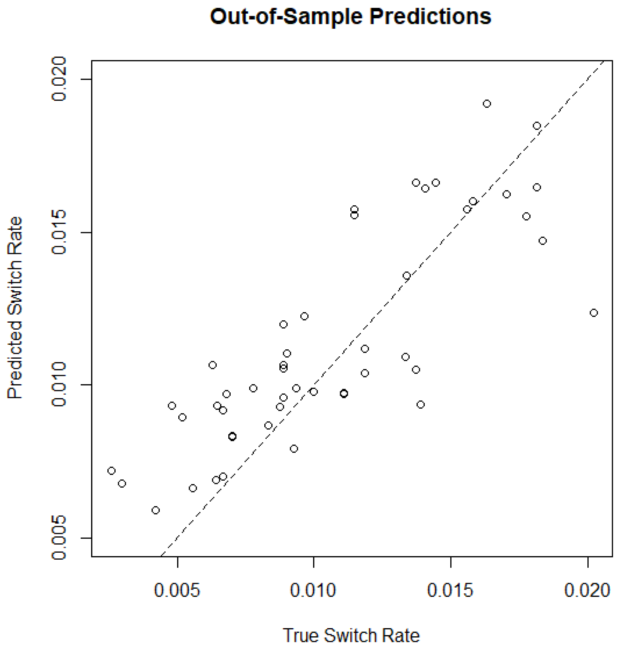


Figure 20. Comparison of true switch rate to predicted switch rate for the test dataset using regression RF model

4.2.2. Strategy Clusters

This section discusses the results of the clustering analysis run on three different sets of predictors, to identify the model that best described the similarities in participants and that predicted the best cluster assignment of out-of-sample data.

4.2.2.1. All predictors

Unsupervised random forest models using all 94 predictors on the training dataset identified mean task interest, mean structure, mean task difficulty, and tool ties to role 9 as the most important variables related to similarity between the participants (i.e., the dissimilarity matrix). The unsupervised random forest did not include switch rate; both the importance measures and the dissimilarity matrix were independent of switch rate. Using the uRF dissimilarity matrix in the clustering analysis to generate 10 clusters, which was the optimal number of clusters determined from the total within clusters sum of squares, gave results that performed almost the same as random chance (Table 7), whether using PAM clustering on the raw uRF dissimilarity matrix, PAM clustering on the cMDS of the uRF dissimilarity matrix, or Ward hierarchical clustering on the raw uRF dissimilarity matrix. The model was rerun using 2, 4, or 6 clusters with no improvement in the results. Figure 21 visualizes the lack of distinction in the clusters and Figure 27a confirms the poor clustering fits for these models. Due to the poor results with the in-sample data no out-of-sample predictions were generated.

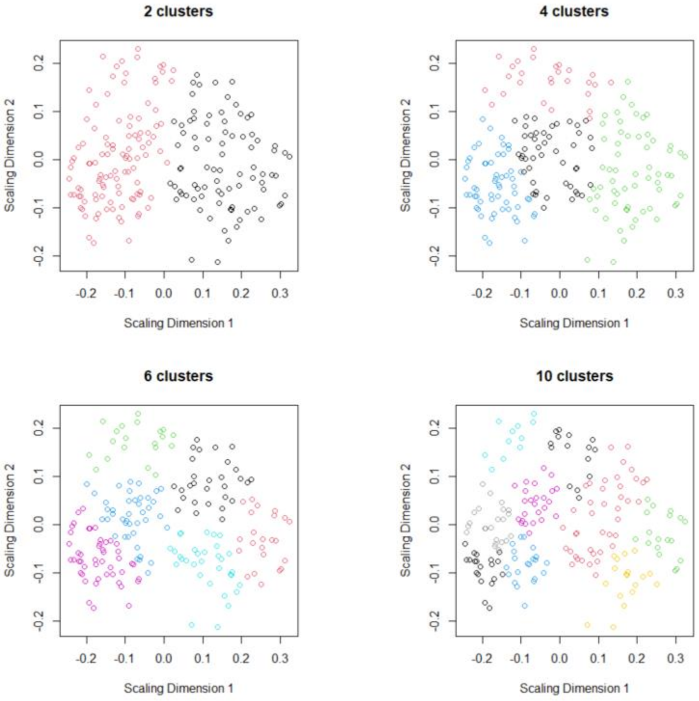


Figure 21. PAM clusters of cMDS scaling coordinates based on uRF dissimiliarity matrix for 2, 4, 6, and 10 clusters using full dataset.

Table 7. Unsupervised RF clustering metrics (in-sample) – 10 clusters using all predictors

Cluster Type	ARI	Error Rate
cMDS (PAM)	0.00	0.78
uRF distance (PAM)	0.03	0.75
uRF distance (HC)	0.01	0.77
Random	0.00	0.80

4.2.2.2. PCA Reduced Predictors

An unsupervised random forest was also run using the PCA reduced set of 27 predictors on the training dataset. The optimal number of clusters using this dataset, based on the total within clusters sum of squares, was also 10, but models were also generated using 2, 4 and 6 clusters. The uRF identified mean task interest, mean task structure, PC1 of tool ties, mean task difficulty, and PC1 of functional ties as the most important variables, which mostly matched the most important variables from the all-predictor uRF; both the importance measures and the dissimilarity matrix were again independent of switch rate. PAM clusters, for 2, 4, 6, and 10 clusters, of cMDS scaling coordinates based on the uRF dissimilarity matrix, shown in Figure 22, visualize the lack of distinction in the clusters.

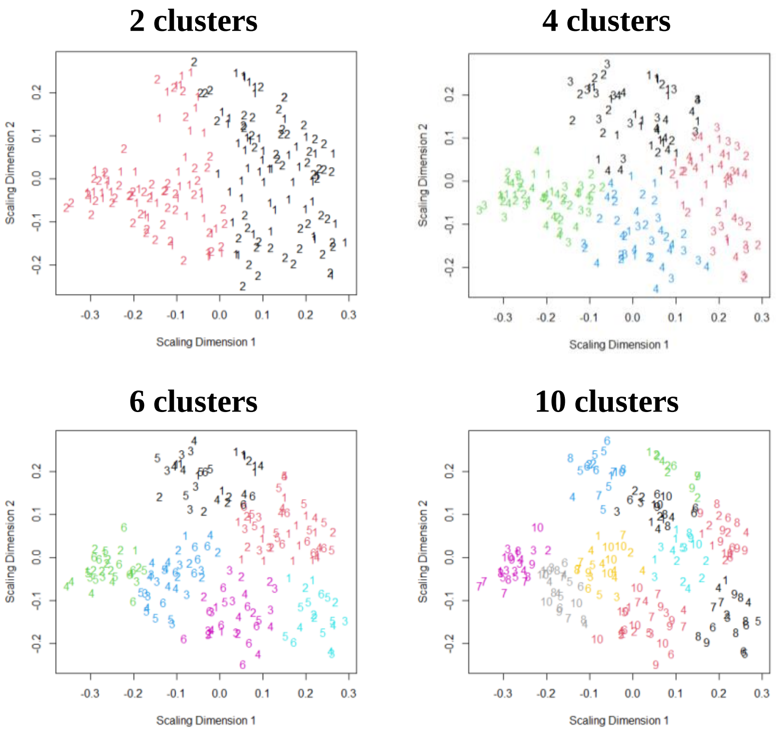


Figure 22. PAM clusters of cMDS scaling coordinates based on uRF dissimiliarity matrix for 2, 4, 6, and 10 clusters using PCA reduced dataset. Colors show cMDS clusters and numbers show 'true' clusters

This dataset performed slightly better than the full set of predictors, but most ARI values were still close to zero and the error rates were approximately the same as random chance, as shown in Table 8; Figure 27b visually confirms the poor fits. Hierarchical clustering algorithms, as well as using 10 clusters with PAM clustering, gave results just slightly better than random chance, but still too close to random chance to generate any out-of-sample predictions with the models.

Table 8. Unsupervised RF clustering metrics (in-sample) – PCA reduced predictors

Cluster Type	Number of Clusters	ARI	Error Rate
cMDS (PAM)	2	0.00	0.48
	4	0.01	0.66
	6	0.02	0.71
	10	0.01	0.76
uRF distance (PAM)	2	0.00	0.48
	4	0.02	0.65
	6	0.01	0.73
	10	0.03	0.72
uRF distance (HC)	2	**0.04**	0.40
	4	**0.03**	0.64
	6	**0.03**	0.70
	10	**0.03**	0.75
Random	2	0.00	0.50
	4	0.00	0.69
	6	0.01	0.72
	10	0.00	0.80

The supervised random forest, which was run using the PCA reduced set of 27 predictors on the training dataset, also generated a dissimilarity matrix to input into the different clustering algorithms. The most important variables from the supervised random forest were different than the most important variables from the unsupervised random forest (Table 9). It identified typical task switches, explorer indicator variable, mean task priority, and mean task salience as the most important variables; for this model the importance measures and the dissimilarity matrix were not independent of switch rate.

Table 9. Most important predictors using unsupervised RF and regression RF

Rank	Unsupervised RF	Regression RF
1	Mean task interest (6)	Typical tasks (14)
2	Mean task structure (9)	Explorer (27)
3	PC1 tool ties (12)	Mean task salience (9)
4	Mean task difficulty (5)	Mean task priority (10)

Note: Numbers in parentheses are the importance of that factor in the other model, determined by percent included MSE

Like the uRF models, the optimal number of clusters with the supervised RF dissimilarity matrix, based on the total within clusters sum of squares (WSS), was 10. However, unlike the uRF models, the WSS dropped steeply until it reached 3 clusters (Figure 23), which indicated that 3 clusters should give class consistency that is about the same as 10 clusters. The cMDS plots (Figure 24), color coded by PAM cluster and number-coded by ‘true’ grouping, showed overlap in the switch rates assigned to each cluster, but there was greater separation in the 2-cluster output. Based on the WSS and cMDS plots, plus the fact that dissimilarity matrix was not independent of switch rate and more clusters can lead to overfitting, the clustering algorithms were run twice to generate outputs with 2 and 3 clusters using PAM clustering and average hierarchical clustering. Table 10 lists the performance metrics for the in-sample predictions using all the clustering techniques to generate 2 or 3 clusters. Hierarchical clustering with 2 clusters performed the best on the in-sample data (ARI = 0.25, error = 0.40). Out-of-sample predictions using the model on the test dataset (i.e., other 20% of the ‘not withheld’ data) were the same as random chance using both PAM clustering (ARI = 0, error = 0.6) and hierarchical clustering (ARI = 0, error = 0.6) methods on the RF dissimilarity matrix. While the in-sample results were better than both uRF models (i.e., using both the full set of predictors and the PCA reduced set), the out-of-sample predictions were not better than a random assignment model.

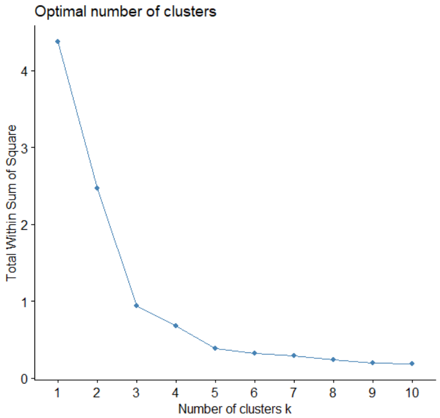


Figure 23. Optimal number of clusters for the regression RF dissimilarity matrix

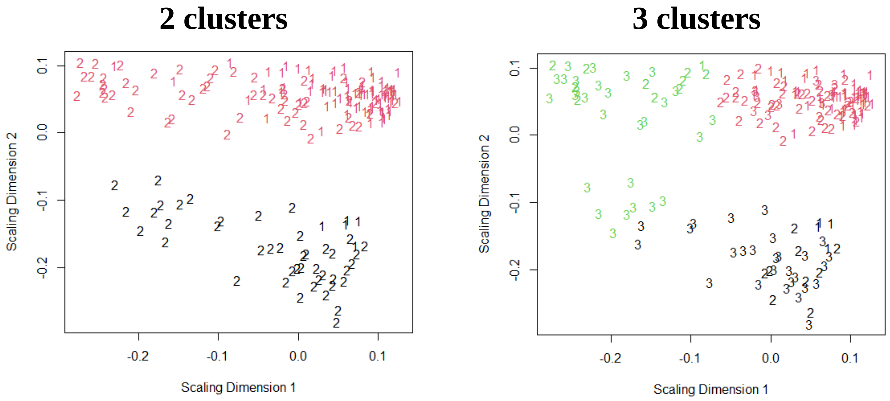


Figure 24. PAM clusters of cMDS scaling coordinates based on supervised RF dissimilarity matrix for 2 and 3 clusters using PCA reduced dataset. Colors show cMDS clusters and numbers show 'true' clusters.

Table 10. Regression RF clustering metrics (in-sample) – PCA reduced predictors

Cluster Type	Number of Clusters	ARI	Error Rate
cMDS (PAM)	2	0.22	0.41
	3	0.19	0.53
uRF distance (PAM)	2	0.18	0.46
	3	0.17	0.54
uRF distance (HC)	2	**0.25**	**0.40**
	3	0.21	0.53
Random	2	0.00	0.5
	3	0.00	0.66

4.2.2.3. ‘Top 4’ Predictors

As Table 9 shows, the most important predictors were different when using the unsupervised versus the supervised random forests. The results showed that using dissimilarity measure from supervised regression random forest in the clustering algorithms produce higher ARIs and lower than error rates than using the dissimilarity matrix from the uRF, which produces in-samples predictions only slightly better than chance. The clusters based on the supervised random forest dissimilarity measure were also more tightly clustered. Additionally, generating clusters using the uRF dissimilarity matrix from the PCA reduced set of variables performed slightly better than using all the variables. Therefore, the set of variables was further reduced to include only the most important variables; the values from the regression RF were chosen because that model provided the best in-sample predictions.

An unsupervised random forest was run using the set of ‘top 4’ regression RF predictors on the training dataset. The optimal number of clusters using this dataset, based on the total within clusters sum of squares, was again 10, but models were also generated using 2, 3 and 4 clusters. The cMDS plots (Figure 25) show a different pattern in the dissimilarity measures than the other models; using 2 clusters provided the most distinction between the clusters where the points in one cluster have a positive slope and the points in the other have a negative slope. Running the different clustering algorithms on the dissimilarity matrix, average hierarchical clustering using 2 clusters gave the highest ARI and lowest error rate for the in-sample predictions while PAM clustering gave the next best ARI and error rate, shown in Table 11. These values are better than any generated using full set of predictors or the PCA reduced set of predictors. Figure 27d visually confirms this as well. Out-of-sample predictions on the 20% test data using these two models perform better than chance with error rates of 0.31 (error rate of random chance = 0.50) and an ARI greater than zero but below the in-sample values (ARI_{HC} = 0.12, ARI_{PAM} = 0.12). Visual assessment of the boxplots (Figure 26) shows greater separation in center values of the clusters than random chance, but also a large amount of overlap in the values. Running leave-one-out cross-validation on the entire 'not withheld' dataset to generate out-of-sample predictions slightly increases ARI while error rates are approximately the same as the test dataset (ARI_{HC} = 0.17, error = 0.29).

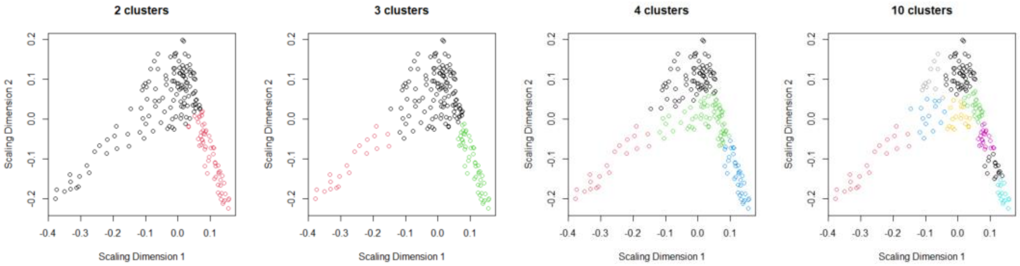


Figure 25. PAM clusters of cMDS scaling coordinates based on uRF dissimiliarity matrix for 2, 3, 4, and 10 clusters using 'top 4' dataset.

Table 11. Unsupervised RF clustering metrics (in-sample) – 'top 4' predictors

Cluster Type	Number of Clusters	ARI	Error Rate
cMDS (PAM)	2	0.16	0.30
	3	0.11	0.46
	4	0.10	0.54
	10	0.06	0.71
uRF distance (PAM)	2	**0.29**	**0.23**
	3	0.20	0.42
	4	0.16	0.50
	10	0.09	0.68
uRF distance (HC)	2	**0.31**	**0.22**
	3	0.21	0.43
	4	0.16	0.48
	10	0.11	0.72

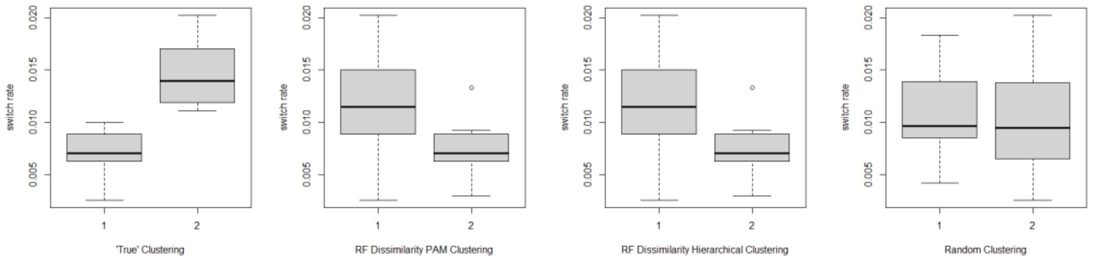
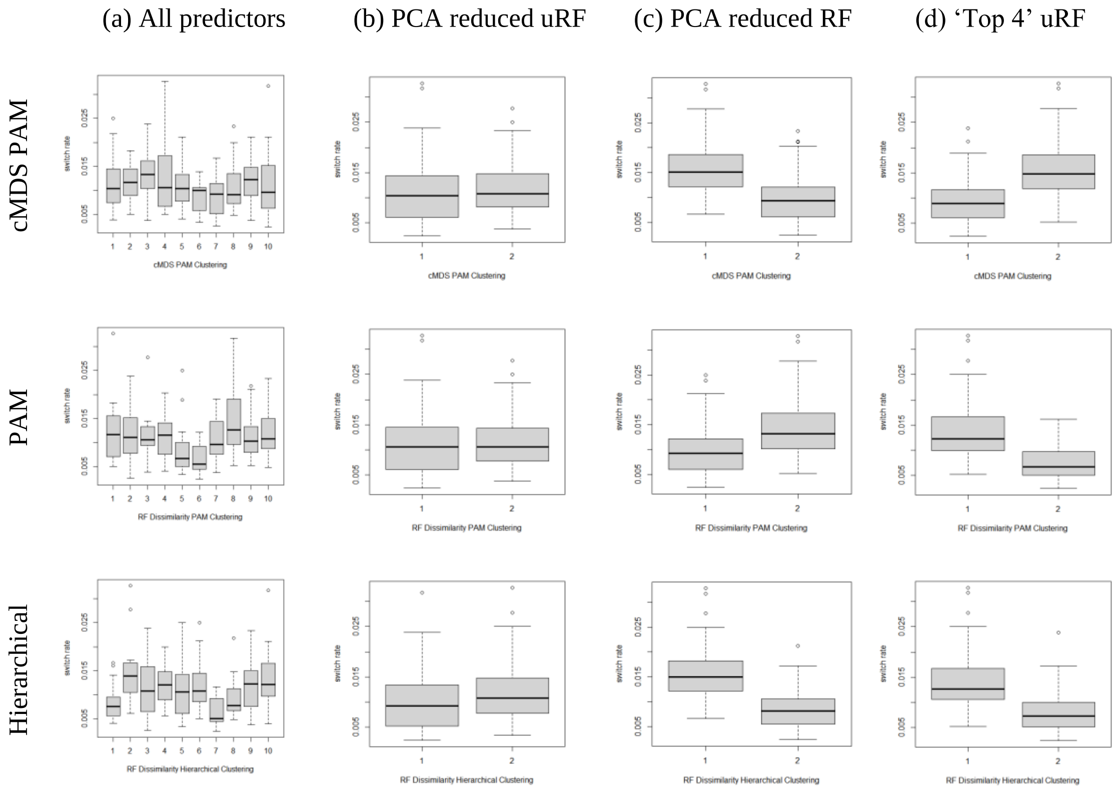


Figure 26. Clustering of 'top 4' uRF out-of-sample predictions


(a) All predictors (b) PCA reduced uRF (c) PCA reduced RF (d) 'Top 4' uRF

cMDS PAM

PAM

Hierarchical

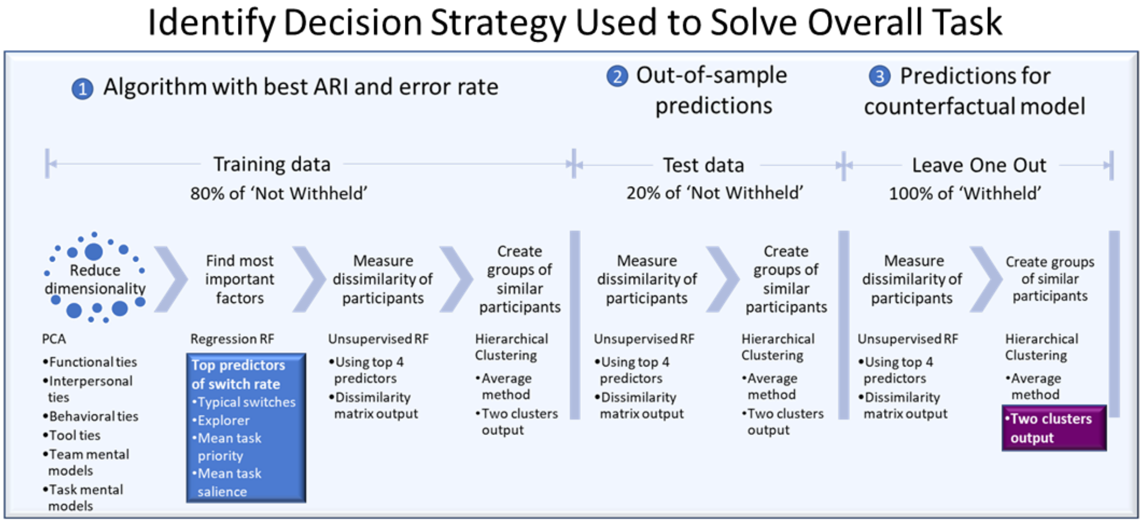
Figure 27. Clustering of in-sample predictions for (a) all-predictors uRF, (b) PCA reduced uRF, (c) PCA reduced RF, and (d) 'top 4' reduced uRF using PAM clustering on cMDS scaling coordinates (top row), PAM clustering on uRF/RF dissimilarity distance (second row), hierarchical clustering on uRF/RF dissimilarity distance (bottom row)

By performing better than random assignment to predict out-of-sample categories, this model confirmed the first hypothesis and addressed the research question of how participants group together in their strategies of completing the overall task. The model built using the ‘not withheld’ data was applied to the ‘withheld’ data to identify the clusters to use in the counterfactual prediction models.

4.3. Summary

The results from the decision strategies analysis support the hypothesis that patterns identified using machine learning predict the out-of-sample decision strategy category better than random assignment. The machine learning analysis identified patterns that indicated differences in clusters of participants completing the overall task. Assuming participants used different strategies to complete the task, the clusters represented the different strategies used by the participants. The model that used the top 4 most important predictors from the regression random forest in the unsupervised random forest (uRF) model combined with hierarchical clustering of 2 clusters on the uRF dissimilarity matrix gave the best in-sample ARI and classification error rate results for the training data (i.e., 80% of ‘not withheld’ data). The results showed that decreasing the number of variables (i.e., IVs for the uRF) increased clustering performance. When there were more variables, increasing the number of clusters slightly increased ARI, but also increased error rate, while using less variables with a smaller number of clusters gave higher ARIs as well as smaller error rates. Using this best model to predict out-of-sample results on the test data (i.e., 20% of ‘not withheld’ data) produced predictions better than random assignment. When this best model was run using a leave one out method to produce out-of-sample predictions the results were about as accurate as the out-of-sample predictions on the test data and better than random assignment. The results from this modeling identified commonalities between the participants that were included in the counterfactual modeling to try to improve the predictions of switch rate, individual performance, team performance, and MTS performance for the within-subject and other counterfactual question. The best model was run on the 'withheld' dataset, as shown in the far-right panel of Figure 28, to generate the cluster assignments that were used as inputs in the counterfactual prediction models.

Variables that were measured during task completion were needed in order to assign a participant to a cluster, but counterfactual predictions are intended to only use data that is available before the task is performed. Instead of a true counterfactual prediction, the results from the machine learning modeling were used to make predictions assuming that the cluster assignments were determined prior to task performance.



Identify Decision Strategy Used to Solve Overall Task

1 Algorithm with best ARI and error rate

2 Out-of-sample predictions

3 Predictions for counterfactual model

Training data
80% of 'Not Withheld'

Test data
20% of 'Not Withheld'

Leave One Out
100% of 'Withheld'

Reduce dimensionality

PCA
- Functional ties
- Interpersonal ties
- Behavioral ties
- Tool ties
- Team mental models
- Task mental models

Find most important factors

Regression RF

Top predictors of switch rate
- Typical switches
- Explorer
- Mean task priority
- Mean task salience

Measure dissimilarity of participants

Unsupervised RF
- Using top 4 predictors
- Dissimilarity matrix output

Create groups of similar participants

Hierarchical Clustering
- Average method
- Two clusters output

Measure dissimilarity of participants

Unsupervised RF
- Using top 4 predictors
- Dissimilarity matrix output

Create groups of similar participants

Hierarchical Clustering
- Average method
- Two clusters output

Measure dissimilarity of participants

Unsupervised RF
- Using top 4 predictors
- Dissimilarity matrix output

Create groups of similar participants

Hierarchical Clustering
- Average method
- Two clusters output

Figure 28. Decision Strategy Model pipeline

5. COGNITIVE PROCESS MODEL RESULTS

The research questions identified several possible cognitive mechanisms that participants could use to decide when to switch to a new task as part of a larger overall task. This research adapted and applied the multi alternative, multi-attribute linear ballistic accumulator models to describe and explain the cognitive mechanisms related to task-switching behaviors, to address the research questions. The models were compared to determine the mechanism most favored given the Project RED and simulated task-switching data. The results from the Project RED data were contradictory across the two campaigns, but the simulated data provided clearer results to determine the cognitive process that best described the task-switching behavior. The estimated parameters from this model explained that individual variations in task-switching behavior for Project RED were related to a bias to avoid switching as well as the attractiveness of the alternative tasks. This chapter first discusses the results from a parameter recovery and model investigation study, then compares the results of the models on the simulated data and the Project RED data, and concludes with a summary of the results.

5.1. Parameter recovery and model investigation

This research investigated the effect that various changes, including varying the number of trials, varying the number of response options, constraining the drift rate parameters to be positive, and changing the number of free parameters in the model, had on the results from each of the models. In addition, a parameter recovery study was run to ensure that the models were able to estimate a unique set of known parameters before running the models on the Project RED data with unknown parameters. Three of the four models used in this analysis were developed as part of this research and the model investigation and parameter recovery helped to understand the strengths and limitations of the models before applying them to real data. This section first describes the results using MLE and then using Particle Metropolis within Gibbs sampling.

5.1.1. Maximum Likelihood Estimation

Maximum likelihood estimation (MLE) was only run for the single-stage versions of the model using equations 1, 2, and 3 for the cumulative prospect theory (CPT) version and equations 6, 7, and 8 for the information foraging theory (IF) version. The CPT version and the IF version of the model each included 5 parameters of interest: threshold (*b*), starting point parameter (*A*), and non-decision time (t_0) for both versions plus the parameters related to value and attention, α and γ in the drift rate equation for the CPT version and α and β in the drift rate equation for the IF version, respectively. Threshold and one or both of the drift rate parameters (i.e., α and/or γ for the CPT version and α and/or β for the IF version) were left free for the MLE fitting. *A*, t_0, the initial drift rate (I_0) and the drift rate scaling factors (c_d and c_m) were fixed for all versions of the model. The MLE fitting considered three configurations of free parameters to fit for each model. For the CPT model the configurations were identified as CAG, where *b*, α and γ were allowed to vary; CA, where *b* and α were allowed to vary; and CG, where *b* and γ were allowed to vary. For the IF model the configurations were identified as IAB, where *b*, α and β were allowed to vary; IA, where *b* and α were allowed to vary; and IB, where *b* and β were allowed to vary.

To examine the ability of the single-stage CPT and IF models to recover the generating values of the model parameters (i.e., the values used to generate simulated data), two 20-subject simulated datasets were created for the CPT and the IF version of the model, as described in section 3.2.6, where the generating values for both drift rate parameters (i.e., α and γ for CPT; α and β for IF) varied for each subject. All other parameters (b, A, t_0, I_0, c_d, and c_m) were set constant across all subjects to the values used to generate the simulated data. These datasets did not allow for comparison between the CPT and IF models because different values were used to generate the datasets and the data differed in each dataset. The MLE method was used to find the best fit drift rate and threshold parameters for the three configurations of each model (i.e., CAG, CA, CG, IAB, IA and IB) to data simulated for 3- and 5-choice tasks. A, t_0, I_0, c_d, and c_m were fixed to the generating value for the model fitting.

For the 3-choice and 5-choice data, using the MLE best-fit parameters, a visual assessment determined that all configurations of the CPT model (i.e., CAG, CA, and CG) and IF model (i.e., IAB, IA, and IB) recovered the choice and RT distributions for most of the subjects (Figure 77 & Figure 78 in Appendix B), with the CPT models deviating less from the original distributions than the IF models. Using quantitative metrics, the Kullback-Leibler (K-L) divergence values (Figure 29) were less than the median of the baseline (D_{KL}=0.133) for all configurations of both the CPT and IF models. The K-L divergence values were lower using the results from the IF models than the CPT models for the 3-choice datasets and most of for the 5-choice datasets. The majority of the Kolmogorov-Smirnov (K-S) test statistic p-values were above 0.05 for the CPT models, as shown in Figure 30, indicating that the null hypothesis (i.e., the two samples are from the same distribution) should not be rejected. The RT distribution using the best-fit parameters was the sufficiently similar than the original simulated data. However, the IF models had many or all subjects with K-S test statistic p-values below 0.05 (Figure 30), indicating that the null hypothesis should be rejected. The CPT models, in general, had low K-L divergence and higher p-values for the K-S test statistic, indicating that responses generated with the best-fit MLE parameters from the CPT models more closely matched the responses and response times of the simple simulated dataset than using the IF models.

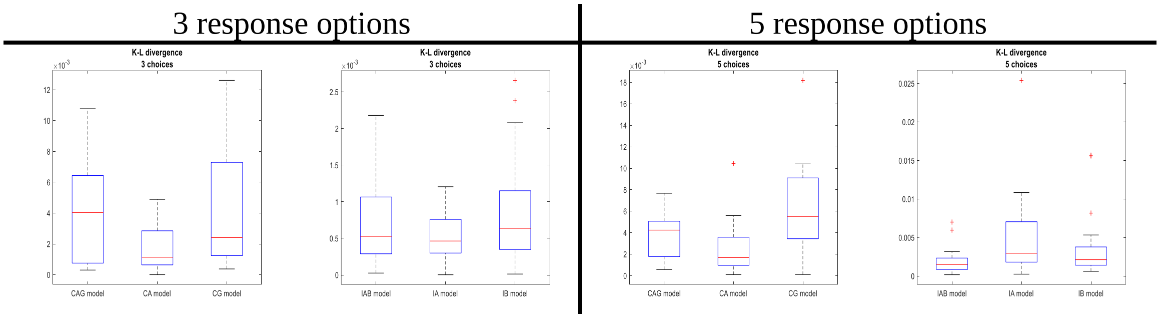


Figure 29. K-L divergence values for 3-choice and 5-choice simple simulated datasets

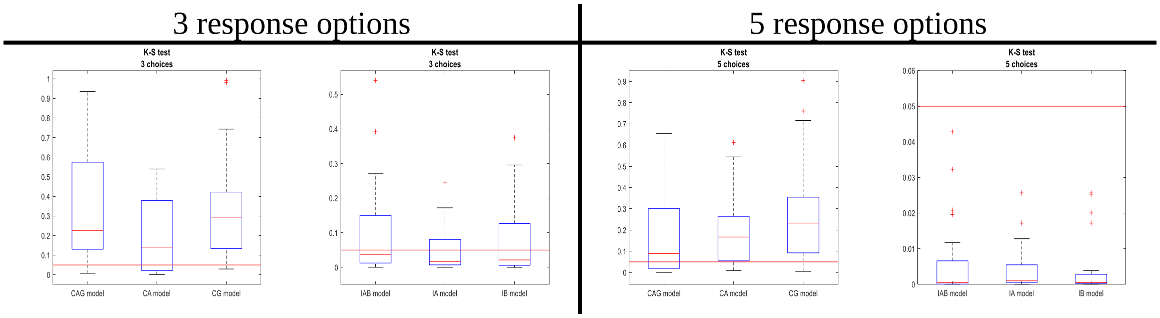


Figure 30. K-S test statistic p-values for 3-choice and 5-choice simple simulated datasets

None of the configurations of the CPT models or the IF models recovered all the generating values for the parameters. Table 37 and Table 38 in Appendix B include the best-fit values for each parameter for each subject for each model configuration. Table 12 provides a summary of the RMSE for best-fit values. For the CPT models (Figure 31), the best fit value for *b* and α were close to the generating value, but γ was not for any of the configurations. For the IF models (Figure 32), *b* was close to the generating value, but α and β were not. This seems to be a limitation of the MLE fitting method. The PMwG method was able to estimate values for the threshold and drift rate parameters, as well as the starting point and non-decision time parameters, that were close to the generating values. Detailed results are described in the next section.

Table 12. MLE best-fit parameter values (RMSE)

	CAG	CA	CG	IAB	IA	IB
3-choice dataset						
b	0.37	0.63	0.51	0.95	0.95	0.95
α	0.07	0.11	NA	177	1.49	NA
γ	1.66	NA	1.66	NA	NA	NA
β**	NA	NA	NA	218	NA	115
5-choice dataset						
b	0.15	0.37	0.40	0.94	0.94	0.94
α	0.12	0.04	NA	197	85	NA
γ	1.63	NA	1.61	NA	NA	NA
β**	NA	NA	NA	321	NA	2.09

** excludes values > 1e6

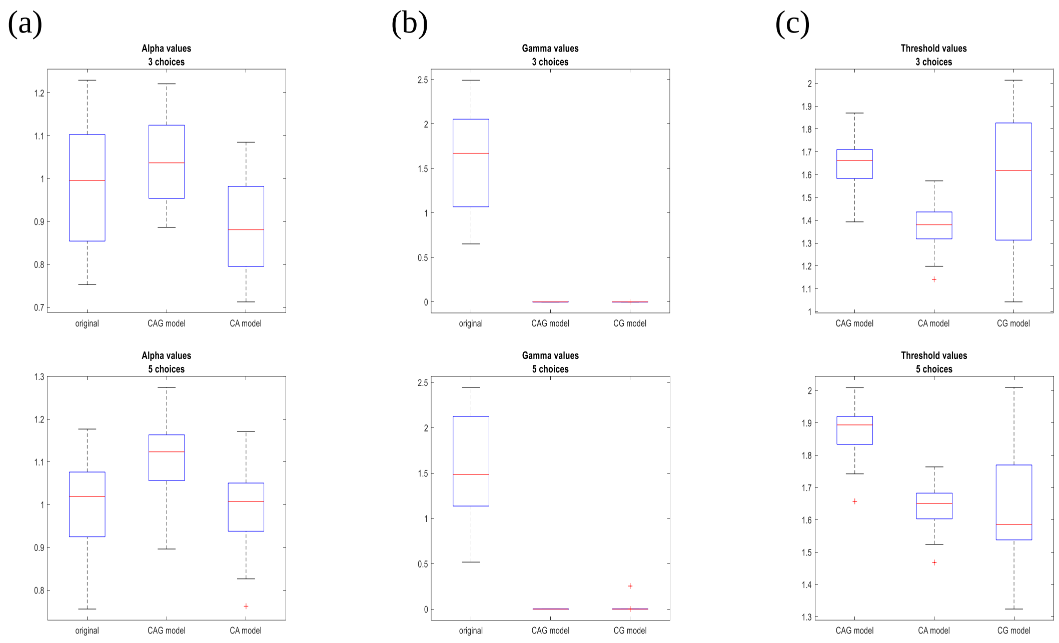


Figure 31. Original and best-fit values from the CPT models configurations with the 3-choice dataset (top row) and the 5-choice dataset (bottom row). (a) Original, CAG, and CA alpha values for all subjects. (b) Original, CAG, and CG gamma values for all subjects. (c) CAG, CA, and CG threshold values for all subjects. The generating value of threshold was 2 for all subjects for all the models.

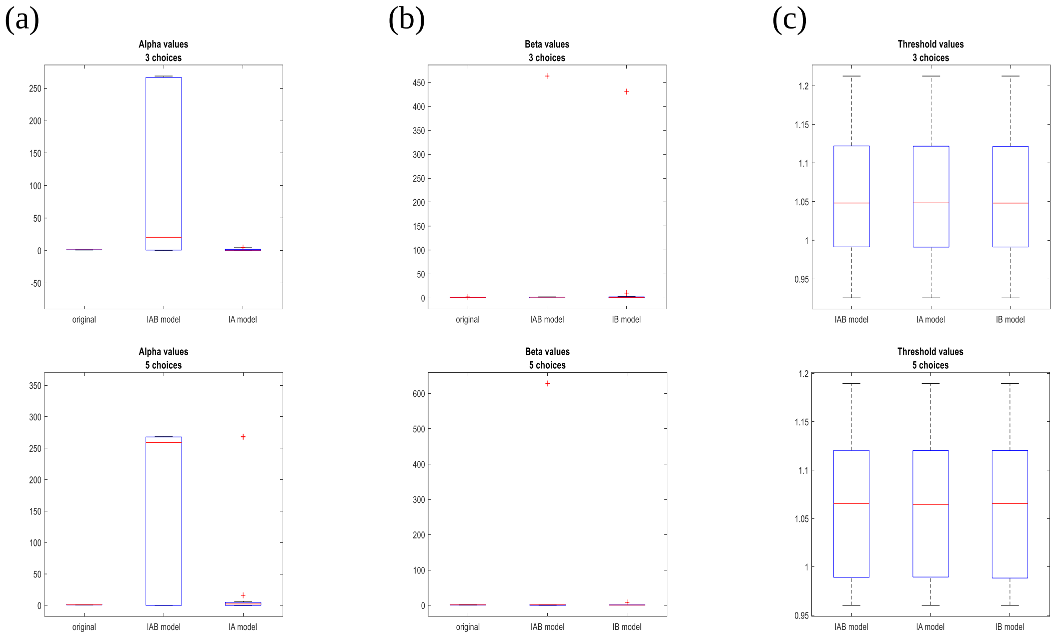


Figure 32. Original and best-fit values from the IF models configurations with the 3-choice dataset (top row) and the 5-choice dataset (bottom row). (a) Original, IAB, and IA alpha values for all subjects. (b) Original, IAB, and IB gamma values for all subjects, excluding value greater than 1e6. (c) IAB, IA, and IB threshold values for all subjects. The generating value of threshold was 2 for all subjects for all the models.

5.1.2. Particle Metropolis within Gibbs

The model investigation and parameter recovery study were run using Particle Metropolis within Gibbs sampling to estimate parameters for both the single-stage and two-stage versions of the MLBA model. The single-stage versions used equations 1, 2, and 3 for the cumulative prospect theory (CPT) version and equations 6, 7, and 8 for the information foraging theory (IF) version. The two-stage versions used equation 12 in addition to the other equations from the single-stage model.

5.1.2.1. Single-stage models

PMwG parameter estimation was run multiple times using the same initial (i.e., generating) values for all 5 parameters of the single-stage CPT model to investigate several different assumptions about the parameters. Two subjects were used to keep the computational run-times reasonable. Table 13 summarizes the RMSE to quantify how the estimated parameter values compared to the generating values for each of the model configurations, the Kullback-Leibler (K-L) divergence to measure the similarity of the choice distributions, and the Kolmogorov-Smirnov (K-S) test to measure the similarity of the response time distributions. First, the model was run to determine if there was an effect on the estimated parameters of allowing α and γ to be negative (M1) or restricting α and γ to be positive (M2). Both models recovered the choice and RT distributions, while both a visual assessment and the quantitative metrics (Table 13) determined that the first model, which allows α and γ to be negative, recovered the distributions better. Additionally, the shape of the γ distribution appeared to vary based on whether or not γ was allowed to be negative, as shown in Figure 79 in Appendix B. M1 produced posterior samples of γ with a normal distribution while M2 did not. Next, the model was run to compare the effect of number of response options on the estimated parameters by running one model with 5 response options (M1) and another with 15 response options (M3); α and γ were allowed to be negative for both. Both models recovered the choice and RT distributions and both recovered the generating parameters equally well. As expected, the 15-choice dataset took much longer to run. Finally, the model was run to compare the effect on the estimated parameters of varying the number of trials within the simulated dataset. This comparison showed that the estimated parameters for the dataset with 1000 trials (M1) were closer to the generating values than estimated parameters for the dataset with 100 trials (M4) and quantified the difference using the distribution of parameter values. As expected, the dataset with 100 trials had a larger difference between the estimated values and the generating values as shown using the RMSE in Table 13. The 100-trial dataset also had more error in the choice and response time distributions, as shown with the K-L divergence values and the K-S test statistic and its associated p-value. Of the 4 CPT models run, models 1 and 3 have the smallest RMSE values across all the parameters providing evidence that allowing α and γ to be negative values resulted in better estimated parameters. The rest of the models were run allowing α & γ to be negative values.

Table 13. PMwG estimated parameter metrics for CPT single-stage models

	Initial Values					K-L divergence (mean ± se)	K-S test statistic and p-value (mean ± se)
	b = 5	A = 15	α = 1	γ = 1.5	t_0 = 10		
	Estimated parameters (RMSE)						
M1	0.30	1.09	0.04	0.19	0.11	0.001 ± 0.0008	0.03 ± 0.0005 p = 0.66 ± 0.04
M2	0.17	0.48	1.04	1.16	0.07	0.0006 ± 0.0003	0.03 ± 0.003 p = 0.30 ± 0.09
M3	0.31	0.21	0.07	0.55	0.02	0.002 ± 0.002	0.02 ± 0.0002 p = 0.82 ± 0.01
M4	3.38	9.34	0.30	0.15	0.93	0.04 ± 0.03	0.06 ± 0.001 p = 0.87 ± 0.04

M1: α & γ can be <0, 5 choices, 1000 trials
M2: α & γ cannot be <0, 5 choices, 1000 trials
M3: α & γ can be <0, 15 choices, 1000 trials
M4: α & γ can be <0, 15 choices, 100 trials

A parameter recovery simulation was run using the version of the CPT model that allows α and γ to be negative. Multiple 5-choice datasets were created by varying the values of *b*, α and γ within this model, and then the same model was used to estimate the *b*, A, α, γ and t_0 parameters. Generating values of α varied from -2 to 2, γ varied from -2 to 2, and *b* varied from 4 to 15 with 8 different combinations of b, α and γ run. Every value of α and most values of γ were recovered, as shown in Figure 33. The threshold value was recovered for values that were farther from the starting point value. As the threshold value approached the starting point value the estimated parameters were further from the generating value. However, when *b* and *A* were added together, the sum of the estimated parameter values was again close to the sum of the original threshold and starting point. This provided confidence that the single-stage CPT model was identifiable and could be used to test the hypotheses relating to what cognitive mechanism best described the decision-making process.

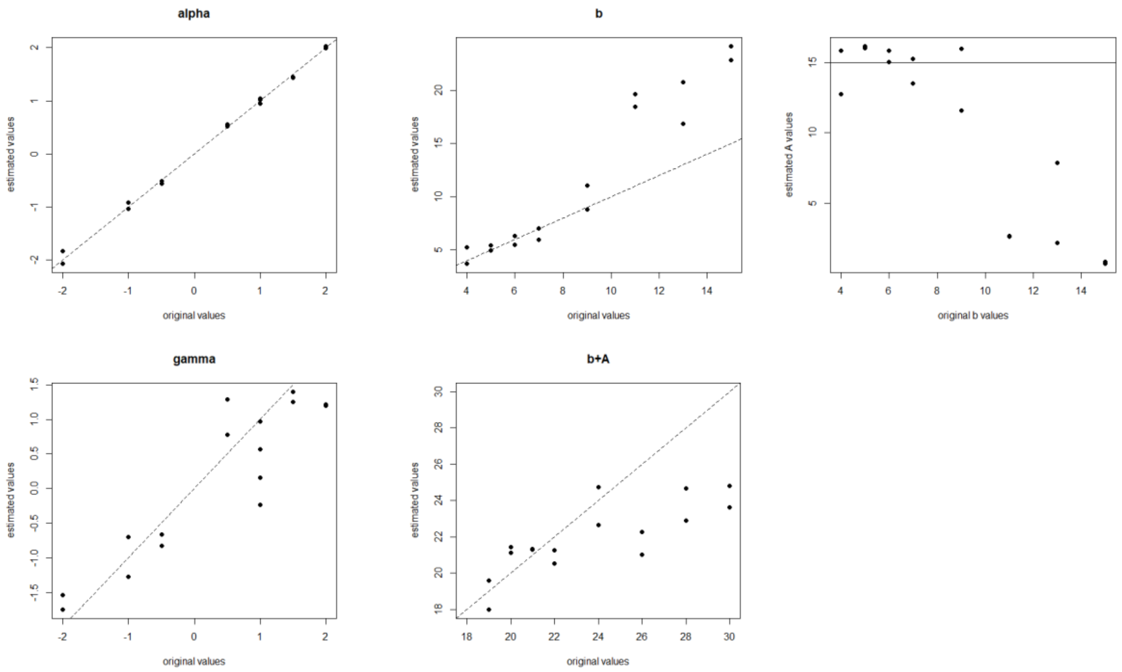


Figure 33. Original and estimated parameters values from the CPT model for threshold, starting point, α, and γ.

This research also investigated assumptions about the parameters of the single-stage IF model. Table 14 summarizes the comparisons, again using RMSE to quantify how the estimated parameter values compared to the generating values for each of the model configurations, the K-L divergence to measure the similarity of the choice distributions, and the K-S test to measure the similarity of the response time distributions. First, the model was run with a version of the model that let the profitabilities vary for each of the response options. This was run twice, once where the profitabilities (π_i) and α were allowed to be negative (M5) and a second time where they were forced to be positive (M6). All of the generating values were positive and both simulated datasets had five response options. Both versions recovered the threshold, starting point and non-decision time values. The version where π_i and α were forced to be positive (M5) output estimated parameters with smaller RMSE values, that were closer to recovering the generating values. A third version was then run where the profitabilities were again estimated, this time for each of 15 response options, where both α and π_i were allowed to be negative. This model (M7) only recovered the original t_0 value and the posterior samples were not normally distributed (Figure 80 in Appendix B), which was likely due to the tuning of the PMwG sampler. Two more versions of the model were run where the profitabilities are determined from the data, as described in section 3.2.2, and the β parameter was allowed to vary; one with five response options (M8) and the other with 15 response options (M9). Both these models were able to recover the generating parameters. The three versions of the model that estimated the profitability values had lower K-L divergence, lower K-S test statistics, and higher K-S test p-values, indicating that the posterior predictive distributions of choice and response time using the estimated parameters from these models best matched the original data. The two versions of the model the used the β parameter had lower RMSE, K-L divergence values that were below the baseline median value of 0.133, and K-S test statistic p-values that were mostly above 0.05. The version with 15 choices that used the β parameter (M9) actually had a lower K-L divergence than the model that estimated 15 π values (M7). To reduce the number of fit parameters and use a model that is identifiable, the version of the model that includes β was used for a parameter recovery simulation and for estimating parameters from the simulated and real data.

Table 14. PMmG estimated parameter metrics for IF single-stage models

Generating values	**M5**	**M6**	**M7**	**M8**	**M9**
b = 5	0.50	0.99	0.79	0.69	0.90
A = 15	1.76	0.58	1.56	0.29	2.08
α = 1/1/1.1/0.5/1	0.003	0.11	0.27	0.06	0.02
π_1 = 0.2	0.47	0.15	3.81	-	-
π_2 = 0.1	0.23	0.07	7.50	-	-
π_3 = 0.25	0.51	0.18	0.70	-	-
π_4 = 0.4	0.86	0.29	7.86	-	-
π_5 = 0.05	0.13	0.03	0.77	-	-
π_6 = 0.11	-	-	10.09	-	-
π_7 = 0.17	-	-	4.81	-	-
π_8 = 0.45	-	-	10.98	-	-
π_9 = 0.6	-	-	15.78	-	-
π_{10} = 0.02	-	-	2.88	-	-
π_{11} = 0.22	-	-	7.74	-	-
π_{12} = 0.3	-	-	4.18	-	-
π_{13} = 0.35	-	-	10.26	-	-
π_{14} = 0.7	-	-	17.53	-	-
π_{15} = 0.03	-	-	0.63	-	-
β = 0.2	-	-	-	0.11	0.03
t_0 = 10	0.07	0.16	0.07	0.20	0.28
K-L divergence (mean ± se)	0.001 ± 0.001	0.0004 ± 0.0002	0.011 ± 0.006	0.05 ± 0.03	0.006 ± 0.003
K-S test statistic (mean ± se)	0.022 ± 0.002	0.028 ± 0.004	0.020 ± 0.003	0.05 ± 0.009	0.028 ± 0.002
K-S test statistic p-values (mean ± se)	0.78 ± 0.08	0.52 ± 0.20	0.78 ± 0.14	0.05 ± 0.04	0.47 ± 0.11

M5: α & π_i can be <0, 5 choices, 1000 trials
M6: α & π_i cannot be <0, 5 choices, 1000 trials
M7: α & π_i can be <0, 15 choices, 1000 trials
M8: α & β can be <0, 5 choices, 100 trials
M9: α & β can be <0, 15 choices, 100 trials

Another parameter recovery simulation was run using the version of the IF model that includes α and β parameters, which are allowed to be negative, to create multiple 5-choice datasets by varying the values of *b*, *A*, α, β and t_0. The same model was then used to estimate the *b*, *A*, α, γ and t_0 parameters. The generating values of α varied from -1.5 to 1.5, β varied from -1.5 to 1.5, *b* varied from 2 to 6, *A* varied from 15 to 30, and t_0 varied from 5 to 15 with 10 different combinations of *b*, *A*, α, β and t_0 run. All values of t_0, most values of α and β, and most threshold and starting point values were recovered for values, as shown in Figure 34. Like the CPT version of the model, as the threshold value approached the starting point the estimated parameters are further from the generating value. However, when *b* and *A* were added, the sum of the estimated parameter values was approximately equal to the sum of the generating threshold and starting point. These results provided confidence that the single-stage IF model was identifiable and could be used to test the hypotheses.

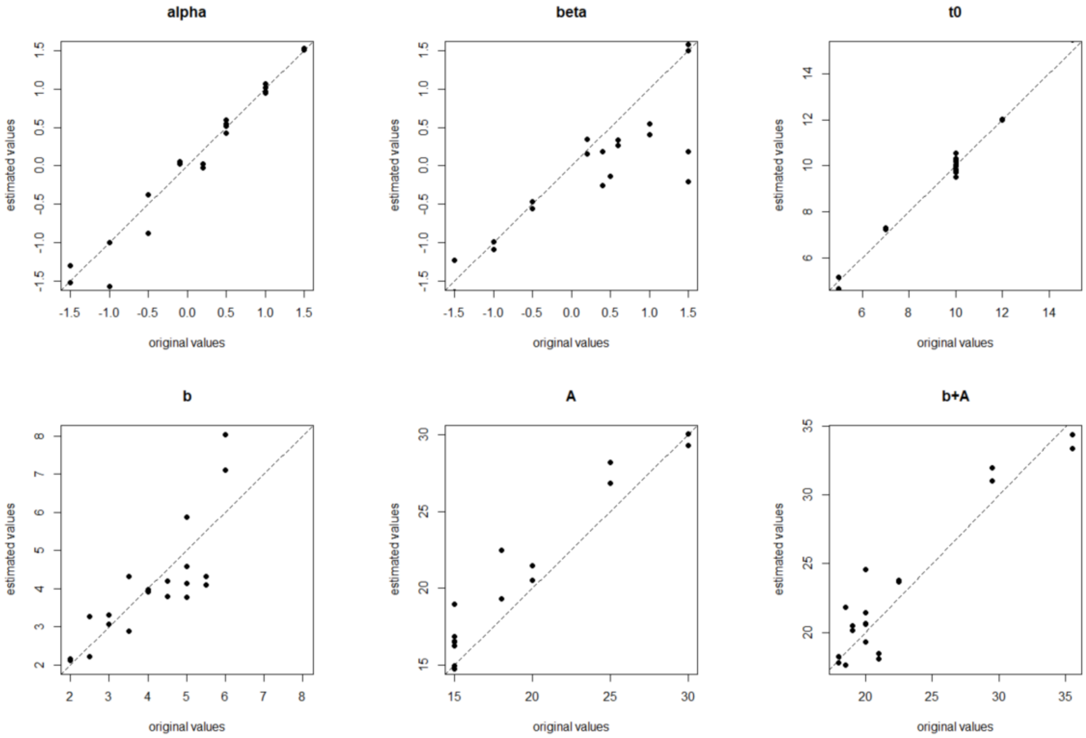


Figure 34. Original and estimated parameters values from the IF model for threshold, starting point, α, β, and non-decision time.

5.1.2.2. Two-stage models

Two different structures of the two-stage MLBA models, as described in section 3.2.4, were investigated for this research. The two-stage models, one using the CPT based equations and the other using the IF based equations, described the decision to switch tasks, and which task to switch to, as a serial process that used task type as a criterion to reduce the number of response options before making the final decision. The model investigation used 5 response choices and 1000 trials for all versions of the two-stage models, and let all drift rate parameter estimates be either negative or positive. The initial structure has 9 total parameters: b_1, b_2, A, α_1, α_2, γ_1, γ_2, t_0, and t_s for the CPT version and b_1, b_2, A, α_1, α_2, β_1, β_2, t_0, and t_s for the IF version. Table 15 summarizes the comparisons of the initial two-stage structure, using RMSE to quantify how the estimated parameter values compared to the generating values for each of the model configurations, the K-L divergence to measure the similarity of the choice distributions, and the K-S test to measure the similarity of the response time distributions. Figure 81 in Appendix B plots the trace plots of the samples from each initial structure two-stage model, along with the generating value, for each parameter.

Letting all the parameters vary within the CPT version of the model resulted in estimated parameters that did not equal the generating values for all the parameters, as shown in Table 15. For the IF version of the model the estimated values of the starting point and α from the first stage were close to the values used to generate the data (i.e., the generating values), but the other parameters were not. The first stage threshold, non-decision time, and stage delay parameter approached zero while the second stage threshold was much larger than the generating value. Additionally, the models produced correlated γ values for the CPT version (r_{CPT} = -0.20, n=2) and correlated β values for the IF version (r_{IF} = -0.5, n=2); these parameters were fixed to the generating values and the models are run again.

Letting the other 7 parameters vary (MT3) resulted in many estimated parameters that were not equal (or even close to equal) to the generating values for the CPT version; only A was close to the generating value. The first stage threshold was much smaller than the generating value while the second stage threshold was much higher. The non-decision time and the stage delay parameter were also much smaller than the generating values. Posterior predictions using the estimated parameters gave a choice distribution that visually was the same as the original distribution (Figure 35a), but response time distributions that predicted longer response times (median RT_{MT3} = 23.0 sec) than the original data (median RT = 13.9 sec). The K-L divergence is smaller than the baseline median value and the K-S test statistic p-values are 0. These metrics indicate that the estimated parameters produce the same choice distribution as the generating parameters, but a different response time distribution.

Fixing β_1 and β_2 as constant values and letting the other 7 parameters vary, the IF two-stage model (MT4) estimated parameters that also were far from equal to the generating values for any of the parameters. Again, the first stage threshold, non-decision time, and stage delay parameter approached zero while the second stage threshold was much larger. Posterior predictions using the estimated parameters from this model gave a choice distribution that was visually different from the original distribution (Figure 35b) as well as response time distributions that predicted longer response times (median RT_{MT4} = 50.4 sec) than the original data (median RT = 30.4 sec). The K-L divergence was larger than the baseline median value and the K-S test statistic p-values were 0. These metrics indicated that the estimated parameters produced both different choice and response time distributions than the generating parameters.

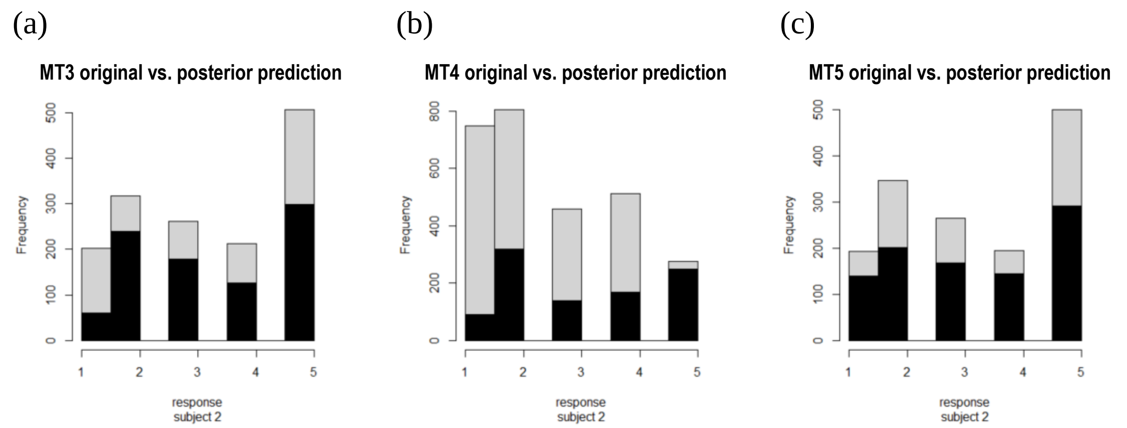


Figure 35. Choice distributions using original values vs. posterior predictions for (a) MT3, (b) MT4, and (c) MT5. Original data in black, posterior predictions in grey.

The model was further reduced to only let b_1, b_2, A and t_0 vary (MT5). This was only run once, using the CPT equations for drift rate to generate data, since none of the drift rate parameters varied in the model. The second stage threshold was large, like the previous versions, but the first stage threshold was only a little below the generating value. The starting point was smaller than the generating value and non-decision time was larger. Like the previous CPT model, posterior predictions using the estimated parameters gave a choice distribution that visually was the same as the original distribution (Figure 35c), but response time distributions that predicted longer response times (median $RT_{MT5} = 58.5$ sec) than the original data (median RT = 44.0 sec). The K-L divergence was smaller than the baseline median value and the K-S test statistic p-values are 0. These metrics indicated that the estimated parameters produced the same choice distribution as the generating parameters, but a different response time distribution. The results from all attempts to estimate parameters that recover the generating values indicated that while the models fit the response data reasonably well, the modeled response times were longer than the original data. Rather than continue to reduce the number of parameters with this structure, the two-stage model structure was revised to include only a single threshold and to use an overall drift rate equation in the log likelihood function, as described in section 3.2.4.

Table 15. PMmG estimated parameter metrics for two-stage model initial structure

Generating values	**MT1**	**MT2**	**MT3**	**MT4**	**MT5**
$b_1 = 2$	1.7	2.0	2.0	2.0	0.8
$b_2 = 3$	2.3	18.5	17.8	19.9	47.7
A =	40.3	1.0	0.9	26.2	45.4
α_1 =	1.3	0.2	0.2	1.1	-
α_2 =	0.7	0.4	0.5	0.8	-
γ_1 =	3.5	-	-	-	-
γ_2 =	2.3	-	-	-	-
β_1 =	-	2.4	-	-	-
β_2 =	-	1.9	-	-	-
$t_0 = 5$	4.9	5.0	5.0	5.0	4.4
$t_s = 10$	5.9	10.0	5.0	10.0	-
K-L divergence (mean ± se)	-	0.204 ± 0.12	0.019 ± 0.01	0.166 ± 0.01	0.004 ± 0.0003
K-S test statistic (mean ± se)	-	0.50 ± 0.002	0.51 ± 0.005	0.39 ± 0.03	0.29 ± 0.01
K-S test statistic p-values (mean ± se)	-	0	0	0	0

MT1: CPT equations, b_1, b_2, A, α_1, α_2, γ_1, γ_2, t_0, and t_s vary
MT2: IF equations, b_1, b_2, A, α_1, α_2, β_1, β_2, t_0, and t_s vary
MT3: CPT equations, b_1, b_2, A, α_1, α_2, t_0, and t_s vary
MT4: IF equations: b_1, b_2, A, α_1, α_2, t_0, and t_s vary
MT5: CPT equations, b_1, b_2, A and t_0 vary

The revised two-stage model structure had 8 total parameters: *b, A*, α_1, α_2, γ_1, γ_2, t_0, and t_s for the CPT version and *b, A*, α_1, α_2, β_1, β_2, t_0, and t_s for the IF version. Table 16 summarizes the comparisons of the revised two-stage structure, using RMSE to quantify how the estimated parameter values compared to the generating values for each of the model configurations, the K-L divergence to measure the similarity of the choice distributions, and the K-S test to measure the similarity of the response time distributions. Figure 83 in Appendix B plots the posterior distribution of the samples from each revised structure two-stage model, along with the generating value, for each parameter. By allowing anywhere from 1 to 8 of the variables to be free, considering the different possible combinations for each number of free variables, gave a total of 255 combinations of free variables that could be investigated. Instead of investigating all the combinations, this research first let all 8 parameters vary, allowing the drift rate to be less than zero and restricting the other parameters to be positive values. When running the PMwG sampler on the CPT model where all the parameters vary (MT6) the sampling phase continued to select new values for the estimated parameters for approximately the first 2000 samples in the sampling phase (Figure 82 in Appendix B). Starting at about sample 4000 the estimated values were consistent for each parameter. The portion of the sampling phase where the samples continued to vary was not used in determining the estimated parameters for the CPT version of the model; the last 1500 samples were used. The IF version of the model (MT7) also had some bad samples at the start of the sampling phase so the estimated parameter values were determined using the last 2000 samples. The resulting parameter estimates had large RMSE compared to the generating values. Both the CPT version of the model (MT6) and the IF version (MT7) had small K-L divergence and a small K-S test statistic with an associated p-value greater than 0.05. The metrics and a visual assessment (Figure 36 and Figure 37) confirmed that the posterior estimates produce choice and response time distributions that are the sufficiently similar to the original distribution.

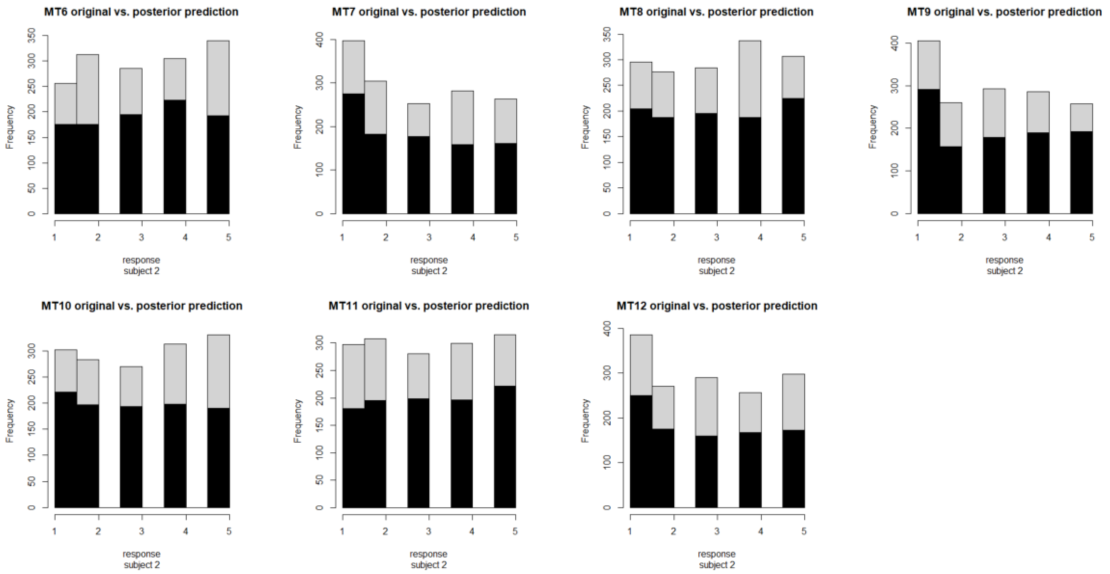


Figure 36. Choice distributions using original values vs. posterior predictions for MT6, MT7, MT8, MT9, MT10, MT11, and MT12. Original data in black, posterior predictions in grey.

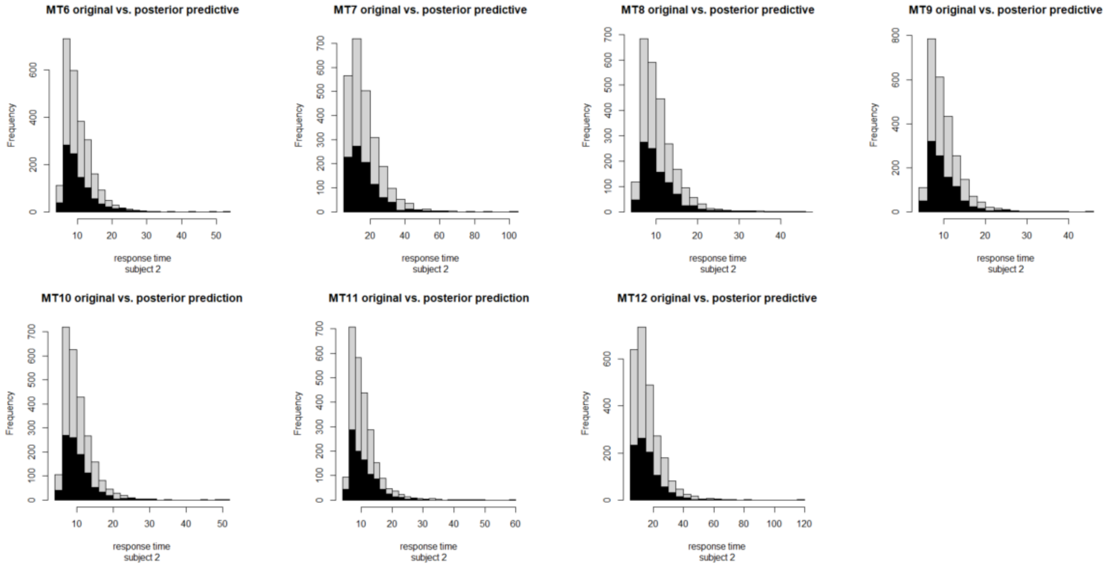


Figure 37. Response time distributions using original values vs. posterior predictions for MT6, MT7, MT8, MT9, MT10, MT11, and MT12. Original data in black, posterior predictions in grey.

Next, the model was restricted to only allow the threshold, starting point maximum value, and non-decision time (MT8 and MT9). These parameters were all recovered in the first revised models (MT6 and MT7) so they were chosen as a reasonable set of free parameters that are recoverable. A visual assessment of the posterior samples (Figure 83 in Appendix B) along with the RMSE values in Table 16 indicated that the estimated parameters were equal to the generating values. However, this configuration eliminated any parameters related to task attributes, in the drift rate equations, and to the number of stages. There was no difference in the structure of the model between the CPT and IF versions of the model, and the results provided the same estimated parameters regardless of whether the CPT equations or the IF equations were used to generate the data. Both versions were run and the results confirmed that both produce the same output. This configuration did not provide any information to address any of the research questions. Both the CPT version of the model (MT8) and the IF version (MT9) have small RMSE, small K-L divergence and small K-S test statistic with a p-value greater than 0.05 for that statistic. The metrics and a visual assessment (Figure 36 and Figure 37) confirmed that the posterior estimates produced choice and response time distributions that were sufficiently similar to the original distribution, and that the estimated parameters were also sufficiently similar to the generating values. It was of some concern that such a constrained model was needed to recover the generating values from the data, but the research questions focus on model comparisons that do not rely on the values of the estimated parameters, so less constrained models that contain more parameters of interest were still able to provide metrics to compare the models and test the hypotheses.

Several other combinations of more than 3 but less than all 8 parameters were run; all recovered both the choice and response distributions, but only some of the generating parameters were recovered for any configuration. Three configurations were looked at in more detail: b, A, t_0 and t_s as free parameters (MT10); $\alpha_{1,}$ $\alpha_{2,}$ $\gamma_{1,}$ and γ_2 as free parameters using the CPT drift rate equations (MT11); and $\alpha_1, \alpha_2, \beta_1,$ and β_2 as free parameters using the IF drift rate equations (MT12). Like MT8 and M9, the first configuration (MT10) also did not allow any of the drift rate parameters to vary, again eliminating any difference between the CPT and IF versions of the model. However, this configuration did include the stage delay parameter which was related to the two-stage process. This configuration recovered some of the generating values, but has RMSEs larger than MT8 or MT9. One limitation of varying only *b*, *A*, and t_0 or even *b*, *A*, t_0 and t_s was that none of the drift rate parameters were estimated so the other two configurations included only α_1, α_2, γ_1, and γ_2 for the CPT version (MT11) and α_1, α_2, β_1, and β_2 for the IF version (MT12). This again resulted in the distributions being recovered but not the parameters. The parameter RMSEs for MT10, MT11 and MT12 were approximately the same as the parameter RMSEs for MT6 and MT7. Including any parameters related to the drift rate or the two-stage structure resulted in approximately the same amount of error in the parameter estimates.

It did not make sense to run a parameter recovery simulation for the two-stage model. The only version that recovered all the parameters was the version that let only threshold, starting point and non-decision time vary, which were all part of the traditional LBA model and have been shown previously (Visser & Poessé, 2017) to recover their generating values given sufficient data. All other configurations of free parameters investigated created too much flexibility in the model to estimate a single true value for each of the parameters. All versions that include any parameters other than *b*, *A*, and t_0 resulted in approximately the same amount of error in the parameter estimates, which supported running models that included all 8 parameters on the simulated and real data. The parameter estimates could not be used in the counterfactual models, but the models were used to test the hypotheses relating to what cognitive mechanism best described the decision-making process.

Table 16. PMmG estimated parameter metrics for two-stage model revised structure

Generating values	**MT6**	**MT7**	**MT8**	**MT9**	**MT10 (CPT / IF)**	**MT11**	**MT12**
b = 1	0.3	0.1	0.07	0.2	0.3 / 0.3	-	-
A = 20	0.9	1.6	0.7	0.3	1.5 / 1.7	-	-
$\alpha_1 = 2$	3.1	0.7	-	-	-	0.3	0.2
$\alpha_2 = 1$	0.3	2.8	-	-	-	1.5	0.6
$\gamma_1 = 0.8$	0.5	-	-	-	-	0.9	-
$\gamma_2 = 0.2$	0.1	-	-	-	-	0.6	-
$\beta_1 = 0.2$	-	1.0	-	-	-	-	0.1
$\beta_2 = 0.8$	-	1.4	-	-	-	-	0.7
$t_0 = 5$	0.1	0.07	0.05	0.07	0.1 / 0.1	-	-
$t_s = 10$	2.9	28	-	-	2.7 / 2.3	-	-
K-L divergence (mean ± se)	0.002 ± 0.0008	0.003 ± 0.0009	0.002 ± 0.001	0.002 ± 0.0008	0.002 ± 8e-6 0.002 ± 0.0007	0.002 ± 0.0002	0.002 ± 0.0002
K-S test statistic (mean ± se)	0.03 ± 0.001	0.02 ± 0.002	0.02 ± 0.0003	0.03 ± 0.003	0.03 ± 0.006 0.02 ± 1e-6	0.03 ± 0.005	0.02 ± 0.002
K-S test statistic p-values (mean ± se)	0.59 ± 0.06	0.85 ± 0.06	0.98 ± 0.004	0.45 ± 0.11	0.58 ± 0.28 0.88 ± 0.01	0.40 ± 0.20	0.77 ± 0.10

MT6: CPT equations, b, A, α_1, α_2, γ_1, γ_2, t_0, and t_s
MT7: IF equations, b, A, α_1, α_2, β_1, β_2, t_0, and t_s vary
MT8: CPT equations, b, A, and t_0 vary
MT9: IF equations: b, A, and t_0 vary
MT10: CPT equations, b_1, b_2, A and t_0 vary
MT11: CPT equations, α_1, α_2, γ_1, and γ_2 vary
MT12: IF equations, α_1, α_2, β_1, and β_2 vary

5.2. Simulated Data Model Comparison

Simulated data, with a larger number of trials for each participant, was generated to provide a second dataset for model comparison. The resulting estimated parameters from the cognitive process models were not used to explain behavior nor in the counterfactual models. This analysis was intended to mitigate the limitations of using only the Project RED data.

5.2.1. Maximum Likelihood Estimation

Even though the MLE method did not recover the original parameter values for many of the parameters, using both the CPT and IF versions of the model, the method was used to find the best fit parameters for a simple set of simulated data, as described in section 3.2.6, and the log likelihood values from each configuration (i.e., CAG, CA, CG, IAB, IA, and IB) were used to determine the BICs and Bayes Factors. A baseline model was also run that did not use either drift rate equation, and instead included the drift rate mean as a parameter. While the parameter values themselves cannot be evaluated, the BICs and Bayes Factors were compared to determine which model configuration best represents the observed simulated data.

Figure 38a shows the BICs for each model, where the model using the CPT equation that lets both alpha and gamma vary best fit the data, when not including the baseline model. Bayes Factors, shown in Table 39 in Appendix B, comparing the CAG model pairwise to each of the other models, except the baseline model, strongly favored the CAG model over all models except the CA model for all subjects. When compared to the CA model, there was strong evidence in favor of the CAG model for a little over half the subjects, but only moderate or weak evidence for the rest. The CAG model assumed that subjects use the individual attribute values in deciding which task to perform and that both the value of the attributes and the weight of attention given to the attributes are considered. Figure 39 shows the choice distribution for the original data and data simulated using the best fit parameters from each model. Only response option 4 was slightly underestimated for the CAG model. Figure 38b shows the K-L divergence values for all the models; all values were below the baseline median of 0.133.

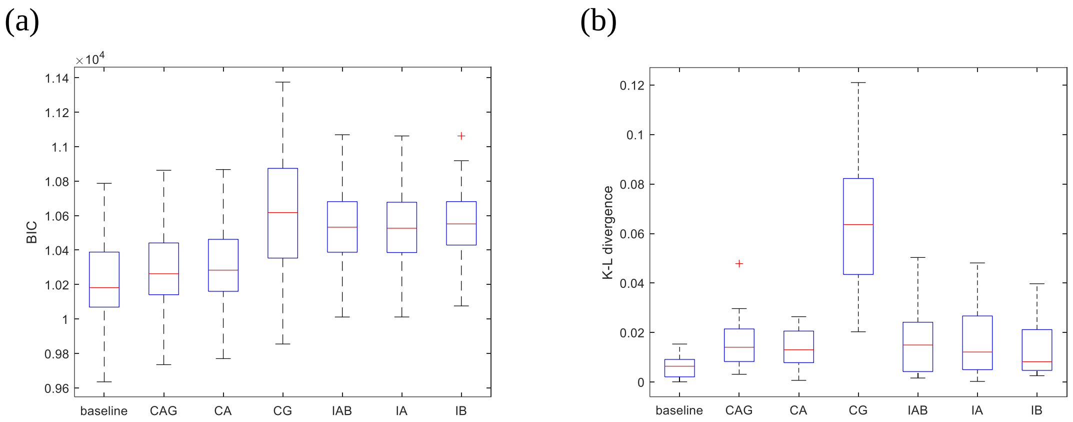


Figure 38. (a) BIC values and (b) K-L divergence values of each single-stage model fit to simple simulated data

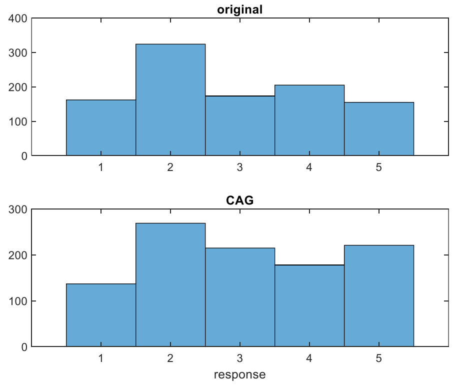


Figure 39. Choice distribution for simple simulated data for CAG (best-fit) model

5.2.2. Particle Metropolis within Gibbs

Particle Metropolis within Gibbs sampling was used to estimate parameters for both the single-stage and two-stage versions of the MLBA model using the simulated data. The results from the single-stage models are presented first, followed by the results from the two-stage models.

5.2.2.1. Single-stage Models

Single-stage model parameters were estimated for the second set of simulated data, described in section 3.2.6, using Particle Metropolis within Gibbs (PMwG) sampling, a

Bayesian hierarchical method of parameter estimation, for both the CPT and IF versions of the model. Both versions of the model estimated values for threshold, starting point, the drift rate parameters (i.e., α and γ for the CPT version, α and β for the IF version) and non-decision time. The drift rate scaling factors, c_d and c_m, were set constant to 0.1 and the initial drift rate, I_0, was set to zero, for both versions of the model.

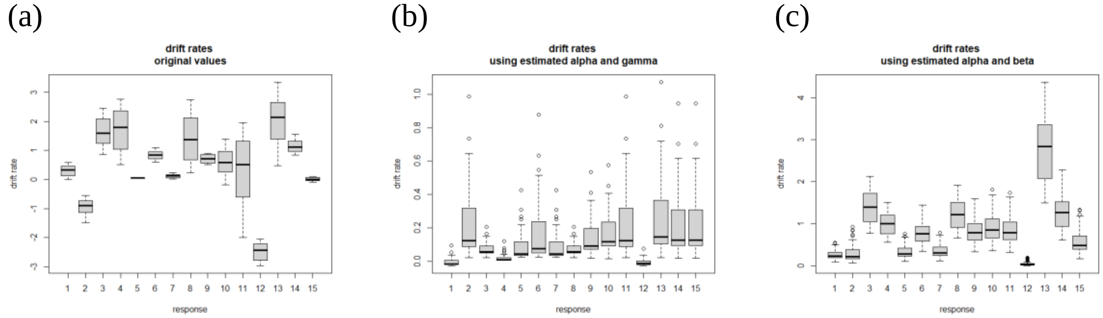


Figure 40. Drift rate using (a) original parameter values, (b) CPT model estimated parameters, and (c) IF model estimated parameters

Using the estimated values for attention weight (i.e., γ for CPT and β for IF) and subjective value (i.e., α), the resulting drift rate for each response option from each model was calculated as shown in Figure 40b and Figure 40c. Visually comparing these to the generating values for drift rate in Figure 40a, the IF version of the model had drift rates that better matched the pattern of original drift rates. In particular, the IF model captures the larger drift rate value for response 13 and more of the variation in drift rates across responses. However, drift rate only provided information about the rate of information processing to reach the threshold. It was not the only parameter of interest for task-switching and there were trade-offs between the LBA parameters. Additionally, participants had large variation in their individual responses, but Figure 40 includes the summed counts across all participants that does not show that variation.

Posterior predictive samples were generated for each version of the model to use for comparing the models. Figure 41 shows choice distributions of the original and (100 samples from the) posterior data for 3 participants from each version of the model. Figure 42 shows the response time distributions for the same participants. The estimated parameters for both versions of the model produced choice and response time distributions that were visually almost identical to the original distributions. The K-L divergence values (Figure 43a) favored the IF version of the model as producing posterior predictions that more closely matched the original data; a t-test of the values also shows that there is a difference in the mean K-L divergence values (t=10.0, p<0.001, n=48). The K-S test statistic is the same for both versions of the model ($d_{K\text{-}S,CPT} = 0.021 \pm 0.001$, $d_{K\text{-}S,IF} = 0.022 \pm 0.001$, t = -0.8, p = 0.40, n=48). The associated p-values (Figure 43b) were above 0.05; the estimated parameter values from both versions of the model produced a response time distribution that was sufficiently similar to the distribution of the original simulated data. Overall, visually and quantitatively, while both models produced posterior predictions of response time that matched the original data, the IF models predicted responses that better matched the original data.

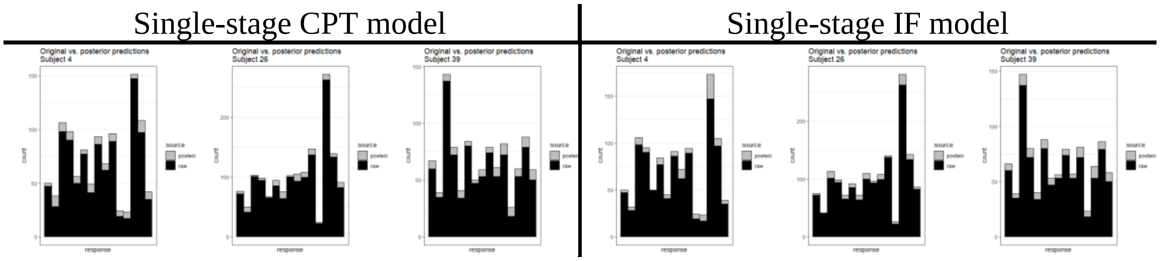


Figure 41. Choice distributions for simulated subjects 4, 26, and 39 using estimated parameters from the single-stage models

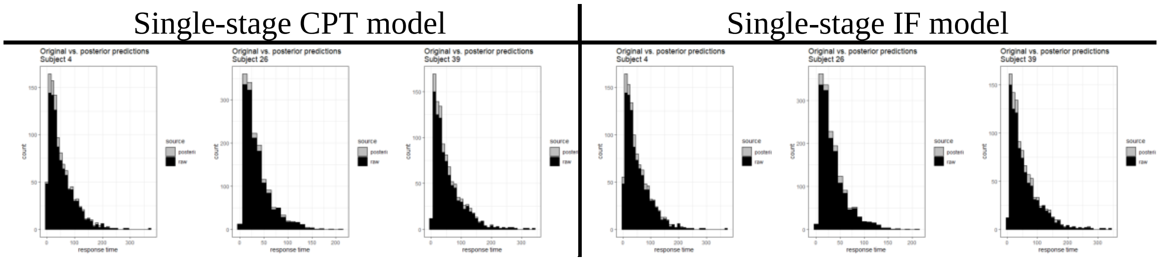


Figure 42. Response time distributions for simulated subjects 4, 26, and 39 using estimated parameters from the single-stage models

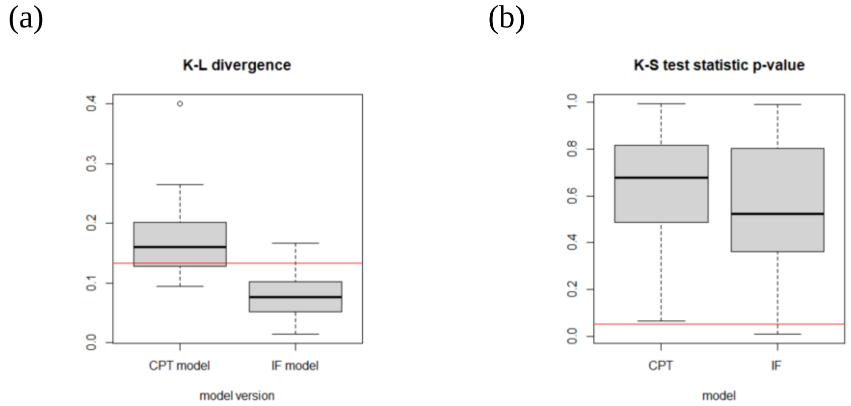


Figure 43. Metrics for single-stage models with simulated data (a) K-L divergence values. Red line is median value of baseline distribution (0.133). (b) K-S test statistic p-values. Red line is 0.05.

The trace plots from each of the models, and the associated boxplot of sampling phase parameter values, across all participants are available in Figure 84 and Figure 85 in Appendix B. The sampling phase was stable for each of the models. Figure 44 and Figure 45 provide plots of the posterior distribution from the sampling phase, for each parameter, using the CPT and IF versions of the model, respectively. All of the parameters had distributions where the mean value and the range of values varied by participant. None of the parameters had values that were consistent for all the participants. The individual differences in participants were included in all of the parameters indicating that individual variation is explained by a combination of the model parameters for the simulated data.

This result was consistent with how the data was generated, by varying all of the model parameters, and cannot be generalized to the Project RED data.

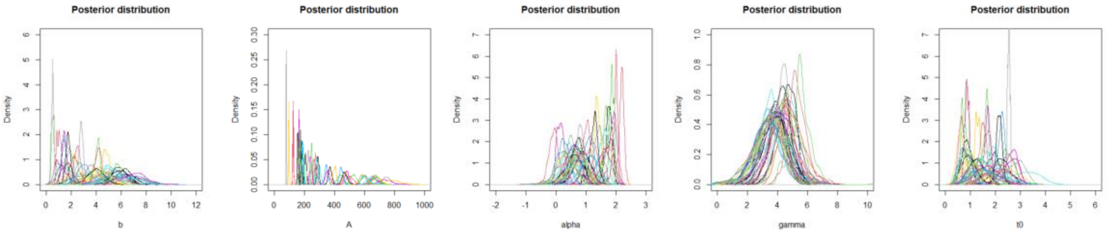


Figure 44. Posterior distributions of estimated parameters value from the CPT single-stage model

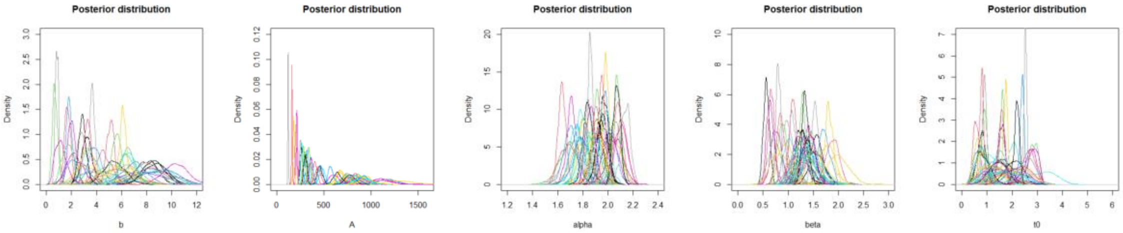


Figure 45. Posterior distributions of estimated parameters value from the IF single-stage model

The model using the information foraging theory drift rate equations better predicted the simulated data. This was determined using both the K-L divergence values and the WAIC values. K-L divergence values for the IF version of the model were less than the baseline (t=18.3, p<0.001, n=48) and the CPT version of the model (t=10.0, p<0.001, n=48). Table 17 showed that the difference in WAIC between the models was 17185 with a standard error of 2848, which was about one-fifth of the difference, so the models were easy to distinguish by expected out-of-sample accuracy. The WAIC values favored the single-stage IF model over the single-stage CPT model for the simulated dataset.

Table 17. WAIC – Simulated data, single-stage CPT vs. IF models

Model	**WAIC**	**SE_{WAIC}**	**dWAIC**	**SE_{dWAIC}**	**pWAIC**
SS IF	1.22e6	79784	-	-	120.8
SS CPT	1.23e6	81830	17185	2848	99.5

Looking at each of the parameters estimated by the single-stage IF model, only the starting point parameter was related to the switch rate (Figure 46). The results of a Bayesian linear regression, in Table 18, shows that there was strong evidence for a number of different models that use a linear combination of starting point parameter, α, β, threshold, the interaction of threshold and starting point, and the interaction of β and/or starting point to describe the switch rate values. The model with the highest Bayes Factor, compared to the intercept only model, included a linear combination of all these parameters (SB1). However, there was only weak evidence (BF = 1.1) that favored including β and the two interactions in addition to starting point, α, and threshold (SB1) over a model that only included starting point, α, and threshold (SB2). There was also weak evidence (BF = 1.2) against including β and the interaction of β and starting point in addition to starting point, α, and threshold (SB4) over a model that only included starting point, α, and threshold (SB2). Additionally, there was weak evidence (BF = 1.4) against including the interaction of threshold and starting point in addition to starting point, α, and threshold (SB3) over a model that only included starting point, α, and threshold (SB2). Finally, there was only weak evidence (BF = 1.4) for a model that included all these parameters (SB1) over a simple model with only starting point and α. From this, plus using the plots in Figure 46 that show switch rate as a function of each parameter individually, it appeared that both the starting point parameter and α are important factors for predicting switch rate in the simulated data. The other factors provided only a small improvement above this.

Table 18. Bayesian Linear Regression – SS IF model estimated parameters, simulated data

Model		**BF**
SB1	A + alpha + threshold + A:threshold + beta + A:beta	3.2e12
SB2	A + alpha + threshold	2.9e12
SB3	A + alpha + threshold + A:threshold	2.7e12
SB4	A + alpha + threshold + beta + A:beta	2.4e12
SB5	A + alpha	2.3e12

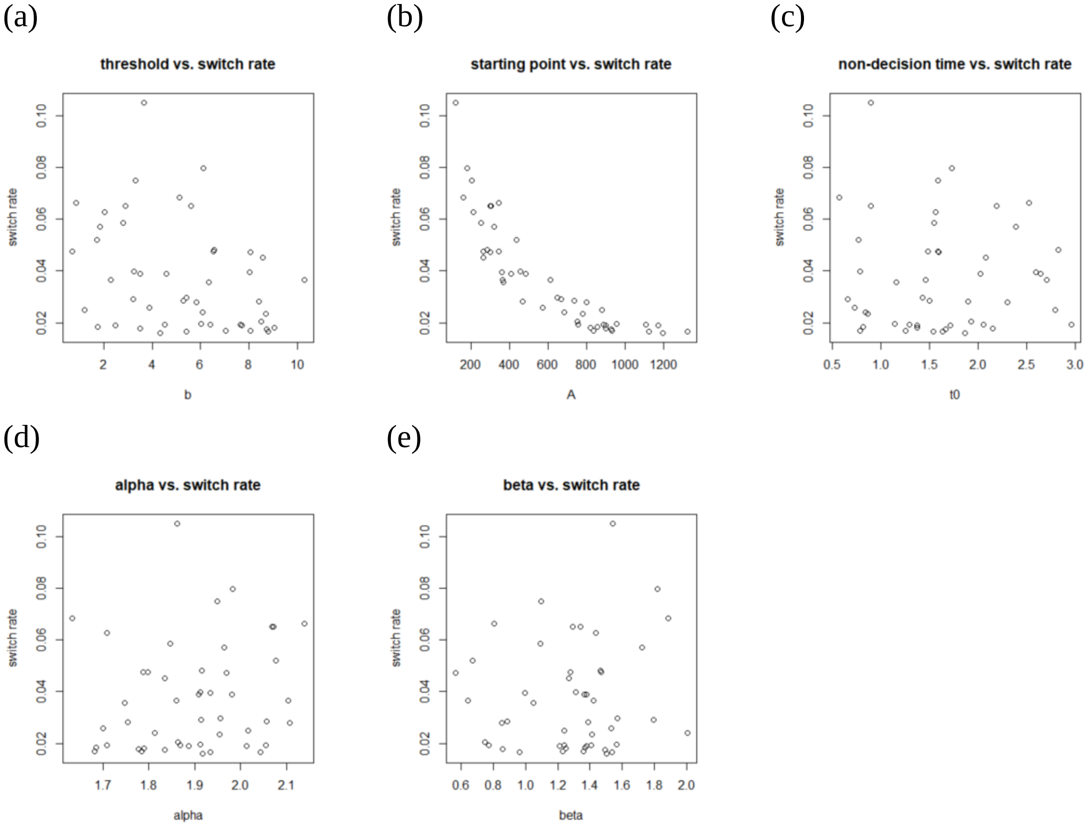


Figure 46. Parameter vs. switch rate plots using single-stage IF model estimated parameters for simulated data

5.2.2.2. Two-Stage Models

Parameters for the two-stage versions of the model were also estimated using PMwG methods, but these parameters cannot be compared directly and their values cannot be compared to the switch rates. The trace plots from each of the models, and the associated boxplot of sampling phase parameter values, across all participants are available in Figure 84 and Figure 87 in Appendix B. The sampling phase was stable for each of the models.

The log likelihood values were used to determine the WAICs and compare the models to determine which set of estimated parameter values best represented the observed simulated data.

Posterior predictive samples were generated for each version of the two-stage model. Figure 47 shows choice distributions of the original and posterior data for 3 participants from each version of the two-stage model; these are the same 3 participants as from the single-stage models. Figure 48 shows the response time distributions for the same participants. The estimated parameters for both versions of the model produced choice and response time distributions that were visually almost identical to the original distributions. The K-L divergence values (Figure 49a) favored the IF version of the model as producing posterior predictions that more closely matched the original data, which was also supported by a t-test of the values which shows there was a difference in the mean K-L divergence values (t=2.7, p=0.007, n=48). The K-S test statistic also favored the IF version of the model ($d_{k\text{-}S,CPT}$ = 0.026 ± 0.001, $d_{K\text{-}S,IF}$ = 0.021 ± 0.001, t = -2.7, p = 0.008, n=48) and the associated p-values (Figure 49b) were above 0.05. The estimated parameter values from the IF version of the model produced a response time distribution that was the same as the distribution of the original simulated data. Quantitatively, the IF models predicted responses and response times that better matched the original simulated data.

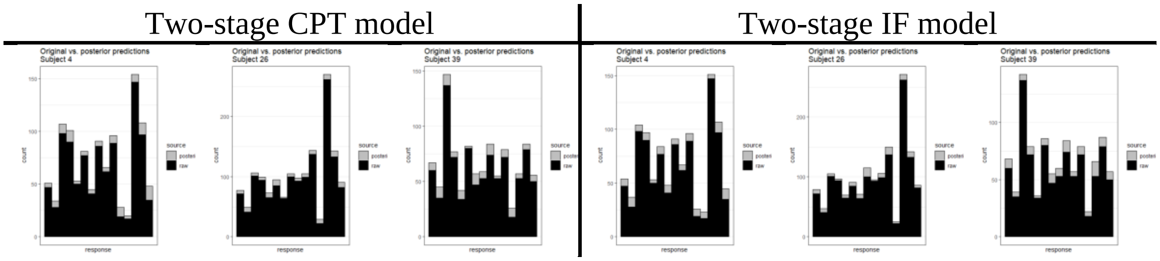


Figure 47. Choice distributions for simulated subjects 4, 26, and 39 using estimated parameters from the two-stage IF models

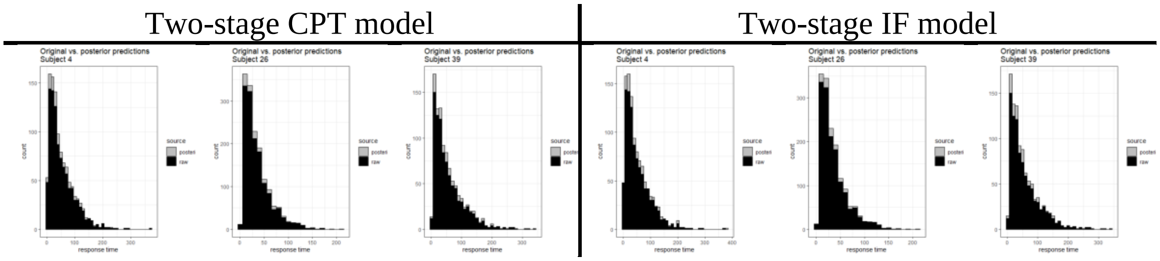


Figure 48. Response time distributions for simulated subjects 4, 26, and 39 using estimated parameters from the two-stage models

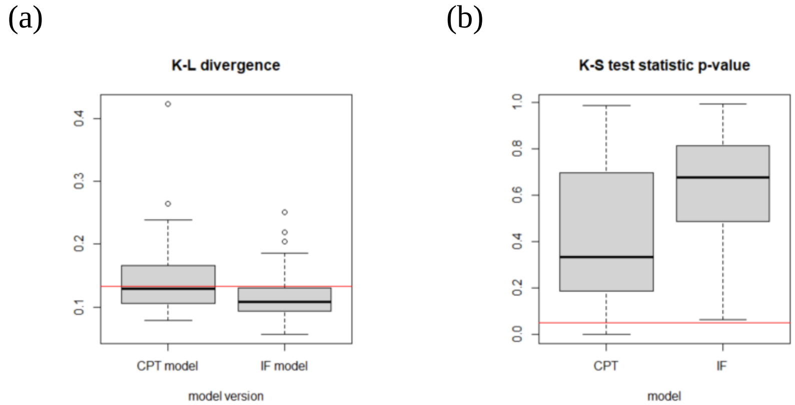


Figure 49. Metrics for two-stage models with simulated data (a) K-L divergence values. Red line is median value of baseline distribution (0.133). (b) K-S test statistic p-values. Red line is 0.05.

Table 19. WAIC – Simulated data, two-stage CPT vs. IF models

Model	**WAIC**	**SE$_{\text{WAIC}}$**	**dWAIC**	**SE$_{\text{dWAIC}}$**	**pWAIC**
TS IF	1.23e6	8.14e4	-	-	1284
TS CPT	1.25e6	8.25e4	16378	4251	7676

The two-stage model using the information foraging theory drift rate equations predicted the simulated data better than the cumulative prospect theory version. This was determined using both the K-L divergence values and the WAIC values. K-L divergence values for the IF version of the model were less than the CPT version of the model (t=2.7, p=0.007, n=48). The WAIC values favored the two-stage IF model over the two-stage CPT model for the simulated dataset. The difference in WAIC between the models was 16378 with a standard error of 4251, as shown in Table 19, which was about one-quarter of the difference, giving a 95% confidence interval of the difference between the models of [8046, 24710], so models were easy to distinguish by expected out-of-sample accuracy. The relatively small amount of error (i.e., uncertainty) in the WAIC provided confidence that the two models are different.

5.2.2.3. Comparing All Four Models

When considering only the information foraging theory versions, the WAIC values favored the single-stage version, as shown in Table 20. The same was true when only considering the cumulative prospect theory models; the single-stage version was favored over the two-stage model, as shown in Table 21. When considering all four MLBA models together, as shown in Table 22, both single-stage models were favored over the two-stage models, and both information foraging theory models were favored over the cumulative prospect theory models. For the simulated data, which had a large number of trials for each participant, the results supported the hypothesis for research question 2; the model that uses the value and attention weight of the task as a whole was preferred over the model that considers the value and weight of attention given to each attribute. The results did not support the hypothesis for research question 3; the model that assumed the decision process was a single stage where all tasks are considered simultaneously was preferred over the model that assumed a down-select process. The two-stage model used to estimate parameters for the simulated data was more complicated, with more parameters and more flexibility, than the single-stage model, which may not accurately represent a serial decision process. Chapter 8 discusses opportunities to better model and test a serial decision process.

Table 20. WAIC – Simulated data, single-stage vs. two-stage IF models

Model	**WAIC**	**SE_{WAIC}**	**dWAIC**	**SE_{dWAIC}**	**pWAIC**
SS IF	1.22e6	7.98e4	-	-	120.8
TS IF	1.23e6	8.14e4	12317	2269	1284

Table 21. WAIC – Simulated data, single-stage vs. two-stage CPT models

Model	**WAIC**	**SE_{WAIC}**	**dWAIC**	**SE_{dWAIC}**	**pWAIC**
SS CPT	1.23e6	8.18e4	-	-	99.5
TS CPT	1.25e6	8.25e4	11050	4237	7676

Table 22. WAIC – Simulated data, all four MLBA models

Model	**WAIC**	**SE_{WAIC}**	**dWAIC**	**SE_{dWAIC}**	**pWAIC**
SS IF	1.22e6	7.98e4	-	-	120.8
TS IF	1.23e6	8.14e4	12317	2269	1284
SS CPT	1.23e6	8.18e4	17185	2848	99.5
TS CPT	1.25e6	8.25e4	28236	4735	7676

5.3. Project RED Data Model Comparison

The campaign 3 (C3) and campaign 4 (C4) Project RED data were fit to each model separately. There were two reasons for doing this. Primarily to reduce the run time for the computations, but also because each campaign had different conditions, such as number of sessions completed, so the parameters for each were estimated separately to give more insight into the specific campaign and if there were differences in the estimated parameters for the two campaigns.

5.3.1. Single-Stage Models

The single-stage versions of the CPT and IF models were initially run estimating seven parameters: b, A, α, γ/β, t_0, c_d and c_m. The models were then rerun using the group level estimates for c_d and c_m from the initial model as fixed values and estimating the remaining five parameters (b, A, α, γ/β, and t_0). The scaling factor values were consistent between the two datasets. The CPT model used the same c_d value for both sets of data (c_d

= 0.6 for campaign 3 and campaign 4 data), but different c_m values for each campaign (c_m = 4.0 for campaign 3 and c_m = 3.8 for campaign 4). The IF model used different c_d values for each campaign (c_d = 1.9 for campaign 3 and c_d = 0.9 for campaign 4) and the same c_m value for both sets of data (c_m = 0.9 for campaign 3 and campaign 4 data). Only the models with five parameters were compared to determine the model that best described each dataset since including the scaling factors caused problems with the log likelihood values, as described in section 3.2.3.

Using the estimated drift rate parameters for attention weight and subjective value, the drift rate for each response option was calculated as shown in Figure 50. Visually comparing this to the aggregated responses across all participants for each campaign, neither version of the model had drift rates that matched the pattern of responses. The drift rates from the IF model included larger values for task options 3, 14 and 15, which all had higher response count, especially in campaign 4, but had smaller values for options 6, 7, 10 and 13, which also had higher response counts. The CPT models had higher drift rate values for most response options, but less differentiation in values between the responses. However, drift rate only provided information about the rate of information processing to reach the threshold. It was not the only parameter of interest for task-switching and there are trade-offs between the LBA parameters, so this was not used to evaluate model performance. Additionally, participants had large variation in their individual responses, but Figure 50 included the summed counts across all participants which did not show that variation.

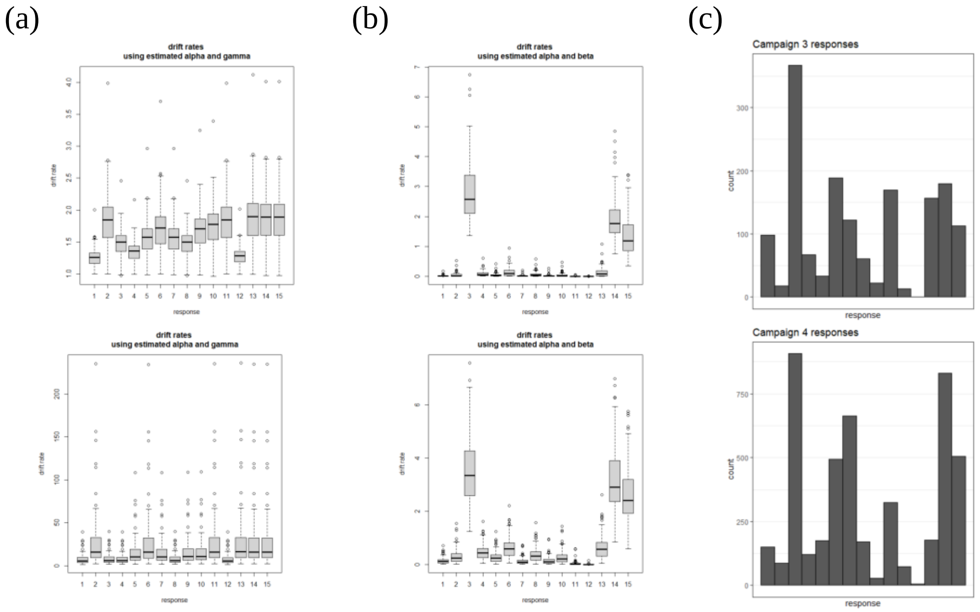


Figure 50. Drift rates for C3 data (top row) and C4 data (bottom row) calculated using the estimated parameters from (a) the single-stage CPT model (b) and the single-stage IF model. (c) Distribution of responses for each campaign.

The trace plots from each of the models, and the associated boxplot of sampling phase parameter values, across all participants are included in Appendix B, Figure 88 - Figure 91 and Figure 93 - Figure 95. The sampling phase was stable for each of the models. Figure 51 provides plots of the posterior distribution from the sampling phase, for each parameter, using each version of the model, for each campaign.

Looking at the estimated parameters found using the campaign 3 data, for the CPT model, most participants had similar threshold, γ and non-decision time values, while the starting point parameter and α varied between participants, suggesting that the individual variation in responses was mostly included in those two parameters for that model. For the IF model, again most participants have similar threshold and non-decision time values, but the α and β values as well as the starting point parameter varied between participants (Figure 51). The parameter for attention weight (γ in the CPT model and β in the IF model) changed behavior between the two models; that is, for the CPT model it was consistent between participants while for the IF model it was not. There are several differences between the models, including the equation for attention weight within the drift rate equation and the scale of the w_k and π_i parameters, so while it was possible to identify differences between the resulting parameter estimates the reason for the differences could not be determined.

There were a larger number of participants and resulting posterior samples drawn for the campaign 4 data. Overall, the campaign 4 values were within the same range as the campaign 3 parameter estimates. The estimated values for the non-decision time and threshold parameters found using the campaign 4 data were fairly constant across participants while individual variation in responses for campaign 4 was primarily included in starting point, subjective value, and attention weight for both the CPT and IF versions of the model. The starting point and subjective value parameter estimates varied to a degree that was consistant with the results from the campaign 3 data. However, the differences in the amount of individual variation in the attention weight parameter between the CPT and IF versions of the model found using the campaign 3 data was not repeated using the campaign 4 data. Along with the differences in the models mentioned above there were also intentional differences in the set-up of campaign 3 and campaign 4 so while the differences in the attention weight parameter were identified, the reasons for the differences could not be determined using the Project RED data. Further research is needed to explain the differences.

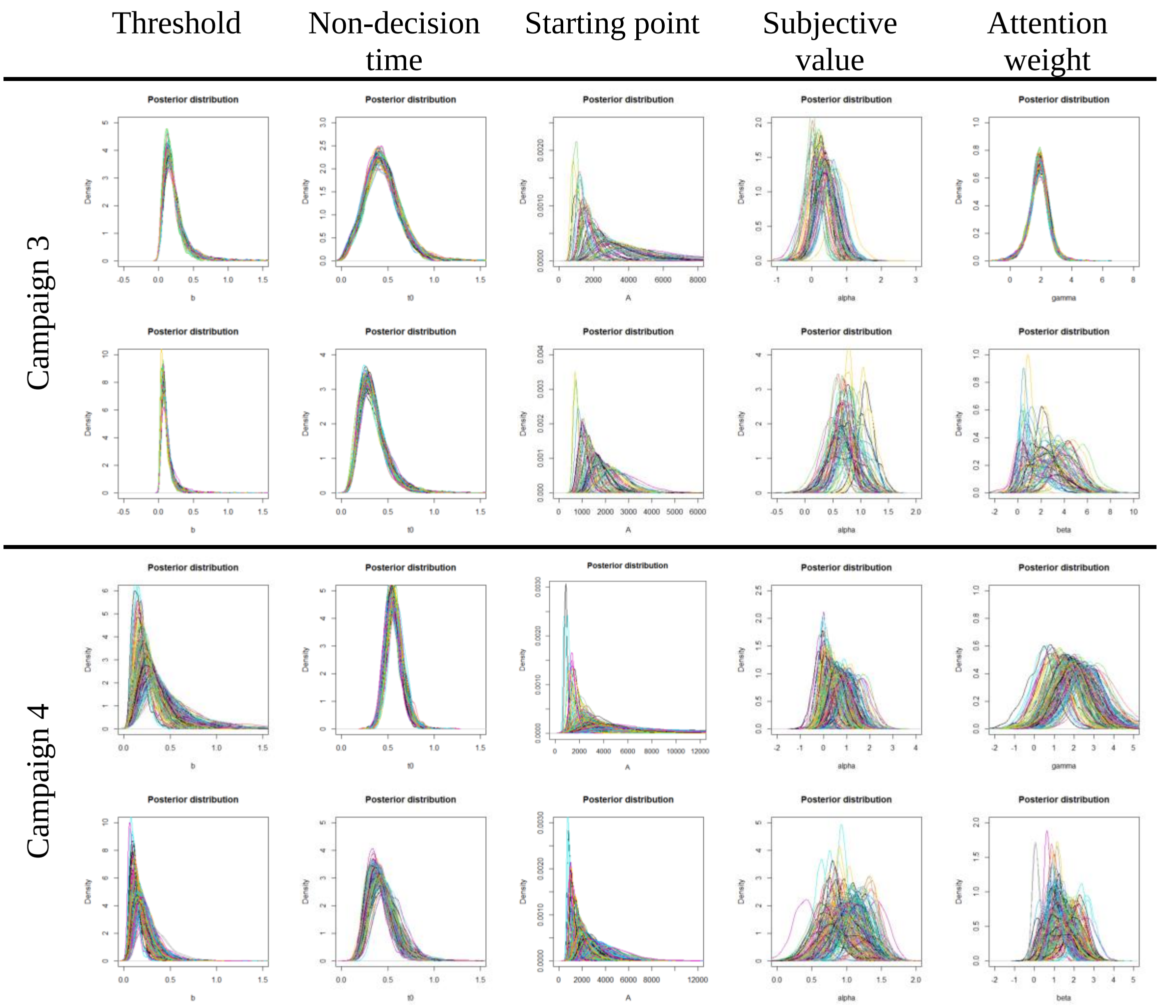


Figure 51. Posterior distributions of estimated parameters show the consistency of threshold and non-decision time values in both the CPT (first and third rows) and IF (second and fourth rows) versions of the single-stage model, the variation of starting point parameter and α by participant for both models, and that the parameter for attention weight (γ in the CPT model and β in the IF model) changes behavior between the two models.

Posterior predictive samples were generated for each version of the model, for each campaign. Figure 52 shows examples of the choice distributions of the original and (100 samples from the) posterior data, for 3 participants in campaign 3, from each version of the model. Figure 54 shows the response time distributions for the same participants. Appendix B, Figure 96 and Figure 97, contains additional plots for participants from campaign 4 that demonstrated the same patterns. The models generated posterior predictions that participants selected all the response options, including those with zero actual responses. The posterior predictions also seemed to more closely match the original data for participants with a larger number of trials, which was expected since parameter estimates generally improve when more data is included. The K-L divergence values (Figure 53) appeared to favor the IF version of the model as producing posterior predictions that more closely matched the original data, which was also supported by a t-test of the values which shows that there was a difference in the mean K-L divergence values between the two models (t=-4.7, p<0.001, n=72 for C3; t=-4.8, p<0.001, n=168 for C4). Additionally, comparing these K-L values to the baseline (Figure 53c) placed the IF version of the model in the 60th percentile of the baseline while the CPT version was in the 74th (for C3) and the 51st vs the 64th (for C4). Visually and quantitatively, the IF models produced posterior predictions of responses that better matched the original data.

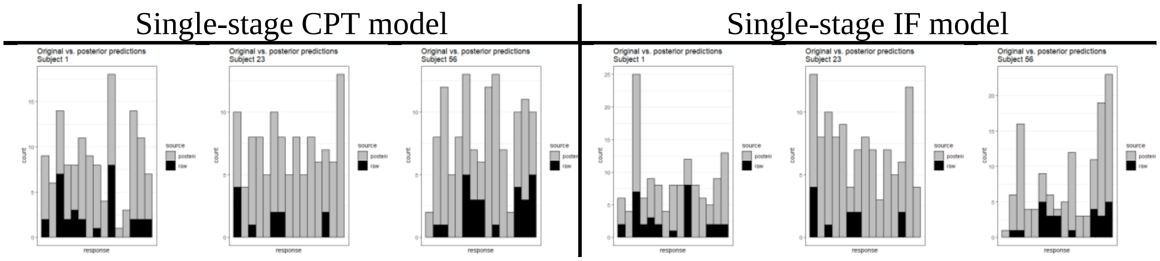


Figure 52. Choice distributions for subjects 1, 23, and 56 from campaign 3 showing the difference between the original (black) and posterior (grey) data single-stage models.

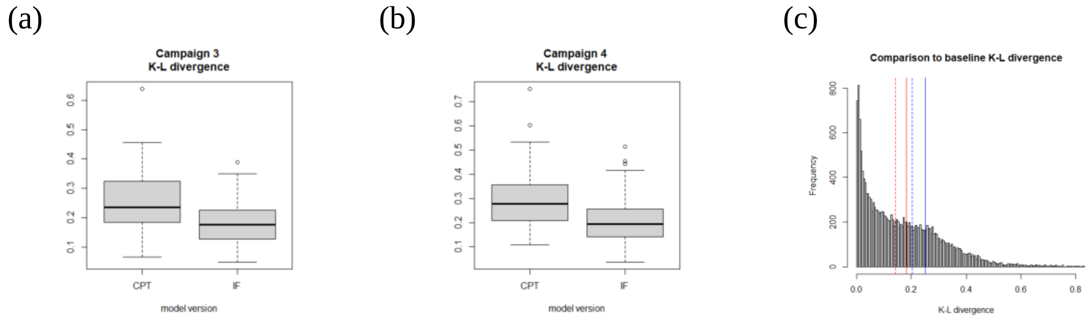


Figure 53. Comparison of K-L divergence values for single-stage CPT and IF models of (a) campaign 3 and (b) campaign 4 data. (c) Comparison to baseline. Red lines are median value for IF model and blue are median value for CPT model, solid line is C3 data and dashed line is C4 data.

Parameter estimates from both versions of the model capture the longer response times in both the campaign 3 and campaign 4 data (Figure 54). For campaign 3, the Kolmogorov-Smirnov (K-S) test statistic was slightly lower for the CPT version of the model (d_{CPT} = 0.20, d_{IF} = 0.21), but the difference was very small (t=-0.42, p=0.68, n=72) and the p-values of the K-S statistic (Figure 55a) for both models rejected the alternative hypothesis that the posterior predictions were drawn from a different distribution than the original data (and vice versa). For campaign 4, the K-S statistic for the IF model also was not statistically different than the CPT model (d_{CPT} = 0.17, d_{IF} = 0.18, t=-0.8, p=0.42, n=168) and the p-values of the K-S statistic for both versions of the model were mostly above 0.05 (Figure 55b). The posterior predictions were from a distribution that was sufficiently similar to the original data.

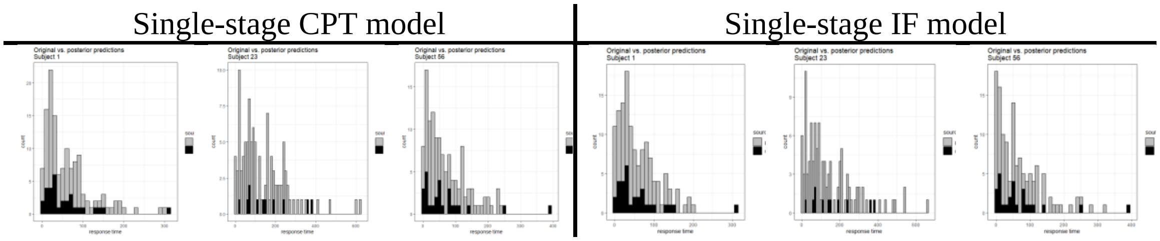


Figure 54. Response distributions for subjects 1, 23, and 56 from campaign 3 showing the difference between the original (black) and posterior (grey) data single-stage models.

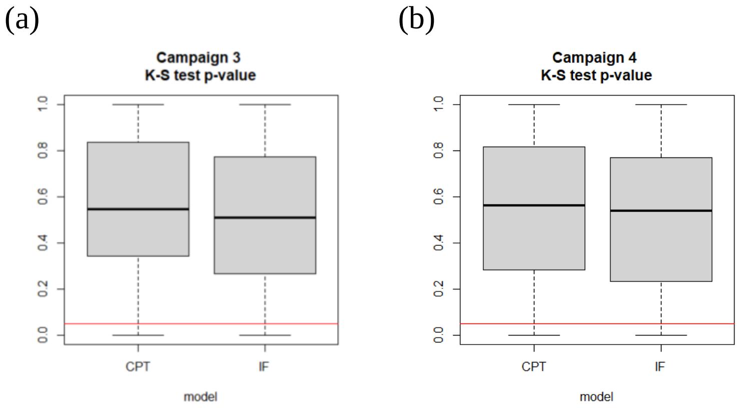


Figure 55. Comparison of associated p-values for single-stage CPT and IF models of (a) campaign 3 and (b) campaign 4 data. The red line is p=0.05.

The single-stage model using the information foraging theory drift rate equations better predicted the campaign 3 data. This was determined using both the K-L divergence values and the WAIC values. The K-L divergence values for the IF version of the model were less than the CPT version of the model (t=-4.7, p<0.001, n=72). The WAIC values favored the single-stage IF model over the single-stage CPT model for the campaign 3 dataset. The difference in WAIC between the models was 544 with a standard error of 97, shown in Table 23, which was about one-fifth of the difference, giving a 95% confidence interval of the difference between the models of [354,734], so the models were easy to distinguish by expected out-of-sample accuracy. The relatively small error (i.e., uncertainty) in the WAIC increased the confidence that the two models were different.

Table 23. WAIC – Campaign 3 data, single-stage CPT vs. IF models

Model	**WAIC**	**SE_{WAIC}**	**dWAIC**	**SE_{dWAIC}**	**pWAIC**
SS IF	25277	1201	-	-	96.1
SS CPT	25821	1267	544	96.7	58.2

The model using the information foraging theory drift rate equations also better predicted the campaign 4 data. This was again determined using both the K-L divergence values and the WAIC values. The K-L divergence values for the IF version of the model were less than the CPT version of the model (t=-4.8, p<0.001, n=168 for C4). The WAIC values favored the single-stage IF model over the single-stage CPT model for the campaign 4 dataset. The difference in WAIC between the models was 1810 with a standard error of 172, as shown in Table 24, which was about one-tenth of the difference, giving a 95% confidence interval of the difference between the models of [1473, 2147], so the models were easy to distinguish by expected out-of-sample accuracy. The relatively small error (i.e., uncertainty) in the WAIC increased the confidence that the two models are different.

Table 24. WAIC – Campaign 4 data, single-stage CPT vs. IF models

Model	**WAIC**	**SE_{WAIC}**	**dWAIC**	**SE_{dWAIC}**	**pWAIC**
SS IF	74778	2318	-	-	205.6
SS CPT	76588	2410	1810	172	151.8

There was not a strong relationship between any individual parameter from the IF model and switch rate (Figure 92 and Figure 98 in Appendix B). Using a Bayes factor analysis of each possible linear combination of parameters and the interactions between parameters to predict the switch rate resulted in the linear combination of starting point, threshold, the interaction of starting point and threshold, β, the interaction of starting point and β, and the interaction of threshold and β having the higher Bayes factor for the campaign 3 data. A model that used a linear combination of starting point, α, non-decision time and the interaction of α and non-decision time had the highest Bayes factor for the campaign 4 data. The campaign 3 data led to much larger Bayes factors than campaign 4 and these models included more interactions. However, for both campaigns, models that included starting point had the highest Bayes factors with strong evidence in favor of each model, respectively, over the intercept-only model. When interactions were not included, the linear combination of starting point, α, and non-decision time had the highest Bayes factor for both campaigns with moderate evidence for this model over the intercept-only model. However, there was moderate evidence (BF = 0.08) against this model compared to the best campaign 4 model (RB2) and strong evidence (BF < 0.001) against this model compared to the best campaign 3 model (RB7).

The best model for campaign 3 included all of the same parameters as the second-best model plus non-decision time. Comparing these two models (RB7 vs. RB8), there was weak evidence against including non-decision time (BF = 0.5). There was also weak evidence in favor of including the interaction of threshold and β (BF = 1.2) and no evidence (BF = 1) for or against including the interaction of threshold and non-decision time.

Four of the five models with the highest Bayes factors for the campaign 4 data include the linear combination of α, non-decision time, and the interaction of αand non-decision time along with one or two other parameters, as shown in Table 25. There was weak evidence in favor of the model that included starting point as the other parameter (RB2) over the models that include threshold (RB3) or β (RB4) as the other parameter. However, there was weak evidence against a model that included threshold (RB6) or β (RB5) in addition to starting point, α, non-decision time, and the interaction of α and non-decision time.

Table 25. Bayesian Linear Regression – SS IF model estimated parameters, real data

Model		Campaign 3 (BF)	Campaign 4 (BF)
RB1	A + alpha + t0	94.0	13.5
RB2	A + alpha + t0 + alpha:t0	50.1	168.7
RB3	alpha + threshold + t0 + alpha:t0		151.1
RB4	beta + alpha + t0 + alpha:t0		133.7
RB5	A + beta + alpha + t0 + alpha:t0	19.4	76.2
RB6	A + threshold + alpha + t0 + alpha:t0	15.8	67.5
RB7	A + threshold + A:threshold + beta + A:beta + threshold:beta	1760674	3.7
RB8	A + threshold + A:threshold + beta + A:beta + threshold:beta + t0	998271	
RB9	A + threshold + A:threshold + beta + A:beta + t0	836585	
RB10	A + threshold + A:threshold + beta + A:beta + t0 + threshold:t0	798396	2.0

The Bayesian generalized linear models to predict task-switching behavior for the counterfactual questions were built using the combined campaign 3 and campaign 4 data, but the model with the best linear combination of parameters from campaign 3 (RB7), in Table 25, had strong evidence against it comparing it to the best campaign 4 model (RB2), using the campaign 4 Bayes factors. Similarly, the model with the best linear combination of parameters from campaign 4 (RB2) had strong evidence against it comparing it to the best campaign 3 model (RB7), using the campaign 3 Bayes factors. There are two possible explanations for these differences. The first is that they are due to random variation in the data. The datasets were small leading to more noise in the data. The other possible explanation is that there is a systematic difference between campaign 3 and campaign 4 leading to different parameters being part of the best model. Because of these differences between the campaigns, only the individual parameter values were included in the counterfactual model. The individual parameters of starting point (A), threshold (b), attention weight (β), and non-decision time (t_0) were in most of the best models from both campaigns, and value (α) was in all the best campaign 3 models.

5.3.2 Two-Stage Models

Parameters for the two-stage versions of the model were also estimated using PMwG methods, but these parameters cannot be compared directly and their values cannot be compared to the switch rates. The log likelihood values were used to determine the WAICs and compare the models to determine which set of estimated parameter values best represented the observed simulated data. Both the CPT and IF versions of the model used the revised two-stage structure with all 8 parameters free. The drift rate scaling factors, c_d and c_m, were set constant to the same values that were used in the respective single-stage models (e.g., campaign 3 two-stage CPT model $c_d = 0.6$ and $c_m = 4.0$). The trace plots from each of the models, and the associated boxplot of sampling phase parameter values, across all participants are included in Appendix B, Figure 99 - Figure 102. The sampling phase was stable for each of the models.

Posterior predictive samples were generated for each version of the model, for each campaign. Figure 56 shows examples of the choice distributions of the original and (100 samples from the) posterior data, for 3 participants in campaign 3, from each version of the model. Figure 58 shows the response time distributions for the same participants. Appendix B, Figure 103 and Figure 104, contains additional plots for participants from campaign 4 that demonstrated the same patterns. The two-stage models also generated posterior predictions that participants selected all the response options, including those with zero actual responses. The K-L divergence values (Figure 57) appeared to favor the CPT version of the model as producing posterior predictions that more closely matched the original data, which was also supported by a t-test of the values shows that there is a difference in the mean K-L divergence values between the two models (t=-7.3, p<0.001, n=72 for C3; t=-

2.7, p=0.006, n=168 for C4). Visually and quantitatively, the CPT models produced posterior predictions (of responses) that better matched the original data. This contradicted the results found when comparing the single-stage versions of the models.

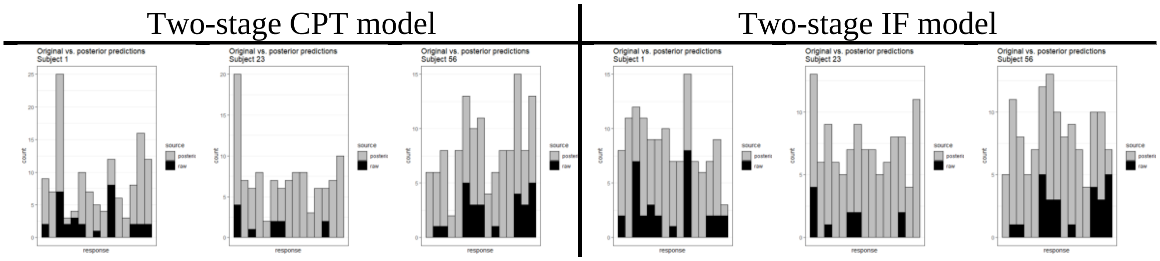


Figure 56. Choice distributions for subjects 1, 23, and 56 from campaign 3 showing the difference between the original (black) and posterior (grey) data for two-stage models.

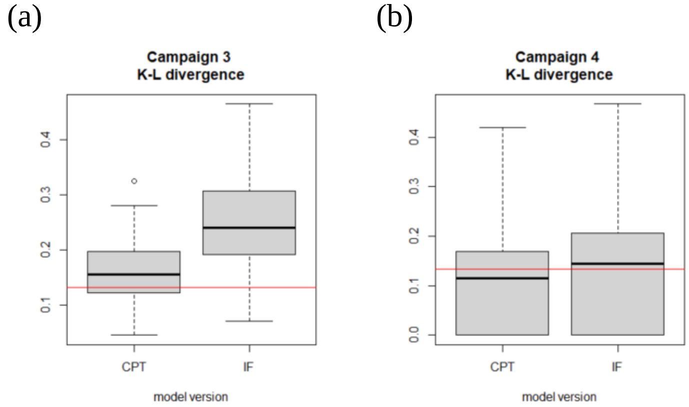


Figure 57. Comparison of K-L divergence values for two-stage CPT and IF models of (a) campaign 3 and (b) campaign 4 data. Red line is baseline median value (D_{KL}=0.133).

Parameter estimates from both versions of the model captured the longer response times in both the campaign 3 and campaign 4 data (Figure 58). For campaign 3, the Kolmogorov-Smirnov (K-S) test statistic was slightly lower for the IF version of the model (d_{CPT} = 0.203, d_{IF} = 0.200), but the difference was very small (t=-0.15, p=0.88, n=72) and the p-values of the K-S statistic (Figure 59) for both models did not reject the null hypothesis that the posterior predictions were drawn from the same distribution as the original data (and vice versa). For campaign 4, the K-S statistic for the IF model was also not statistically different than the CPT model (d_{CPT} = 0.181, d_{IF} = 0.176, t=-0.56, p=0.58, n=168) and both versions of the model had p-values of the K-S statistic (Figure 59) that supported the conclusion that the posteriors were from the same distribution as the original data.

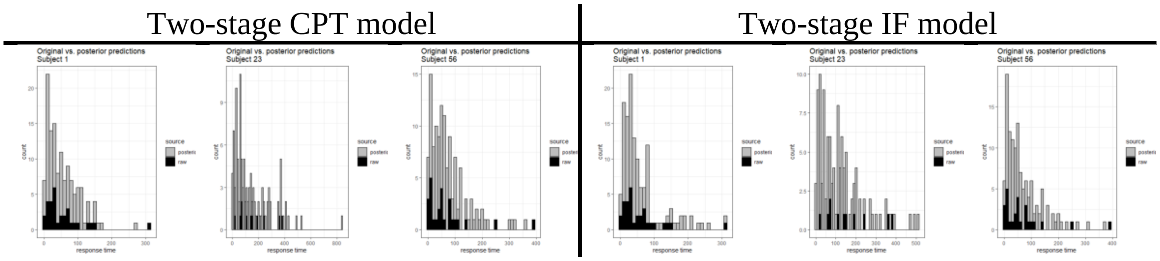


Figure 58. Response distributions for subjects 1, 23, and 56 from campaign 3 showing the difference between the original (black) and posterior (grey) data for two-stage models

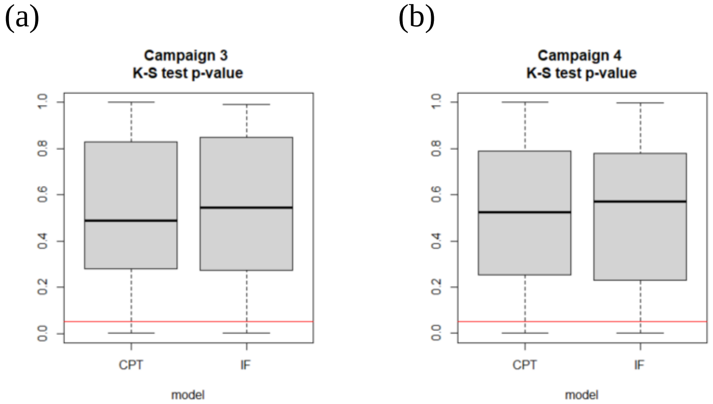


Figure 59. Comparison of associated p-values for two-stage CPT and IF models of (a) campaign 3 and (b) campaign 4 data. The red line is p=0.05.

The two-stage model using the cumulative prospect theory drift rate equations predicted the campaign 3 data slightly better than the information foraging theory version. This was determined using both the K-L divergence values and the WAIC values. The K-L divergence values for the CPT version of the model were less than the IF version of the model (t=-7.3, p<0.001, n=72). The WAIC values favored the two-stage CPT model over the two-stage IF model for the campaign 3 dataset. The difference in WAIC between the models was 471 with a standard error of 238, as shown in Table 26, which was about one-half of the difference, giving a 95% confidence interval of the difference between the models of [4.5,937], so models were somewhat easy to distinguish by expected out-of-sample accuracy. However, the larger error (i.e., uncertainty) in the WAIC reduced the confidence that the two models were different.

Table 26. WAIC – Campaign 3 data, two-stage CPT vs. IF models

Model	WAIC	SE$_{WAIC}$	dWAIC	SE$_{dWAIC}$	pWAIC
TS CPT	25374	1209	-	-	260.7
TS IF	25845	1338	471	238	207.3

The model using the cumulative prospect theory drift rate equations also better predicted the campaign 4 data. This was again determined using both the K-L divergence values and the WAIC values. The K-L divergence values for the CPT version of the model were less than the IF version of the model (t=-2.7, p=0.006, n=168 for C4). The WAIC values favored the two-stage CPT model over the two-stage IF model for the campaign 4 dataset. The difference in WAIC between the models was 1786 with a standard error of 173, as shown in Table 27, which was about one-tenth of the difference, giving a 95% confidence interval of the difference between the models of [1447,2125], so models were easy to distinguish by expected out-of-sample accuracy. The relatively small error (i.e., uncertainty) in the WAIC increased the confidence that the two models were different.

Table 27. WAIC – Campaign 4 data, two-stage CPT vs. IF models

Model	WAIC	SE$_{WAIC}$	dWAIC	SE$_{dWAIC}$	pWAIC
TS CPT	73875	2262	-	-	463
TS IF	75661	2365	1786	173	789

5.3.3. Comparing All Four Models

When considering only the information foraging theory versions, the WAIC values favored the single-stage version, as shown in Table 28 and Table 30. The opposite was true when only considering the cumulative prospect theory models; the two-stage version was favored over the single-stage model, as shown in Table 29 and Table 31. All four MLBA models were compared using the WAIC to see which model was favored for each dataset, in Table 32 and Table 33. For campaign 3, the single-stage information foraging theory model had the best WAIC values, but the difference between it and the two-stage cumulative prospect theory model was small and the standard error of the difference between the models was about sixth-tenths as large as the difference between the models, which reduced confidence that there was a difference between the two models. The difference between models was clearer for the campaign 4 dataset. The two-stage cumulative prospect theory model had the best WAIC values. This model assumed that participants consider each attribute of each task when making the decision of which task to perform next and when to switch to a new task. It also assumed that participants make this decision in two stages using the task type (i.e., individual, team or multi-team) as a filter in the first stage before making the final decision in the second stage.

Table 28. WAIC – Campaign 3 data, single-stage vs. two-stage IF models

Model	**WAIC**	**SE_{WAIC}**	**dWAIC**	**SE_{dWAIC}**	**pWAIC**
SS IF	25277	1201	-	-	96.1
TS IF	25845	1338	568	236	207.3

Table 29. WAIC – Campaign 3 data, single-stage vs. two-stage CPT models

Model	**WAIC**	**SE_{WAIC}**	**dWAIC**	**SE_{dWAIC}**	**pWAIC**
TS CPT	25374	1209	-	-	260.7
SS CPT	25821	1267	447	84	58.2

Table 30. WAIC – Campaign 4 data, single-stage vs. two-stage IF models

Model	**WAIC**	**SE_{WAIC}**	**dWAIC**	**SE_{dWAIC}**	**pWAIC**
SS IF	74778	2318	-	-	205.6
TS IF	75661	2365	883	119	789

Table 31. WAIC – Campaign 4 data, single-stage vs. two-stage CPT models

Model	**WAIC**	**SE_{WAIC}**	**dWAIC**	**SE_{dWAIC}**	**pWAIC**
TS CPT	73875	2262	-	-	463
SS CPT	76588	2410	2713	107	151.8

Table 32. WAIC – Campaign 3 data, all four MLBA models

Model	**WAIC**	**SE_{WAIC}**	**dWAIC**	**SE_{dWAIC}**	**pWAIC**
SS IF	25277	1201	-	-	96.1
TS CPT	25374	1209	97	59	260.7
SS CPT	25821	1267	544	97	58.2
TS IF	25845	1338	568	236	207.3

Table 33. WAIC – Campaign 4 data, all four MLBA models

Model	**WAIC**	**SE_{WAIC}**	**dWAIC**	**SE_{dWAIC}**	**pWAIC**
TS CPT	73875	2262	-	-	463
SS IF	74778	2318	903	135	205.6
TS IF	75661	2365	1786	173	789
SS CPT	76588	2410	2713	107	151.8

For the Project RED data, which had a small number of trials for each participant, the results did not clearly support the hypotheses for the second or third research questions. Addressing only whether a participant considered alternative or attribute level preferences, the results contradicted each other depending on whether a parallel process (i.e., single-stage model) or serial process (i.e., two-stage model) was used. Assuming a single-stage model led to favoring the model that assumed the task was considered as a whole, while assuming a two-stage model led to favoring the model that assumed each attribute of a task was considered. Addressing only the question of whether all tasks were considered simultaneously or not also produced results that contradicted each other depending on whether an attribute-level (i.e., CPT) or alternative-level (i.e., IF) version of the model was used. For the CPT version, the results favored a model that assumed two-stages, while for the IF version the results favored a model that assumed a single-stage. The small amount of data (along with possible errors in the structure of the two-stage model) increased the uncertainty in the parameter estimates and the ability of the model to accurately represent the data, which could contribute to the contradictory results.

5.4. Summary

The results from the cognitive process analysis supported the second hypothesis that a model based on alternative-level preferences was favored over a model based on attribute-level preferences. The results did not support the third hypothesis related to the processing architecture used. The results favored a single-stage model, which assumed that all tasks were considered simultaneously, over a two-stage model, which assumed a serial process where only a subset of tasks were considered for the final decision.

This research used four different versions of the MLBA to estimate parameters for three different datasets. The models were structured based on the features of interest in the research questions. Table 4, in section 3.2, summarized how the models align to the different parts of the research questions. The single-stage models assumed that all the response options were considered simultaneously while the two-stage models assumed that there was a down-select decision of which type of task to perform made prior to the final decision of which specific task to perform. The models were also structured to assume two different functions for the drift rate equation(s). The version using equations based on cumulative prospect theory assumed that participants considered each attribute for each response option while the version using equations based on information foraging theory assumed that participants considered each task as a whole. The single-stage models were shown to be identifiable while the two-stage models were non-identifiable.

Only the parameters from the single-stage models were used to explain individual differences in behavior. The results suggested that the individual variation in responses was mostly accounted for by the starting point parameter and the parameters related to the drift rate (i.e., the subjective value and the weight of attention). The drift rate parameters measured the attractiveness of each response option and the starting point parameter represented the bias towards switching or not switching for each participant. The starting point parameter was also non-linearly and inversely proportional to the switch rate. Participants with a higher starting point parameter had smaller switch rate.

WAIC was used to compare the two versions of the single-stage models for a simulated dataset, the campaign 3 data, and the campaign 4 data, separately. The two versions of the two-stage model were also compared using WAIC for the simulated dataset, the campaign 3 data, and the campaign 4 data, separately. Comparing only the single-stage models, the information foraging theory model was favored for the simulated, campaign 3, and campaign 4 data. Comparing only the two-stage models, the cumulative prospect theory model was favored for the campaign 3 and campaign 4 data, while the information foraging theory model was favored for the simulated data. This was determined using the WAIC values as well as the K-L divergence values. Comparing only the IF models, the WAIC values favored the single-stage model for the simulated, the campaign 3, and that campaign 4 data. Comparing only the CPT models, the two-stage model was favored for the campaign 3 and campaign 4 data, while the single-stage model was favored for the simulated data. The results from the simulated data, which had a large number of trials for each participant, were clearer in supporting a single-stage information foraging theory version of the MLBA model. Overall, the results supported the second hypothesis, but not the third hypothesis.

Only parameter estimates from single-stage models were used in the counterfactual prediction models as predictors of switch rate. When comparing only single-stage models, the IF version was favored so the individual-level estimated parameters from the single-stage IF version of the model were used in the counterfactual models. These results are discussed in the next section.

6. COUNTERFACTUAL MODEL RESULTS

To address the final research question, Bayesian generalized linear models (GLM) were used to make counterfactual predictions about expected task switching as well as individual, team, and multi-team task performance for the Project RED data. A different model was built and predicted responses were generated to address each counterfactual question (i.e., between-subject, within-subject, and other), as described in section 3.3. All models were built and validated using only ‘not withheld’ data, and these models were then used to make predictions for the ‘withheld’ data. Section 2.1 describes the differences between the ‘not withheld’ and ‘withheld’ data.

The models that included task switching as the dependent variable used a binomial distribution for the likelihood function to describe the relationship between the outcome and the predictors. This distribution was the most conservative distribution when each trial, which for this data is each second of the overall task, must result in a constant chance of either staying with the current task or switching to a new task. The distribution used two parameters to describe its shape: *n*, the number of trials (i.e., total number of seconds) and *p*, the probability of the event (i.e., switching tasks) occuring. The total switch count of a participant was the number of times the participant chose to switch out of the total number of seconds, which was at most 1800 for campaign 3 and 2700 for campaign 4. One limitation of using the binomial distribution was that the total number of seconds needed to be known; it was assumed to be the maximum value for the ‘withheld’ data. Other potential distributions, such as Gaussian or Poisson, have different limitations that resulted in them not being used. Using a Gaussian distribution for switch rate (i.e., switch count divided by total time) allows negative switch rates to be predicted, which were not possible for task switching. A Poisson distribution, like the binomial, is a count distribution, and when the number of trials, n, is very large or unknown and the probability of an event, p, is small the binomial converges to the Poisson distribution with a single parameter to describe its shape, $\lambda = np$. The Poisson distribution was expected to be the better option for this data given a relatively large number of trials compared to the number of switches for each participant. However, comparing two models where each distribution was used in a simple intercept-only model for the entire 'not withheld' dataset favors the binomial distribution, as shown in Figure 60 and using the LOO-IC values (LOO-IC$_{\text{binomial}}$ = 2537, LOO-IC$_{\text{Poisson}}$ = 2602).

The models that included task performance as the outcome used a Gaussian distribution for the likelihood function. Task performance values were normalized to range between zero and one so using a Gaussian distribution in the GLM was less than ideal, but since the calculation of the performance values was not well documented a Gaussian distribution was probably as good as any other option and was simple to implement.

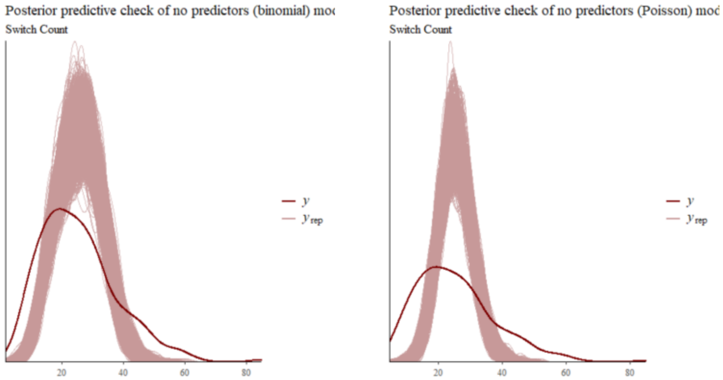


Figure 60. Posterior density plot for in-sample predictions using binomial (left) and Poisson (right) intercept-only model. Thick line indicates true value. Thin lines are posterior predictions using the model.

A baseline intercept-only model was first run for each counterfactual question. Revised models then tested the hypotheses related to the out-of-sample accuracy of the predicted task switching and performance values. The first revised model included only the cluster assignment for each participant, generated using machine learning techniques, as described in section 3.1 and section 4. Since predictions were made using the ‘withheld’ data, cluster assignments were needed for those participants also. The ‘withheld’ participants were not included in the machine learning analysis, to address the first research question, or to build the counterfactual GLM. The cluster assignment for the ‘withheld’ dataset, which also included all the ‘not withheld’ participants, was generated using the same ‘top 4’ predictors in the uRF model from section 4.2.2.3. That uRF dissimilarity matrix is fed into the LOO hierarchical clustering algorithm to generate cluster values for all the participants, as shown in Figure 28. The clustering values generated using only the ‘not withheld’ data were used to build the counterfactual models and the clustering values generated for the ‘withheld’ participants were used to make the counterfactual predictions. The counterfactual predictions were not true counterfactuals since the task had to be completed by the ‘withheld’ participants to know their cluster value. A true counterfactual prediction requires a way to generate a clustering or strategy value (or type) without the participant needing to complete the task, perhaps by experimental manipulation or another testing method.

The second revised model includes the parameter values from the single-stage information foraging theory model. As discussed in the previous section, no interactions between parameters were included in the models. The parameter values from ‘not withheld’ data were used to build the Bayesian GLM and the parameters from the ‘withheld’ data were used to make counterfactual predictions. The results in section 5 are all from ‘not withheld’ participants, run separately on campaigns 3 and 4, to investigate and explain how a participant decides to select a new task or remain at their current task. A repetition of the best model, the single-stage information foraging theory model, estimated parameters using the ‘withheld’ dataset, separately on campaigns 3 and 4, to determine parameter values to use for the ‘withheld’ participants in the counterfactual prediction models. The revised counterfactual models, to address the between-subjects and the within-subject questions, were built using parameters estimated from only the ‘not withheld’ data. The predictions rely on parameter estimates for the ‘withheld’ participants generated using all of the data.

The final revised model used both the cluster values and the cognitive process model parameters. The model was built using values generated with only the ‘not withheld’ data and the predictions are made for the ‘withheld’ participants using values generated using all the data.

Two additional types of predictors are used in the improved models; the ‘top 4’ most important decision strategy factors were used in models to predict task switching for the within-subject, between-subjects and other counterfactual questions, and HERA participant ID was used as a hyperparameter in multilevel models to predict task switching for the within-subject counterfactual question. The ‘top 4’ most important predictors were a result of the machine learning analysis to determine decision strategy and were included as an alternative to cluster value as the counterfactual predictor. Multilevel level modeling, which includes different intercept and slope values for each participant, was appropriate for modeling the within-subject counterfactual question where the same individual performs the task multiple times.

6.1. Baseline models

A total of 12 baseline models were generated: one for each outcome of task switching, individual performance, team performance, & multi-team performance using three different sets of data to address each counterfactual question. The baseline served as the worst-case prediction to compare to the revised models. Figure 61 shows the distribution of predictions for the in-sample ‘not withheld’ participants for each outcome (column of plots) and each counterfactual question (rows of plots). While the baseline models performed fairly well at in-sample predictions for task-switching, the models did not capture the multiple peaks in the performance data. Figure 62 shows the out-of-sample predictions of the ‘withheld’ participants, as light grey circles, for each counterfactual question. Almost all of the predicted values deviated from the dark-grey true values. The LOO-IC values and MAE and RMSE of the out-of-sample data are included in Table 34 to compare to the revised models.

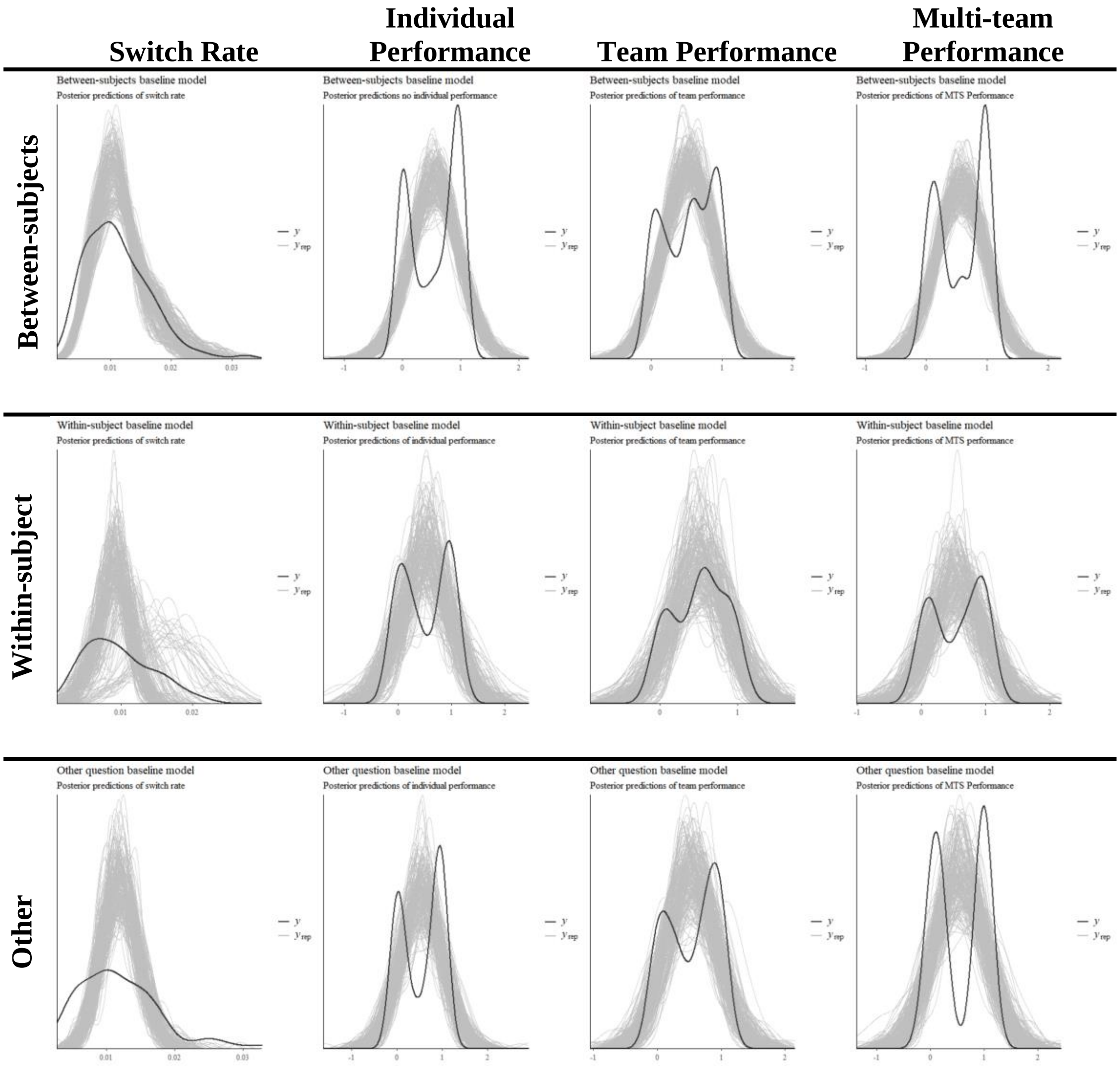


Figure 61. *Posterior density plots for between-subjects, within-subject, and other counterfactual questions. Thick line indicates true values. Thin lines are posterior predictions using model. Model and posteriors generated using 'not withheld' data.*

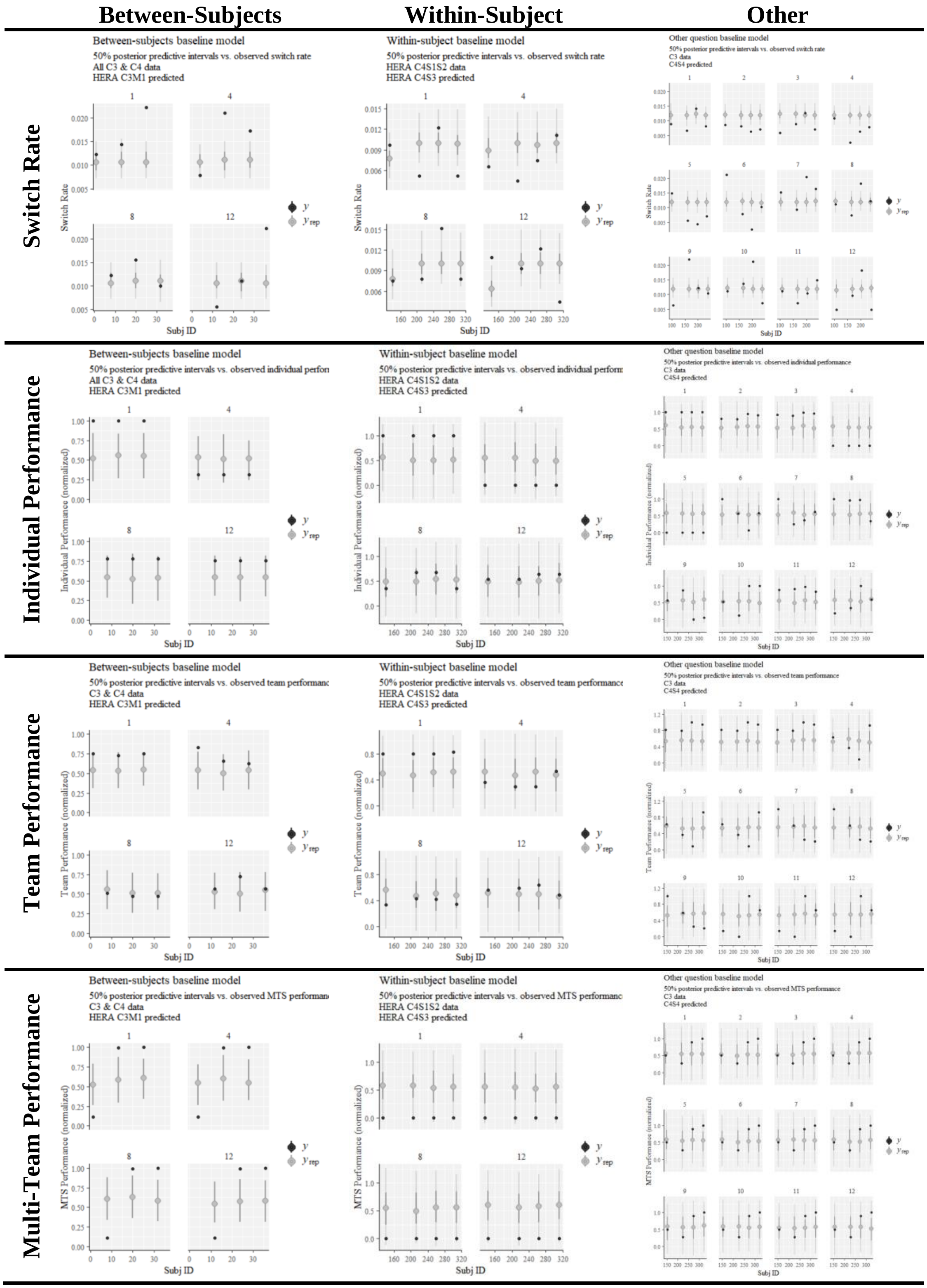


Figure 62. Posterior predictive interval plots, split by role, for between-subjects, within-subject, and other counterfactual questions using baseline (no predictor) models. Models fit using 'not withheld' data and posterior predictions for 'withheld' data.

6.2. Strategy revised models

The first set of revised counterfactual prediction models used the decision strategy, represented either by the cluster assignment or the most important factors, as the predictor of outcome. The models were used to address the within-subject counterfactual question. The assigned clusters taken from the best machine learning results, as described in section 4, for the ‘not withheld’ HERA subjects from campaign 4, sessions 1 & 2 were used as a predictor to build the Bayesian GLM. The resulting alpha and beta parameters from this model were used with the assigned clusters for the campaign 4 session 4 HERA participants, which were identified as the subjects of interest for the within-subject question, to determine their predicted switch rates and performance values.

There was no difference between the baseline model and the model that included the assigned clusters for switch rate (dLOO-IC=3, SE=12.4), while the cluster model had slightly lower RMSE and MAE values for the counterfactual predictions of switch rate than the baseline model. There was little consistency in the cluster values across the sessions for the HERA participants and almost no relationship between the cluster value and the switch rate, as shown in Figure 63, which likely led to there being little to no difference in the predictions between the two models. Comparing the predicted responses between the two models in Figure 64, there were no visual differences in the predicted responses. The decision strategy analysis also identified the most important factors related to switch rate so an additional Bayesian GLM used the three continuous variables – typical task switches, mean priority, and mean salience – as predictors. The indicator variable of explorer status was not included. This model was no different than either the baseline or the cluster model (dLOO-IC=1, SE=22.0; dLOO-IC=2, SE=18.9), but did generate better predicted responses for the within-subject counterfactual question, as shown by the error values in Table 34 and visually in Figure 64.

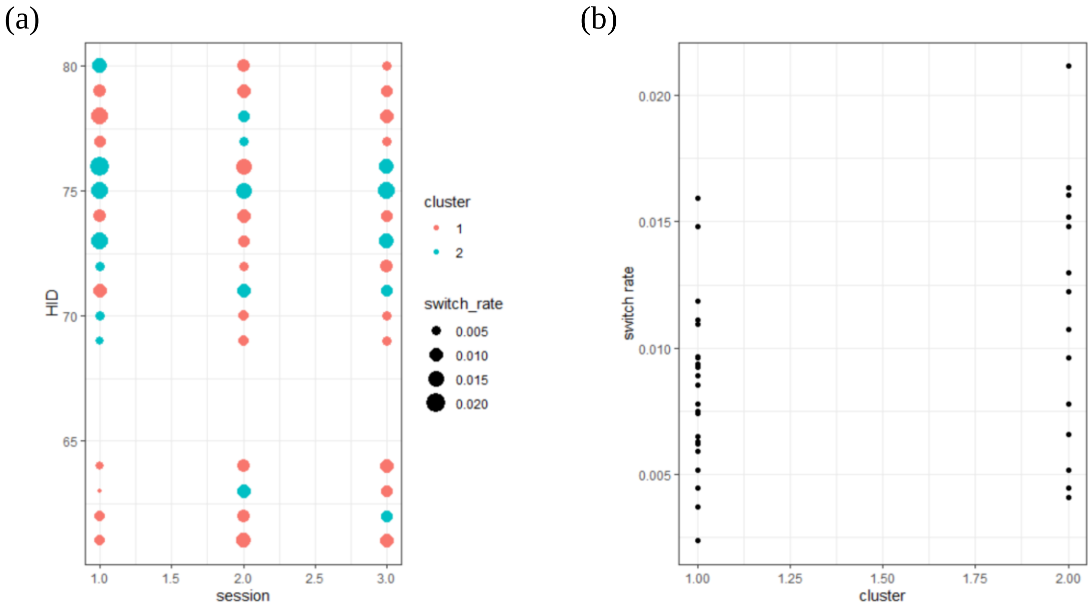


Figure 63. (a) Cluster values by HERA participant for each session (b) cluster values vs. switch rate

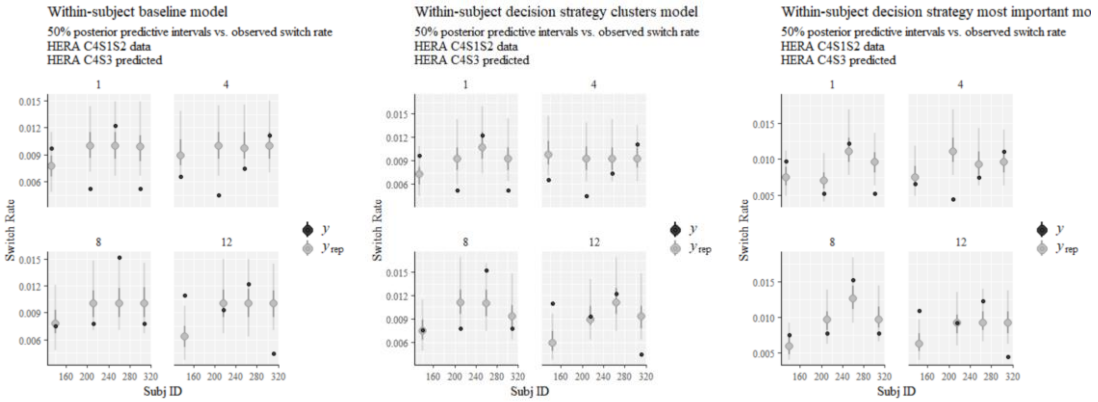


Figure 64. Within-subject counterfactual predictions of switch rate using decision strategy models compared to baseline model

A multilevel model is useful to model results that have a group of measures, in this case the different HERA participants, that naturally differ from one another and it provided an additional method to capture individual variation in output. The decision strategy clusters were included in a multilevel model that used the HERA participant ID as the hyperparameter. This model was favored over the baseline model (dLOO-IC=107, SE=29.6), but the predicted counterfactual responses were no better than those from the baseline model. To determine if the improved model fit was only due to structuring the model as a multilevel model, a multilevel version of the intercept-only baseline was run. That model also produced counterfactual predictions that were no better than the baseline model predictions and the multilevel cluster model was favored over this model (dLOO-IC=25, SE=10.9). Including HERA ID as a hyperparameter in a multilevel model improved the cluster model, but not the baseline model.

There was no difference between the baseline model and the cluster model for individual (dLOO-IC=0), team (dLOO-IC=0), and multi-team (dLOO-IC=0) performance; comparing the density distributions from the baseline model (grey) and the cluster model (blue) in Figure 65 provided a visualization of the similarity between the two models. Both models also generated counterfactual predictions of performance with a large amount of error, with larger error values for the predictions generated using the cluster model (Table 34); this is also visualized in Figure 62b and Figure 66 using the posterior predictive intervals for each model. This similarity between the baseline and decision strategy clusters models could be because the clusters were learned using the more important predictors of switch rate from the regression RF; there may be different factors that are important predictors of performance that produce different decision strategy clusters. Running the algorithm sequence again (i.e., regression RF, top predictors into uRF, uRF dissimilarity matrix into clustering algorithm) using the performance values as the outcome in the regression RF may produce clusters that improve the performance predictions.

Alternatively, the Gaussian distribution wasn't the best choice to model the bimodal distributions of the normalized performance outcomes. One alternative option is to use a different distribution in the Bayesian GLM. Another option is to fit the Bayesian GLM using the non-normalized performance outcomes. However, the non-normalized responses have a very large range of values that could be difficult to fit, as shown in Figure 67, which was what originally led to using the normalized values. Also, the scoring criteria was not repeatable and the scores were difficult to interpret.

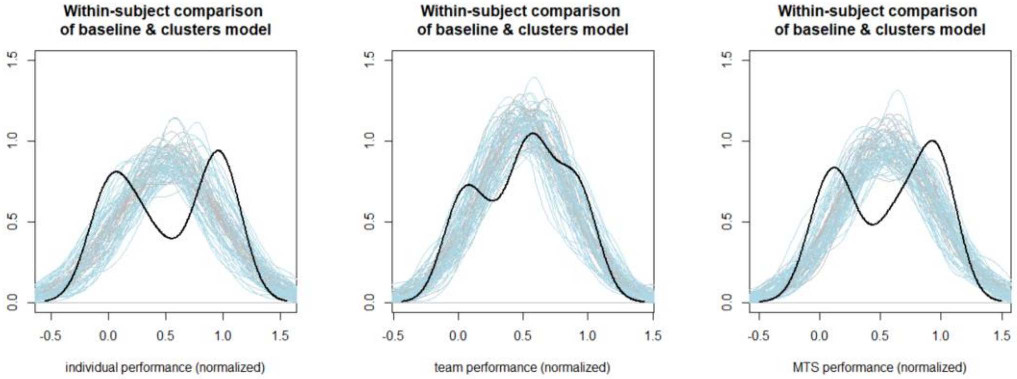


Figure 65. Comparison of within-subject posterior predictions of 'not withheld' campaign 4, sessions 1 & 2 data using baseline (grey) and clusters (blue) model

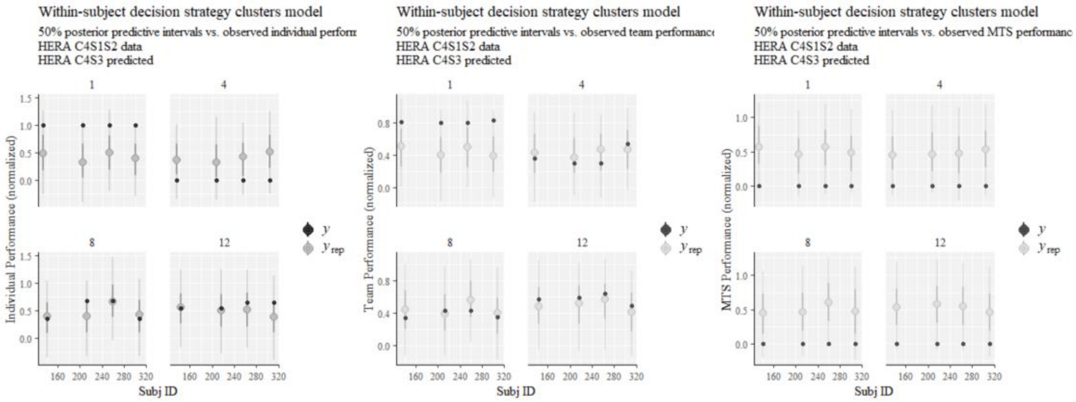


Figure 66. Within-subject posterior predictions of 'withheld' campaign 4, session 3 data using clusters model

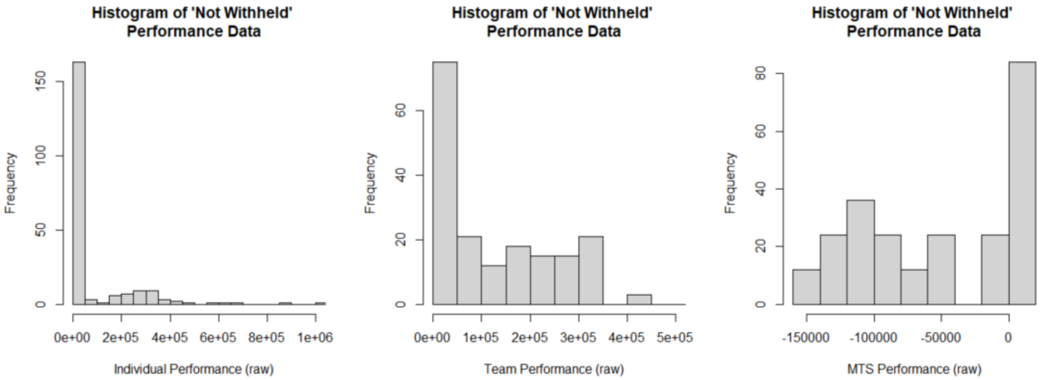


Figure 67. Histogram of performance values

6.3. Cognitive process revised models

The next set of revised models used a linear combination of the parameters from the single-stage information foraging theory based MLBA (i.e., threshold, starting point parameter, attention weight, value, and non-decision time) as the predictor of outcome. The values for the starting point parameter were an order of magnitude larger than any other parameter models so it was scaled to improve the estimation of the GLM alpha and beta parameters. The values from the cognitive process models addressed both the within-subject and between-subjects counterfactual question. Again, for the within-subject question, the values for ‘not withheld’ HERA subjects from campaign 4, sessions 1 & 2 were used to build the Bayesian GLM and the resulting alpha and beta parameters were used with parameter values for the campaign 4 session 4 HERA participants, which were identified as the subjects of interest for the within-subject question, to determine their predicted switch rates and performance values. For the between-subjects question, the model was built using the parameter values from all the ‘not withheld’ participants and the resulting GLM parameters were used with MLBA parameter values for the campaign 3 mission 1 HERA participants, which was the all-female crew identified in the between-subjects question.

For the within-subject question, there was no difference between the baseline model and the model using the MLBA parameters (dLOO-IC=4, SE=19.4), or in the resulting counterfactual predictions of switch rate from the models. A multilevel version of the MLBA parameters model was favored over the baseline model (dLOO-IC=91, SE=31.8), but was no different than the multilevel baseline model (dLOO-IC=13, SE=13.0), and this model generated counterfactual predictions of switch rate with more error than the baseline model. The improvements for the multilevel MLBA model were due to the multilevel structure, not from including the MLBA parameters as predictors. For the between-subject question, the MLBA parameters model was favored over the baseline model (dLOO-IC=118, SE=70.5) and generated counterfactual predictions of switch rate with lower error than the baseline model, as shown in Figure 68.

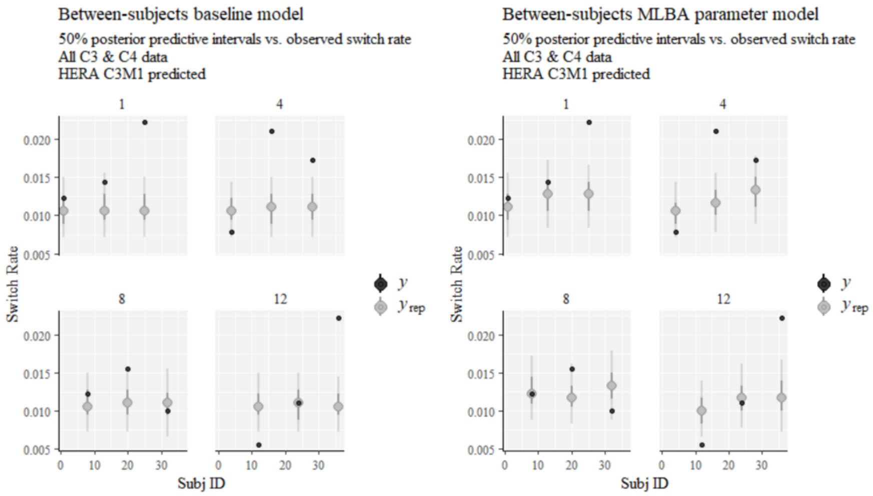


Figure 68. Between-subjects counterfactual predictions of switch rate using cognitive process model parameters compared to baseline model

For the within-subject counterfactual question, the baseline model was favored over the model that used a linear combination of MLBA parameters for the individual performance (dLOO-IC=3.7, SE=2.2) and team performance (dLOO-IC=22.6, SE=5.4) outcomes. For multi-team performance, there was no difference between the baseline and the model using the MLBA parameters (dLOO-IC=4.3, SE=5.1); however, the MLBA parameters model was better able to capture the range of multi-team performance values, as shown in Figure 69, and generated counterfactual predictions of multi-team performance with lower error values than the other models (Table 34). Both models also generated counterfactual predictions of performance with a large amount of error for the other two performance measures (i.e., individual and team performance).

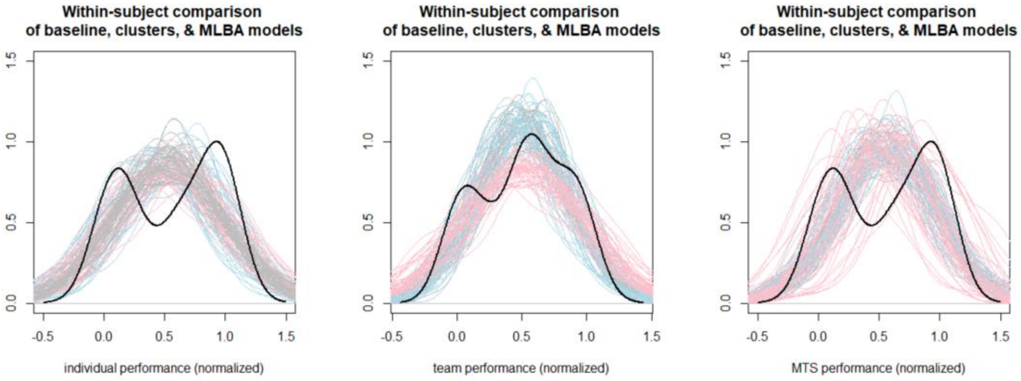


Figure 69. Comparison of within-subject posterior predictions of 'not withheld' campaign 4, sessions 1 & 2 data using baseline (grey), clusters (blue), MLBA (light red) models

For the between-subject counterfactual questions, there was no difference between the baseline and the model that used a linear combination of MLBA parameters for team (dLOO-IC=3.5, SE=4.3) and multi-team (dLOO-IC=0.4, SE=4.8) performance. The baseline model was favored for individual (dLOO-IC=7.3, SE=1.0) performance. The models also generated counterfactual predictions with larger error values than the baseline model for individual, team, and multi-team performance. Including cognitive process model parameters did not improve the predictions of performance over the baseline intercept-only model for the between-subjects counterfactual.

6.4. Strategy & cognitive process revised models

The final set of revised models used a linear combination of the assigned clusters from the decision strategies machine learning model and the parameters from the single-stage information foraging theory based MLBA as predictor of outcome. These models, also referred to as the all-parameters models, were used to address the within-subject, the between-subjects, and the other counterfactual question. The models for the within-subject and between-subjects questions were built using the combination of the values from the previous sections. For the other counterfactual question, the model was built using the parameter values from all the campaign 3 'not withheld' participants and the resulting GLM parameters were used with the assigned cluster and MLBA parameter values for the campaign 4 session 4 participants.

For the within-subject question, there is no difference between the baseline model, a model using a linear combination of assigned cluster plus the MLBA parameters, or a model using a linear combination of the most important factors plus the MLBA parameters (dLOO-IC=4, SE=23.2; dLOO-IC=5, SE=24.2). However, including the most important factors with the MLBA parameters produced counterfactual predictions of switch rate with less error than the baseline model, as shown in Table 34 and Figure 70. For the between-subjects question, model using a linear combination of the most important factors plus the MLBA parameters was favored over both the baseline model and the model using a linear combination of cluster value plus the MLBA parameters (dLOO-IC=871, SE=113.2; dLOO-IC=504, SE=82.7) and generated predictions of switch rate with less error than any other between-subjects model, as shown in Table 34 and Figure 71. The model using a linear combination of cluster value plus the MLBA parameters was also favored when compared only to the baseline model (dLOO-IC=367, SE=86.4). For the other counterfactual question, the model using a linear combination of the most important factors plus the MLBA parameters was favored over both the baseline model and the model using a linear combination of cluster value plus the MLBA parameters (dLOO-IC=232, SE=54.0; dLOO-IC=126, SE=40.6). Both the model using a linear combination of the most important factors plus the MLBA parameters and the baseline model generated out-of-sample counterfactual predictions of switch rate with approximately the same error, as shown in Table 34 and Figure 72.

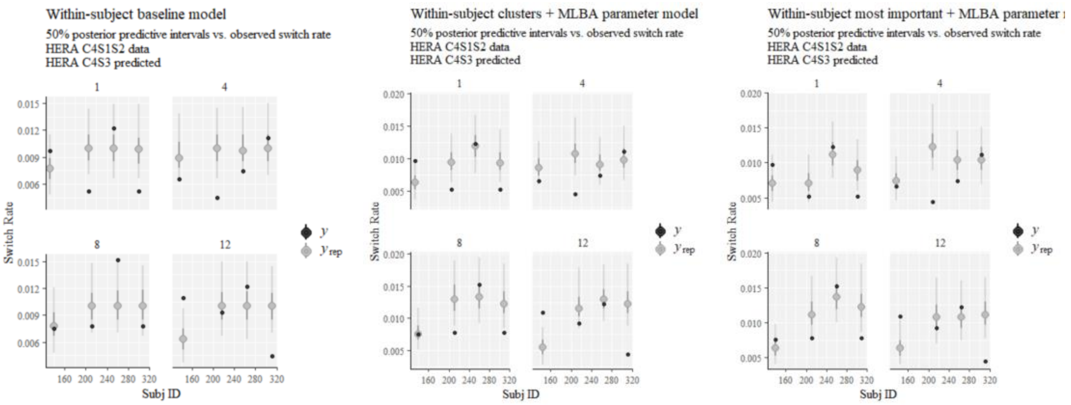


Figure 70. Within-subject counterfactual predictions of switch rate using decision strategy and cognitive process model parameters compared to baseline model

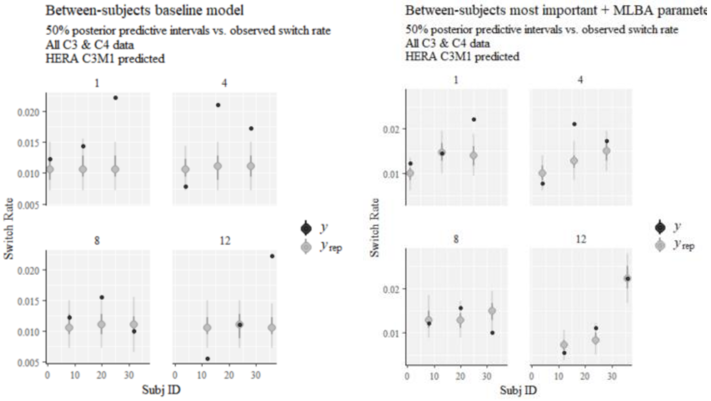


Figure 71. Between-subjects counterfactual predictions of switch rate using decision strategy and cognitive process model parameters compared to baseline model

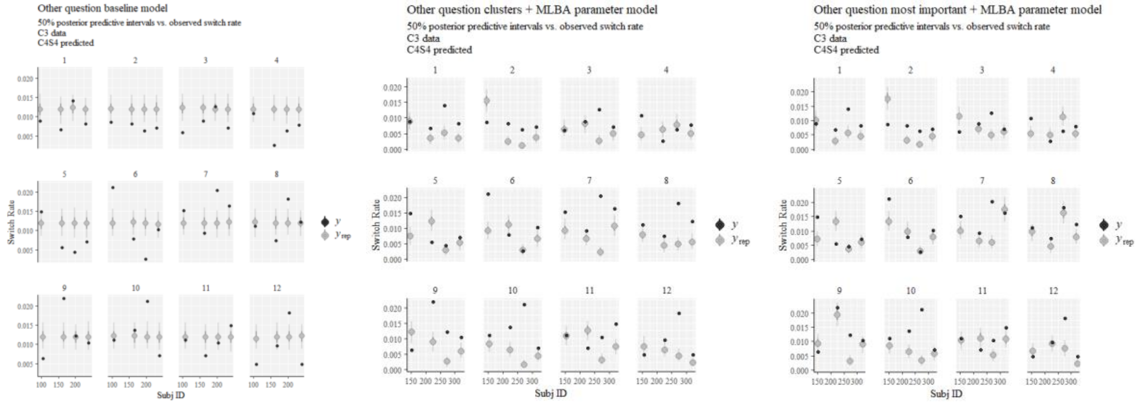


Figure 72. Other counterfactual predictions of switch rate using decision strategy and cognitive process model parameters compared to baseline model

For the within-subject counterfactual question, there was no difference between the baseline model and the model that used a linear combination of decision strategy clusters and MLBA parameters for the individual performance (dLOO-IC=3.1, SE=3.7), team performance (dLOO-IC=3.4, SE=4.3), or multi-team performance (dLOO-IC=1.6, SE=5.3) outcomes. This model also generated counterfactual predictions with higher error than the baseline, and all the other models, for all three performance outcomes.

There also was no difference between the baseline model and the model that used a linear combination of decision strategy clusters and MLBA parameters for team performance (dLOO-IC=5.7, SE=4.3) and MTS performance (dLOO-IC=1.8, SE=4.9) for the between-subjects counterfactual question, but for individual performance the baseline model was favored (dLOO-IC=8.7, SE=2.0). There also was no difference in the error of any of the between-subjects counterfactual predictions of performance.

Finally, for the other counterfactual question, there was no difference between the baseline model and the model that use the cluster and MLBA parameters for individual (dLOO-IC=5.1, SE=4.7), team (dLOO-IC=3.5, SE=3.1), or multi-team (dLOO-IC=1.7, SE=5.9) performance. The model that includes the assigned cluster and MLBA parameters generate counterfactual predictions with larger error values than the baseline model.

Table 34. Summary of Counterfactual Model Results

	LOO-IC	**SE$_{LOO-C}$**	**dLOO-IC**	**SE$_{dLOO-IC}$**	**pLOO**	**CF MAE**	**CF RMSE**
Within-Subject							
Switch Rate							
Multilevel Clusters	**294**	12.2	-	-	31.2	0.0033	0.0039
Multilevel MLBA	306	15.9	12	15.0	36.9	0.0034	0.0043
Multilevel Baseline	319	20.0	25	10.9	33.4	0.0032	0.0041
MLBA params	397	33.5	103	32.7	15.9	0.0032	0.0037
Clusters + MLBA	397	24.7	103	32.9	19.4	0.0029	0.0034
Clusters	398	30.8	104	29.0	9.6	0.0025	0.0029
Most important	400	28.6	106	29.3	15.2	0.**0021**	**0.0027**
Baseline	401	31.8	107	29.6	4.8	0.0028	0.0032
Most imp + MLBA	401	31.0	107	30.6	24.3	0.0025	0.0031
Individual Perf							
Clusters	47.9	4.4	-	-	2.1	0.32	0.42
Baseline	48.4	3.2	0.5	3.4	1.1	0.31	0.36
Clusters + MLBA	51.5	4.6	3.6	2.0	4.1	0.40	0.47
MLBA params	52.1	3.5	4.2	3.7	3.3	0.29	0.35
Team Perf							
Baseline	29.1	5.6	-	-	1.4	0.16	0.19
Clusters	29.1	5.6	-	-	2.3	0.17	0.22
Clusters + MLBA	32.9	6.5	3.8	4.3	4.7	0.26	0.28
MLBA params	51.7	3.5	22.6	5.4	3.1	0.17	0.19
MTS Perf							
MLBA params	35.7	5.9	-	-	3.5	**0.22**	**0.25**
Clusters + MLBA	38.4	6.0	2.7	0.6	4.7	0.63	0.63
Baseline	40.0	3.9	4.3	5.1	1.2	0.56	0.56
Clusters	40.5	4.5	4.8	5.1	2.1	0.50	0.50
Between-Subjects							
Switch Rate							
Most imp + MLBA	**1666**	42.2	-	-	18.2	**0.0030**	**0.0040**
Clusters + MLBA	2170	96.1	504	82.7	30.9	0.0042	0.0052
MLBA params	2419	115.6	753	108.7	32.6	0.0043	0.0055
Baseline	2537	125.6	871	113.2	6.1	0.0049	0.0063

Individual Perf							
Baseline	259.3	9.3	-	-	1.2	0.29	0.30
MLBA params	266.6	9.4	7.3	1.0	4.9	0.32	0.34
Clusters + MLBA	268.0	9.5	8.7	2.0	5.9	0.32	0.35
Team Perf							
Baseline	184.3	12.4	-	-	1.3	0.12	0.15
MLBA params	187.8	13.2	3.5	4.3	5.1	0.21	0.25
Clusters + MLBA	190.0	13.1	5.7	4.3	6.2	0.22	0.25
MTS Perf							
MLBA params	234.3	9.7	-	-	5.2	0.43	0.43
Baseline	234.7	9.0	0.4	4.8	1.2	0.42	0.42
Clusters + MLBA	236.5	9.8	2.2	0.7	6.2	0.43	0.43
Other							
Switch Rate							
Most imp + MLBA	**474**	18.6	-	-	14.3	**0.0043**	**0.0056**
Cluster + MLBA	600	50.8	126	40.6	22.1	0.0056	0.0071
Baseline	706	63.6	232	54.0	5.4	**0.0044**	**0.0051**
Individual Perf							
Baseline	87.0	4.2	-	-	1.1	0.38	0.41
Cluster + MLBA	92.1	4.7	5.1	4.7	4.4	0.45	0.61
Team Perf							
Baseline	68.7	5.8	-	-	1.2	0.32	0.35
Cluster + MLBA	72.2	6.0	3.5	3.1	4.6	0.48	0.59
MTS Perf							
Cluster + MLBA	90.7	6.6	-	-	4.4	0.61	0.68
Baseline	92.4	2.2	1.7	5.9	1.0	0.28	0.32

6.5. Summary

The results from the decision strategies and cognitive process models were used to identify residual information about individual differences in performing the overall well placement task; this information was used within Bayesian generalized linear models as predictors of switch rate, individual task performance scores, team task performance scores, and multi-team task performance scores. The three different counterfactual questions – within-subject, between-subjects, and other – required different subsets of the data to be used to build the model and also to generate the different counterfactual predictions. Table 35 gives a summary of the conclusions compared to the hypotheses, for the models that were listed in the original hypotheses (e.g., the most important factors models are not included since that was not specified in the original hypothesis).

This research identified several conclusions from the counterfactual models that included switch rate as the outcome parameter. Models that included only the parameters related to the decision strategy – either assigned cluster or most important factors – were used to evaluate the first hypothesis within research question 4 (i.e., H4a), which only considered the within-subject counterfactual question. First, model comparison favored the baseline model over the decision strategy clusters model for within-subject question, which does not support the hypothesis. Including the most important decision strategy factors in the model resulted in a model that is no better or worse than the baseline model, using information criteria to compare the models, but generated predictions for the within-subject counterfactual question that have less error than the baseline model, and all other models, which supports the hypothesis. The models that included the parameters from the single-stage information foraging theory based MLBA (i.e., threshold, starting point parameter, attention weight, value, and non-decision time) were used to evaluate the second hypothesis within research question 4 (i.e., H4b), which considered both the within-subject and between-subjects counterfactual question. The model was no different than the baseline model for the within-subject counterfactual question, using both information criteria and the errors of the counterfactual predictions. The model was favored over the baseline for the between-subjects counterfactual question, using information criteria to compare the models, and it generated counterfactual predictions with less error than the baseline. These results support the hypothesis. Finally, the models that included both the assigned clusters as well as the MLBA parameters, referred to as the all-parameters model, were used to evaluate the third hypothesis within research question 4 (i.e., H4c), which considered all three counterfactual questions. The analysis found no difference between the baseline and all-parameters models for the within-subject counterfactual question, but the all-parameters models were favored over the baseline for both the between-subjects and the other counterfactual question. Using the most important decision strategy factors plus the MLBA parameters in the GLM improved out-of-sample predictions for within-subject and between-subjects questions while the baseline model best predicted out-of-sample responses for the other question. Finally, multilevel versions of several models were built using the data for the within-subject question; all were favored over the single-level models, but none improved out-of-sample predictions.

Table 35. Summary of Counterfactual Model Conclusions

Hypothesis			**Conclusions**	
	Predictor(s)	**Counterfactual Question(s)**	**Task-switching**	**Performance**
4a	Decision-making strategy cluster	Within-subject	**Partially supported** * lower LOO-IC, **if using a multi-level model**	**No difference** compared to baseline
4b	Cognitive process model parameters	Within-subject	**No difference** compared to baseline	**Partially supported** * lower RMSE for **multi-team performance**
		Between-subjects	**Supported** * lower LOO-IC and RMSE	**Not supported**
4c	Decision-making strategy cluster & cognitive process model parameters	Within-subject	**No difference** compared to baseline	**Not supported**
		Between-subjects	**Supported** * lower LOO-IC and RMSE	**No difference** compared to baseline
		Other	**Partially supported** * lower LOO-IC, but higher RMSE	**Not supported**

NOTE: Not supported indicates that the baseline model was favored. No difference indicates that both the baseline and tested model perform equally.

7. DISCUSSION

This research investigated four different research questions using three distinct modeling paradigms where the approaches fit together to form an overall analysis pipeline. First, a machine learning approach was used to identify the decision strategy used to complete the overall task and the most important factors contributing to that strategy to address the first research question. Next, multiple cognitive process models were compared to find the best model to describe and explain the cognitive mechanisms used in completing the overall task and the associated task-switching behavior to address the second and third research questions. Finally, a Bayesian generalized linear modeling (GLM) approach was used to generate counterfactual predictions of task-switching rates and overall task performance scores to address the final research question. There were various choices and assumptions made to test the hypotheses associated with each research question, that in some cases affected the generalizability of the results. These are discussed in more detail for each modeling approach in the sections below.

7.1 Decision strategies

The machine learning analysis identified patterns that indicate differences in clusters of participants completing the overall task where the clusters represented the different strategies used by the participants, assuming that participants use different strategies to complete the task. Using the best machine learning model to predict out-of-sample results on the test data sample produced predictions better than random assignment. The results confirm the first hypothesis.

This research expected to see the clusters fall between two extremes of behavior. Highly exploratory behavior is one extreme, where the individual or team members switch(es) tasks often and complete(s) a larger number of tasks, which indicates that an individual was using a compensatory strategy to complete the overall tasks by processing all relevant information and trading off the good and bad aspects of each alternative. Highly exploitive behavior is the other extreme, where the individual or team members switch(es) tasks infrequently and complete(s) only a small number of distinct tasks, which indicates that an individual was using a non-compensatory strategy to reduce information processing demands by ignoring potentially relevant problem information. However, the best model to fit the observed data had only two relatively loosely group clusters without large separation between the clusters. The results are not able to delineate groups with intermediate behavior and do not confirm that there are two extremes in behavior. The assumption that observed task-switching behavior is related to unobserved decision strategy may still be valid, but should be tested using data where the true decision strategy is known.

While the results of this research are specific to the Project RED data, the methodology used to analyze participants' decision strategy in the Project RED task could be generalized to other tasks. The method found patterns of similar participants displaying similar strategies by using the important predictors relating to the observed behavior. For the Project RED task-switching data a key assumption was that the unobserved decision strategies related to the observed task-switching outcome. This assumption could apply to other task-switching tasks, such as an overall monitoring task (e.g., maintaining normal plant operations) or an investigation task (e.g., medical diagnostics), if completed in an environment without unrelated interruptions, but not to other types of tasks such as preferential choice, since multiple strategies could lead to the same observed response (Lee et al, 2019). For other types of tasks, the methodology could be applied using only the unsupervised methods, without any assumption of relation between the underlying strategy and the observed response. However, the results using only the unsupervised method for the Project RED data were approximately the same as random assignment, so applying this method to data from other types of tasks may not provide useful results.

For other types of task-switching data both the supervised and unsupervised methods could be used to analyze the data. While results from this research show that the supervised random forest gives the best predictions of both in-sample and out-of-sample switch rates, the supervised method does not identify any similarities between the participants and using the regression RF dissimilarity matrix with various clustering algorithms produces results that are no better than random assignment. The unsupervised method allows the variables alone to drive the clustering between participants to provide insights into commonalities in participants with both similar and different switching behavior. Because we don't know what strategy (or strategies) a participant used when completing the task, there is no ground truth to compare to. Switch rate is one observed behavior that describes how a person completed a task, but as mentioned above the assumption that the observed switch rate relates to the unobserved strategy should be tested using data where the true decision strategy is known.

7.2 Cognitive mechanisms

The results from the cognitive process analysis favored a model based on alternative-level preferences over a model based on attribute-level preferences as well as a single-stage model, which assumed that all tasks were considered simultaneously, over a two-stage model, which assumed a serial process where only a subset of tasks were considered for the final decision. This confirms the second hypothesis, but not the third hypothesis.

This research used both real and simulated datasets to determine which version of the MLBA model provided the most evidence for the observed data. Because the Project RED data had a small number of trials for many of the participants, leading to increased uncertainty in the modeling results, a simulated dataset was generated to compare the fit metrics between the MLBA models. A generic LBA model was used to generate the data, using parameter values from a range of values that was representative of the real data. Although the response times and the particular responses selected in the simulated data did not allow the dataset to be used in place of the Project RED data to explain the behavior, a visual assessment determined there was sufficient similarity in the response time distribution and the variation in the responses that enabled the simulated data to be used to compare differences in switching behaviors using the different models and address the research questions.

The single-stage models used in this research are generalizable to other tasks, both other types of task-switching tasks as well as other types of tasks. The cumulative prospect theory version of the single-stage model has already been applied to preferential choice and risky choice tasks (Cohen et al., 2017), where there are multiple attributes identified for each response option and the value and weight of each attribute was known. Similarly, the information foraging theory version of the single-stage model could also be generalized to other tasks. The methodology related to determining the profitability would need to be changed to apply it to other types of tasks. Ideally, the profitability of each choice would be identified prior to data collection rather than empirically afterwards. This requires knowing or measuring the gain that an option provides in relation to the cost or processing time of that option. For a task-switching task, the gain was the popularity of the response option and the processing time was the amount of time spent completing it. To repeat this analysis on another task-switching scenario, the task could be set up to quantify the gain as the amount of benefit each response option provides towards completing the overall task and the processing time as a relative measure of the time spent on each task. This would allow the values to be quantified a priori rather than needing to be calculated after the overall task is completed. For a preferential choice task, such as selecting which item to purchase, the gain could be the sum of features or benefits that the option provides while the cost could be the actual monetary cost or, more consistent with information foraging theory, it would be a relative measure of the time required to process the information provided by that option. An option with more features or features that contradict each other (e.g., a phone with more memory, but lower camera quality) should require a longer time to evaluate than one with simple or consistent features. For the information foraging theory based model the attributes still contribute to the decision, but at a higher level. They are not each evaluated separately, but are part of the overall profitability and value that the option provides.

The two-stage model structure is generalizable to other tasks with improvements, discussed in chapter 8, and modifications to the model details. The two-stage models in this research assume that the first stage is reducing the number of response options using the type of task, which is specific to task-switching involving teams. The model structure could still be applied using another heuristic, such as task typicality, to a task-switching scenario, by modifying the models to use this as the criterion for reducing the number of response options for the final decision. The factor used to reduce the number of response options for a task-switching task is not considered to be a task attribute. The structure may also be generalizable to other types of tasks, like preferential choice, if a heuristic is applied to reduce the number of choice options and not to make the final selection. For example, a participant using a take-the-best strategy could use cost to select the final response, where the two-stage model structure cannot be used. Alternatively, a participant could use cost as a factor to reduce the number of response options (e.g., anything less than $X), where the two-stage model could be used to represent the decision-making process. Although, the process may require more than two stages if other factors are used to iteratively reduce the number of options considered until final choice is made (e.g., elimination by aspects), which would require modifications to the model structure. This leads back to the question of how the attributes are used in the decision-making process – if the attributes are used simultaneously to trade-off between all the options or if they are used to systematically reduce the number of options considered until a single option remains.

An additional consideration in applying the cognitive models is that the single-stage models were shown to be identifiable while the two-stage models were non-identifiable. In the two-stage model the drift rate was a function of the threshold, stage one drift rate, stage two drift rate, and stage delay parameter. As the stage delay parameter varies (i.e., decreases or increases), the stage one and stage two drift rates must also vary to maintain the same overall drift rate across both stages, plus the overall drift rate varies with the threshold that was also varying to best explain the data, which was likely the cause of there being too much flexibility in the model. Even holding the drift rate parameters constant, allowing only threshold, starting point, non-decision time, and stage delay to vary, still results in the model not recovering the generating values of the free parameters. If the number of free parameters for the two-stage models are reduced to only include threshold, starting point, and non-decision time, then the models are identifiable, but doing so also removes all the parameters of interest to address the research questions.

This leads to the question of whether the model or the task-switching data were sufficient to address whether a participant considered the response options serially or in parallel. Only the measures of model fit, not the estimated parameter values, were used to test the hypothesis, so the model is still sufficient to use to compare the two-stage models to their single-stage counterparts. Still, a model structure with less dependency between the stages is needed. If the stage one response time was known or could be assumed, then the stage delay parameter could be eliminated and each stage could have a separate threshold and drift rate. Even if the stage delay was still a parameter, a model where the drift rate(s) are independent of the threshold may reduce the dependency between the stages enough to be identifiable. There may also be other manipulations of task-switching, different than how the data were collected during Project RED, that allow the estimated parameters as well as the model fit measures of the less-constrained two-stage model to be used. For example, if some participants were given a manipulation to lead to spending less time on each task or if the total time spent on the overall task was not the same for all the participants.

7.3 Counterfactual models

Generalized linear models, which can be applied to a large variety of datasets to predict outcomes using various independent variables, were used to predict the resulting task-switching behavior and task performance outcomes using the results from the decision strategies and cognitive process models. The results from the counterfactual prediction models are summarized in Table 35 to show how the conclusions compared to the hypotheses, for the models that were listed in the original hypotheses (e.g., the most important factors models are not included since that was not specified in the original hypothesis). Most of the hypotheses related to predicting task-switching behavior were supported or partially supported while most of the hypotheses relating to predicting overall task performance were not supported. The contextual factors used in the counterfactual prediction models (e.g., the decision strategy clusters) probably have a small effect on the resulting task-switching behavior or task performance, and the noise in the Project RED data, especially for the task performance outcomes, may make it difficult to perceive the effect. Including different contextual factors in each of the counterfactual questions may lead to improved predictions. For example, this research did not consider using only the decision-making strategy cluster to make predictions for the between-subjects counterfactual, but the strategy used by different participants is assumed to lead to different task-switching behaviors so it could improve predictions for the between-subjects counterfactual question.

8. FUTURE WORK

There are several areas in which to build upon this research. There are opportunities to perform additional analyses using the Project RED dataset as well as opportunities to further explore the idea of including cognitive process measures as counterfactual predictors of future performance. Additional work can also be completed to improve the two-stage MLBA model developed here or to identify a different model structure to apply to a serial process of deciding to switch tasks. Longer term, further research can be completed on the information foraging theory based MLBA and to investigate decision strategies used for task switching.

One additional area to investigate using the Project RED dataset is using the MLBA cognitive process models to investigate differences between the multiple sessions for the HERA participants. This research assumed that each iteration of the overall task and the resulting task-switching data was independent (i.e., each row of data was for a different participant) to understand overarching differences in individual performance, but the participants in the HERA roles completed the overall task multiple times over the 30- or 45-day mission. Looking at just the data from those participants gives insight to explain how the differences between each session affected their cognitive processes and if there are any mechanisms that change for all participants as a result of the experimental manipulations (e.g., longer communications delay).

There are multiple options to continue analyzing the Project RED data to identify and understand participants' decision-making strategies. One simple analysis is to measure, using an ANOVA, the effect of strategy on overall task performance. While outside the scope of this research, the results would provide insight into whether participants that cluster together also have similar performance measures. The exploratory analysis identified the mean time on task as a factor with a strong non-linear relationship to switch rate. This factor in a Bayesian GLM may be the only one needed to predict task switching. However, it must be measured when an individual completes the task, which makes it difficult to use as a counterfactual predictor. Additional research to understand why that relationship is non-linear may provide additional insights into task-switching strategies and behavior. Additionally, there may be other machine learning and data analysis techniques that can be applied to the data that improve the clustering results (i.e., increase ARI, decrease classification error). Specifically, since task-switching is time-series data, sequence analysis techniques (Ritschard & Studer, 2018) could be used to identify time-series patterns that relate to decision strategy. These patterns could be compared to task performance to determine if there is a sequence or pattern of tasks that results in improved performance. While this analysis would be specific to this dataset, the techniques developed to identify decision strategies from real-world data could be generalized to other tasks where the strategy a person uses to complete the task is undefined and unidentified.

The dataset used in this research has limitations that create challenges in estimating the cognitive process model parameters, and in making and evaluating counterfactual predictions. While the data includes second-by-second results for a 30- or 45-minute period, the actual number of trials (i.e., task-switches) is small, which adds error to the parameter estimates. The MLBA models are usually applied to structured, clean data with hundreds or thousands of trials per subjects whereas the real data used in this research is more complex, noisy, and sparse. The Bayesian estimation techniques, which incorporate priors into the estimates, provide parameter values with the small amount of data, but they require more computation time and still have higher error in the estimates from the small number of trials. Also, the set-up for collecting the data was complex and, since it was completed by another laboratory, was not fully understood, which creates challenges in evaluating the effect of including cognitive process parameters as predictors of behavior. Running the cognitive process model on a less-complex task with a larger number of trials for each participant, either on an existing dataset or a new dataset, would allow for further analysis into the effectiveness of including cognitive process parameters.

This research developed a new MLBA model framework based on information foraging theory and showed that this model is favored over another existing model to describe the task-switching data. This model includes changes both to the model structure (i.e., the attention weight equation based on information foraging theory) as well as to the underlying assumption that attribute measures are considered holistically. Additional research is needed to understand the effect of each of these changes. This could be done by running a version of the model based on cumulative prospect theory that assumes alternative-level decision-making on the simulated dataset and comparing it to the two single-stage models from this research. In addition, more research is needed to apply the information foraging theory based MLBA to data from existing literature for information foraging tasks as well as other decisions (e.g., perceptual choice) to better understand its performance.

The two-stage model developed in this research has a different structure (Figure 12) than originally intended (Figure 11a). The two-stage model needs a structure like the originally intended structure, with less dependency between the stages, which allows the drift rates to be independent of the threshold. To implement the intended structure, where the stage one threshold is the exact starting point of the second stage, requires using approximation methods (e.g., Probability Density Approximation; Holmes, 2015) to estimate the likelihood values. This research unsuccessfully attempted to implement this structure, but additional work is needed to refactor the code to generate the likelihood estimates without using the *rtdists R* package to correctly implement the structure in Figure 11a and evaluate that model to use to represent a serial decision-making process. An advantage to using an evidence accumulation model to represent a serial decision-making process is that it provides psychologically relevant parameters to explain the process. An alternative is to use a different model paradigm that evaluates serial versus parallel architecture for decision-making, like Systems Factorial Technology (SFT; Townsend & Nozawa, 1995), on a task-switching dataset. This could be challenging since SFT requires selective influence manipulations of parameters to create separation of high and low salience conditions, but research has been done using SFT to evaluate consumer choice tasks (Cooper & Hawkins, 2019), another complex decision with many possible strategies.

This research shows that people can be loosely grouped together to identify similarities in how they complete an overall task relating to their task switching behavior. However, additional research is needed to identify the decision strategy(ies) related to task switching independently of completing a primary task and to manipulate the strategy used. More research is also needed to test and model team strategies explicitly. Strategy identification and manipulation is already included in experiments with simpler tasks (e.g., Lee et al., 2019; Rieskamp & Otto, 2006) using single participants. An example applied to a more complicated task looks at the strategy used for task scheduling to understand how people reason about agenda changes (Rosenthal & Hiatt, 2020). Future research related to task-switching can both identify the strategy participants use to switch between multiple tasks (e.g., by self-reporting) and provide incentives to encourage behaviors of divergent thinking, curiosity, and exploring for more information, and then measure the effect of the strategy on task switching and overall task performance. One possibility is to set-up the multi-attribute task battery (MAT-B; Comstock & Arnegard, 1992) to require sequential task performance, rather than concurrent task performance, with different scenarios to encourage or discourage task-switching. Another possibility is to use a simulator of unmanned vehicle planning and operations (e.g., Research Environment for Supervisory Control of Heterogeneous Unmanned Vehicles (RESCHU; Boussemart & Cummings, 2008)) to allow single operators or teams of operators to control multiple unmanned vehicles on missions with different overall goals to manipulate and measure how people switch between the tasks related to unmanned vehicle mission planning and operations.

9. CONCLUSIONS

This research makes several novel contributions to the current knowledge base of decision-making – theoretically, methodologically, and practically. Theoretically, it identified and applied factors to describe the influence of task structure on individual performance. As part of the exploratory analysis, this research developed a measure of typical task switching, that was directly influenced by the overall task structure and was subsequently identified, using machine learning techniques, as one of the most important factors related to task switching. An alternative measure of atypical task switching was also provided and applied to define whether an individual preferred to explore new tasks (i.e., is an explorer) or repeat known tasks (i.e., is not an explorer). This indicator of whether a person was an explorer was also identified as one of the most important factors related to task switching. Additional parameters were derived that integrated principles from information foraging theory into the machine learning analysis of decision strategy. The exploratory analysis also developed a measure of the functional ties, or dependencies between participants, and identified less interaction between participants and less dependency on other participants' behavior on task-switching than expected. The participant dependencies, along with other provided measures of interpersonal ties, behavioral ties, and shared mental models were used to investigate team factors impacting individual decision-making strategies; team factors were not considered in previous research on task-switching behavior.

This research identified a pattern of individual strategies that predicts out-of-sample data better than random assignment. The actual strategies that participants used to complete the overall task were unknown and undefined a priori, making this a difficult problem. By applying machine learning techniques to the data, clusters of similar participants were found and assumed be similar to each other. There was large overlap in the clusters since the techniques did not provide perfect groupings, but the results performed better than randomly assigning participants to a group. The analysis also learned which factors were most important to switching tasks.

Another theoretical contribution is the incorporation of information foraging theory principles, explaining how people gather and exploit information, into the multi-alternative, multi-attribute linear ballistic accumulator (MLBA) model framework of decision-making, specifically by including a measure of the profitability of a task to account for the weight of attention given to a task, to quantify the underlying processes driving exploratory or exploitative behavior. The profitability measure is also directly tied to task structure since the values are determined for the specific tasks completed as part of the overall task. This measure, along with a measure of value, are incorporated at the task-level, assuming that the multiple task attributes are processed as a whole, while existing MLBA models assume each attribute is processed separately. As mentioned in chapter 8, an area of future work is to further investigate this model to understand the effect of the model structure (i.e., the attention weight equation based on information foraging theory) versus the effect of incorporating attribute measures at the task level in explaining decision-making response data. The two aspects are combined in the information foraging theory based models used in this research.

The information foraging theory based model, compared to one based on cumulative prospect theory, better describes and explains task-switching responses and response times. The model describes that people select which task to work on holistically while the resulting parameter values explain the differences in participants. Participants vary in their efficiency in responding, as determined using the parameters that mention the weight of attention given to a task and the value of the task, as well as by their tendency to switch tasks, measured by their range of starting points. Someone with a large range of starting points (i.e., larger *A*) is less likely to switch tasks since they have more opportunities to need a larger amount of evidence to select any next task. This supports the finding by Wickens et al. (2015) of switch-avoidance tendency as an important factor in task switching.

A final theoretical contribution is including these latent, residual measures of individual differences in cognitive strategy and cognitive processing as counterfactual predictors of out-of-sample responses. Only a few of the models that included the measures performed better than the baseline intercept-only model, but the dataset was also very noisy and complex which reduces the ability to detect small effects.

Methodologically, this research developed a sequence of machine learning techniques to process a large number of independent variables to identify clusters of similar participants. The data contains multiple types of variables (e.g., continuous, categorical) so this research applied the idea of unsupervised random forests, from genetics research, to generate a measure of dissimilarity between the participants that could be used to cluster participants. The development of the information foraging theory based model also provides a methodological contribution. The model can be generalized to apply to other tasks where the attributes of a task are less distinct or are conflated, like researching a topic, analyzing information, or monitoring conditions within an environment. Some tasks, especially laboratory tasks, encourage or require looking at each attribute, e.g., preferential choice tasks, so the model may not generalize to those tasks, but those types of tasks provide an opportunity to more rigorously test the results from the model. This research also developed an initial evidence accumulation model structure to explore whether alternatives all considered simultaneously or reduced in a serial process. Applying this model to the task switching data, and comparing it to a model that uses the traditional structure of the MLBA, found that people considered all the alternatives simultaneously. The initial version developed in this research uses the *rtdists R* package, which resulted in a non-identifiable model, but with some improvements, as described within the future work, it provides an alternative to other existing models to evaluate serial versus parallel processing in decision-making.

Finally, practically, the results from this research suggest that people select the next task to work on and switch between tasks by considering the ongoing and alternative tasks as a whole, not by using the specific attributes of the tasks. This idea can be applied to real-world scenarios where multiple tasks are available to switch between, like plant operations or piloting an airplane. There are examples from both domains where people will not select to switch to a more salient, higher priority task, which could be explained by considering that each of the available tasks is considered holistically rather than assuming they are evaluated using each attribute.

REFERENCES

Armstrong, J. S. (2001). *Principles of forecasting*. Norwell: Kluwer Academic.

Boussemart, Y., & Cummings, M. L. (2008). Behavioral recognition and prediction of an operator supervising multiple heterogeneous unmanned vehicles. *Humans operating unmanned systems*.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*(3), 153–178.

Byrne, K. A., Silasi-Mansat, C. D., & Worthy, D. A. (2015). Who chokes under pressure? The Big Five personality traits and decision-making under pressure. *Personality and Individual Differences*, *74*, 22–28. https://doi.org/https://doi.org/10.1016/j.paid.2014.10.009

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

Cohen, A. L., Kang, N., & Leise, T. L. (2017). Multi-attribute, multi-alternative models of choice,: Choice, reaction time, and process tracing. *COGNITIVE PSYCHOLOGY*, *98*, 45–72. https://doi.org/10.1016/j.cogpsych.2017.08.001

Comstock, J. R., & Arnegard, R. J. (1992). The multi-attribute task battery for human operator workload and strategic behavior research.

Cooper, G. J., & Hawkins, G. E. (2019). Investigating consumer decision strategies with systems factorial technology. *Journal of Mathematical Psychology*, *92*, 102258.

Cooper, G., Innes, Reilly I., Kuhne, C., Cavallaro, J.P., Gunawan, D., Hawkins, G., & Brown, S. (2021). pmwg: Particle Metropolis Within Gibbs. R package version 0.2.0. https://CRAN.R-project.org/package=pmwg

Defense Advanced Research Projects Agency. (2019). Teaching AI to Leverage Overlooked Residuals (TAILOR) Artificial Intelligence Exploration (AIE) Opportunity (Notice ID DARPA-PA-18-02-07). Retrieved from https://beta.sam.gov/api/prod/opps/v3/opportunities/resources/files/766a0702c1bd490b00dc1b65dd8760c7/download?api_key=null&status=archived&token=

Evans J.S., Murphy M.A. (2018). rfUtilities. R package version 2.1-3. https://cran.r-project.org/package=rfUtilities

Farrell, S., & Lewandowsky, S. (2018). *Computational modeling of cognition and behavior*. Cambridge University Press.

Geden, M., Smith, A., Campbell, J., Spain, R., Amos-Binks, A., Mott, B., Feng, J., & Lester, J. (2019). Construction and Validation of an Anticipatory Thinking Assessment. *Frontiers in Psychology*, *10*, 2749.

Gigerenzer, G. (2008). Why Heuristics Work. *Perspectives on Psychological Science*, *3*(1), 20–29. https://doi.org/10.1111/j.1745-6916.2008.00058.x

Green, R. F. (1984). Stopping rules for optimal foragers. *The American Naturalist*, *123*(1), 30–43.

Gunawan, D., Fiebig, D., & Kohn, R. (2017). Efficient Bayesian estimation for flexible panel models for multivariate outcomes: Impact of life events on mental health and excessive alcohol consumption. *arXiv preprint arXiv:1706.03953*.

Gunawan, D., Hawkins, G. E., Tran, M. N., Kohn, R., & Brown, S. D. (2020). New estimation approaches for the hierarchical Linear Ballistic Accumulator model. *Journal of Mathematical Psychology*, *96*, 102368.

Harvey, N. (2007). Use of heuristics: Insights from forecasting research. *THINKING & REASONING*, *13*(1), 5–24. https://doi.org/10.1080/13546780600872502

Hendrickson, N. (2009). Applied Counterfactual Reasoning. In *Computational Methods for Counterterrorism* (pp. 249–262). Springer.

Hogarth, R. M., & Makridakis, S. (1981). FORECASTING AND PLANNING - AN EVALUATION. *MANAGEMENT SCIENCE*, *27*(2), 115–138. https://doi.org/10.1287/mnsc.27.2.115

Holmes, W. R. (2015). A practical guide to the Probability Density Approximation (PDA) with improved implementation and error characterization. *Journal of Mathematical Psychology*, *68*, 13–24.

Intelligence Advanced Research Projects Activity. (2018). Forecasting Counterfactuals in Uncontrolled Settings (FOCUS) Opportunity (Notice ID IARPA-BAA-17-08). Retrieved from https://beta.sam.gov/api/prod/opps/v3/opportunities/resources/files/ae28b95f47ea9c6ca2523281b2f05629/download?api_key=null&status=archived&token=

Johnson, J. G., & Busemeyer, J. R. (2010). Decision making under risk and uncertainty. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(5), 736–749.

Juvina, I., Larue, O., Widmer, C., Ganapathy, S., Nadella, S., Minnery, B., Ramshaw, L., Servan-Schreiber, E., Balick, M., & Weischedel, R. (2020a). Computer-supported collaborative information search for geopolitical forecasting. In Wai Tat Fu & Herre van Oostendorp (Eds.) *Understanding and Improving Information Search – A Cognitive Approach.* Human–Computer Interaction Series, Springer Nature. https://www.springer.com/gp/book/9783030388249

Juvina, I., Aue, W. R., Minnery, B., Hitzler, P., Nadella, S., & Sarker, M. K. (2020b). Counterfactual reasoning over large-scale human performance optimization experiments. Virtual poster presented at the annual meeting of the Psychonomic Society.

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis* (Vol. 344). John Wiley & Sons.

Kahneman, D., & Miller, D. T. (1986). Norm theory: Comparing reality to its alternatives. *Psychological Review*, *93*(2), 136.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the american statistical association*, *90*(430), 773-795.

Klein, G., Snowden, D., & Pin, C. L. (2011). Anticipatory thinking. *Informed by Knowledge: Expert Performance in Complex Situations*, 235–245.

Lee, M. D., Gluck, K. A., & Walsh, M. M. (2019). Understanding the complexity of simple decisions: Modeling multiple behaviors and switching strategies. *Decision*, *6*(4), 335.

Lewis, D. (1973). *Counterfactuals*. Blackwell.

Liaw, A. & Wiener, M., (2002). Classification and Regression by randomForest. *R News*, *2*(3), 18-22.

Luce, R. D. (1977). The choice axiom after twenty years. *Journal of mathematical psychology*, *15*(3), 215-233.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2021). cluster: Cluster Analysis Basics and Extensions. R package version 2.1.1.

Mahoney, L., Houpt, J., & Juvina, I. (2021). Explaining Task-Switching Behavior Using Evidence Accumulation Models. Presentation at 54nd Annual Meeting of Society for Mathematical Psychology, Virtual: Society for Mathematical Psychology.

Massey Jr, F. J. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, *46*(253), 68–78.

McElreath, R. (2020). *Statistical rethinking: A bayesian course with examples in R and Stan*.

Mehlhorn, K., Newell, B. R., Todd, P. M., Lee, M. D., Morgan, K., Braithwaite, V. A., Hausmann, D., Fiedler, K., & Gonzalez, C. (2015). Unpacking the exploration--exploitation tradeoff: A synthesis of human and animal literatures. *Decision*, *2*(3), 191.

Meiran, N. (1996). Reconfiguration of processing mode prior to task performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(6), 1423.

Mellers, B., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, S. E., Ungar, L., Bishop, M. M., Horowitz, M., Merkle, E., & Tetlock, P. (2015). The Psychology of Intelligence Analysis: Drivers of Prediction Accuracy in World Politics. *Journal of Experimental Psychology-Applied*, *21*(1), 1–14. https://doi.org/10.1037/xap0000040

Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., Chen, E., Baker, J., Hou, Y., Horowitz, M., Ungar, L., & Tetlock, P. (2015). Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions. *Perspectives on Psychological Science*, *10*(3), 267–281. https://doi.org/10.1177/1745691615577794

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., Scott, S. E., Moore, D., Atanasov, P., Swift, S. A., & others. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, *25*(5), 1106–1115.

Mesmer-Magnus, J., Lungeanu, A., Harris, A., Niler, A., DeChurch, L. A., & Contractor, N. (2020). Working in Space: Managing Transitions between Tasks. In *Psychology and Human Performance in Space Programs* (pp. 179–203). CRC Press.

Monsell, S. (2003). Task switching. *Trends in Cognitive Sciences*, *7*(3), 134–140. https://doi.org/https://doi.org/10.1016/S1364-6613(03)00028-7

Newcastle Cognition Lab (2021). *Particle Based Samplers for MCMC*. https://newcastlecl.github.io/samplerDoc/

Newell, B. R., & Shanks, D. R. (2003). Take the best or look at the rest? Factors influencing" one-reason" decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(1), 53.

Ngufor, Che. (2019). *nguforche/UnsupRF: Unsupervised Random Forest Clustering*. Github. https://rdrr.io/github/nguforche/UnsupRF/

Ohio Supercomputer Center. (1987). Ohio Supercomputer Center, Columbus OH. URL http://osc.edu/ark:/19495/f5s1ph73

Oppenheimer, D. M. (2003). Not so fast! (and not so frugal!): rethinking the recognition heuristic. *Cognition*, *90*(1), B1–B9. https://doi.org/https://doi.org/10.1016/S0010-0277(03)00141-0

Pacala, S. W., Gordon, D. M., & Godfray, H. C. J. (1996). Effects of social group size on information transfer and task allocation. *EVOLUTIONARY ECOLOGY*, *10*(2), 127–165. https://doi.org/10.1007/BF01241782

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(3), 534.

Pearl, J., & Mackenzie, D. (2018). *The book of why: the new science of cause and effect*. Basic Books.

Pirolli, P., & Card, S. (1999). Information foraging. *Psychological Review*, *106*(4), 643.

Pirolli, P., & Card, S. (2005). The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. *Proceedings of International Conference on Intelligence Analysis*, *5*, 2–4.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological methodology*, 111-163.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108. https://doi.org/10.1037/0033-295X.85.2.59

Rieskamp, J., & Otto, P. E. (2006). SSL: a theory of how people learn to select strategies. *Journal of Experimental Psychology: General*, *135*(2), 207.

Ritschard, G., & Studer, M. (2018). *Sequence analysis and related approaches: Innovative methods and applications*. Springer Nature.

Roe, R. M., Busemeyer, J. R., & Townsend, J. T. (2001). Multialternative decision field theory: A dynamic connectionst model of decision making. *Psychological Review*, *108*(2), 370.

Rosenthal, S., & Hiatt, L. M. (2020, May). Human-Centered Decision Support for Agenda Scheduling. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems* (pp. 1161-1168).

Sanderson, K. R. (2012). *Time orientation in organizations: Polychronicity and multitasking*. [Doctoral dissertation, Florida International University]. FIU Electronic Theses and Dissertations. https://digitalcommons.fiu.edu/etd/738

Scrucca L., Fop M., Murphy T. B., & Raftery A. E. (2016). mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, *8*(1), 289-317.

Shi, T., & Horvath, S. (2006). Unsupervised learning with random forest predictors. *Journal of Computational and Graphical Statistics*, *15*(1), 118–138.

Shlens, J. (2014). A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*.

Simon, H. A. (1956). Rational choice and the structure of the environment. *Psychological Review*, *63*(2), 129.

Singmann, H., Brown, S., Gretton, M., & Heathcote, A. (2020). rtdists: Response Time Distributions. R package version 0.11-2. https://CRAN.R-project.org/package=rtdists

Smallman, R., & Summerville, A. (2018). Counterfactual thought in reasoning and performance. *Social and Personality Psychology Compass, 12*(4), e12376.

Tetlock, P. E. (2005). *Expert political judgment: How good is it? How can we know?* Princeton, NJ: Princeton University Press.

Townsend, J. T., & Nozawa, G. (1995). Spatio-temporal properties of elementary perception: An investigation of parallel, serial, and coactive theories. *Journal of Mathematical Psychology, 39*(4), 321-359.

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

Trueblood, J., & Dasari, A. (2017). The Impact of Presentation Order on the Attraction Effect in Decision-making. In *CogSci*.

Trump, D. J. (2017). *National security strategy of the United States of America*. Executive Office of The President Washington DC Washington United States.

Turner, B. M., Schley, D. R., Muller, C., & Tsetsos, K. (2018). Competing theories of multialternative, multiattribute preferential choice. *Psychological Review*, *125*(3), 329.

Turner, B. M., Sederberg, P. B., Brown, S. D., & Steyvers, M. (2013). A method for efficiently sampling from distributions with correlated dimensions. *Psychological Methods*, *18*(3), 368.

Turner, B. M., Steyvers, M., Merkle, E. C., Budescu, D. V, & Wallsten, T. S. (2014). Forecast aggregation via recalibration. *Machine Learning*, *95*(3), 261–289. https://doi.org/10.1007/s10994-013-5401-4

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: the leaky, competing accumulator model. *Psychological Review*, *108*(3), 550.

Usher, M., & McClelland, J. L. (2004). Loss aversion and inhibition in dynamical models of multialternative choice. *Psychological Review*, *111*(3), 757.

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T., & Gelman, A. (2020). *loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models*. https://mc-stan.org/loo

Visser, I., & Poessé, R. (2017). Parameter recovery, bias and standard errors in the linear ballistic accumulator model. *British Journal of Mathematical and Statistical Psychology*, *70*(2), 280-296.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic bulletin & review*, *14*(5), 779-804.

Watanabe, S., & Opper, M. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of machine learning research*, *11*(12).

Wickens, C. D., Santamaria, A., & Sebok, A. (2013). A Computational Model of Task Overload Management and Task Switching. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *57*(1), 763–767. https://doi.org/10.1177/1541931213571167

Wickens, C. D., Gutzwiller, R. S., & Santamaria, A. (2015). Discrete task switching in overload: A meta-analyses and a model. *International Journal of Human-Computer Studies*, *79*, 79–84.

Wübben, M., & Wangenheim, F. v. (2008). Instant customer base analysis: Managerial heuristics often “get it right.” *Journal of Marketing*, *72*(3), 82–93.

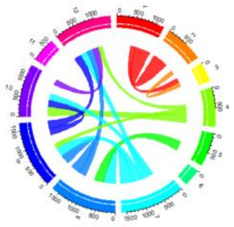
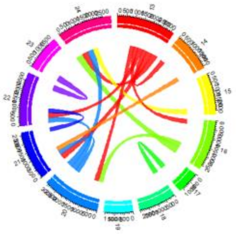
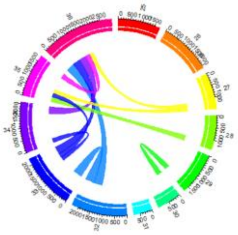
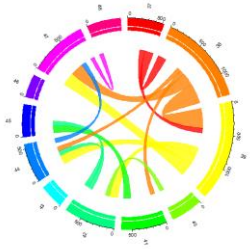
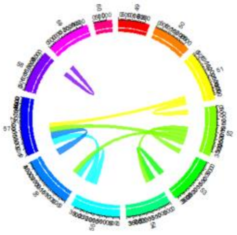
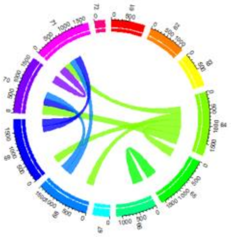
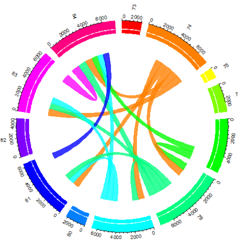
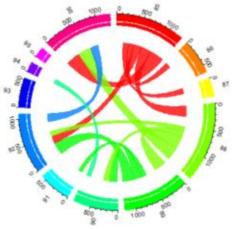
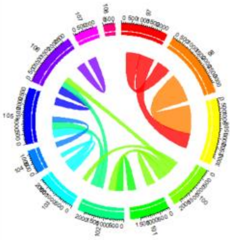
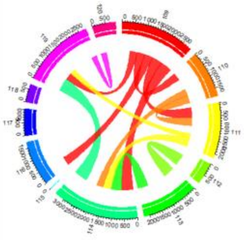
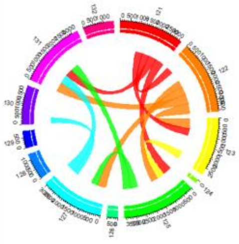
Wylie, G., & Allport, A. (2000). Task switching and the measurement of “switch costs.” *Psychological Research*, *63*(3–4), 212–233.

Yi, S. K. M., Steyvers, M., Lee, M. D., & Dry, M. J. (2012). The Wisdom of the Crowd in Combinatorial Problems. *Cognitive Science*, *36*(3), 452–470. https://doi.org/10.1111/j.1551-6709.2011.01223.x

Appendix A

Table 36. Pearson's r correlation values (n=240) for each factor with ID and switch rate.

	ID	Switch rate
ID	1	-0.0758
Switch rate	-0.076	1
Role	0.050	0.187
HERA	-0.003	-0.163
Campaign	**0.794**	-0.096
Mission	**0.986**	-0.084
Session	0.193	-0.103
Gender	-0.135	0.004
Comm delay	-0.106	-0.115
Extraversion	**0.777**	-0.156
Agreeableness	**0.769**	**-0.251**
Conscientiousness	**0.771**	-0.138
Neuroticism	**0.678**	-0.195
Mean task salience	-0.052	**-0.295**
Mean task interest	0.157	**0.207**
Mean task priority	-0.065	**0.422**
Mean task difficulty	0.102	**0.292**
Mean task category	-0.194	-0.092
Mean task structure	0.087	-0.064
Mean task value	0.222	0.091
Outcome	**0.276**	0.153
Mean total time on task	-0.180	**-0.732**
Explorer	-0.167	**0.465**
Typical task switches	0.116	**0.743**
Percent vertical switches	0.080	-0.119
Mean tool visibility	-0.272	0.011
Mean tool persistence	**0.602**	-0.128
Mean tool editability	0.034	0.065
Mean tool association	-0.413	0.012
Task mental model role 1	-0.051	0.118
Task mental model role 2	0.117	0.036
Task mental model role 3	0.079	0.000
Task mental model role 4	0.104	0.078
Task mental model role 5	0.092	0.038
Task mental model role 6	0.101	-0.001
Task mental model role 7	0.153	-0.087
Task mental model role 8	**0.225**	-0.038
Task mental model role 9	-0.003	-0.093
Task mental model role 10	-0.011	-0.061
Task mental model role 11	0.036	-0.054
Task mental model role 12	-0.114	0.139
Team mental model role 1	-0.088	0.076
Team mental model role 2	-0.003	-0.062
Team mental model role 3	-0.008	-0.051
Team mental model role 4	0.053	0.003
Team mental model role 5	0.049	-0.048
Team mental model role 6	0.001	-0.077
Team mental model role 7	-0.072	-0.112
Team mental model role 8	**0.300**	-0.023
Team mental model role 9	-0.067	-0.160
Team mental model role 10	0.032	-0.077
Team mental model role 11	0.112	-0.144
Team mental model role 12	**-0.252**	0.127
Behavioral tie role 1	0.107	-0.102
Behavioral tie role 2	-0.085	-0.094
Behavioral tie role 3	0.001	-0.109
Behavioral tie role 4	0.113	0.038
Behavioral tie role 5	0.093	-0.051
Behavioral tie role 6	0.132	0.038
Behavioral tie role 7	0.010	0.152
Behavioral tie role 8	0.083	0.109
Behavioral tie role 9	-0.042	0.049
Behavioral tie role 10	0.033	0.014
Behavioral tie role 11	0.087	-0.032
Behavioral tie role 12	0.101	0.003
Functional tie role 1	**0.240**	-0.101
Functional tie role 2	0.060	-0.116
Functional tie role 3	0.057	-0.089
Functional tie role 4	0.122	0.078
Functional tie role 5	0.162	0.034
Functional tie role 6	0.056	-0.133
Functional tie role 7	0.167	-0.083
Functional tie role 8	**0.254**	-0.017
Functional tie role 9	0.113	-0.037
Functional tie role 10	0.095	0.029
Functional tie role 11	0.119	0.027
Functional tie role 12	0.144	0.064
Interpersonal tie role 1	0.005	-0.168
Interpersonal tie role 2	-0.063	-0.108
Interpersonal tie role 3	-0.007	-0.097
Interpersonal tie role 4	0.042	0.116
Interpersonal tie role 5	0.017	0.002
Interpersonal tie role 6	-0.023	0.057
Interpersonal tie role 7	-0.006	**0.226**
Interpersonal tie role 8	-0.088	0.138
Interpersonal tie role 9	-0.099	0.166
Interpersonal tie role 10	0.065	0.024
Interpersonal tie role 11	0.084	0.008
Interpersonal tie role 12	0.036	-0.008
Tool tie role 1	0.379	0.006
Tool tie role 2	0.141	-0.052
Tool tie role 3	0.183	-0.019
Tool tie role 4	**0.356**	0.032
Tool tie role 5	**0.316**	0.002
Tool tie role 6	0.175	-0.149
Tool tie role 7	0.165	-0.122
Tool tie role 8	**0.272**	-0.055
Tool tie role 9	0.162	-0.058
Tool tie role 10	0.173	-0.001
Tool tie role 11	**0.228**	0.008
Tool tie role 12	0.128	0.054
MLBA threshold	0.616	-0.127
MLBA starting point	0.335	**-0.266**
MLBA attention weight	-0.519	0.178
MLBA subjective value	0.376	0.034
MLBA non-decision time	-0.585	0.168

	Session 1	Session 2	Session 3	Session 4
Mission 2				
Mission 3				
Mission 4				
Mission 5				
Mission 6				

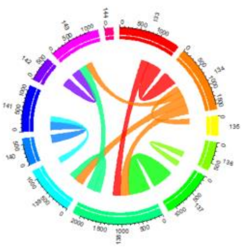
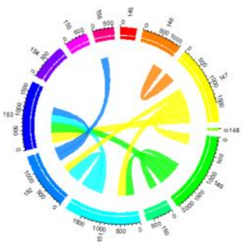
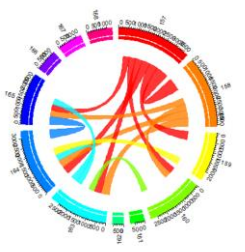
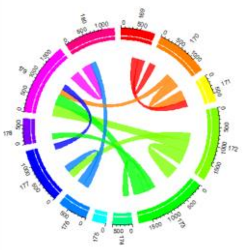
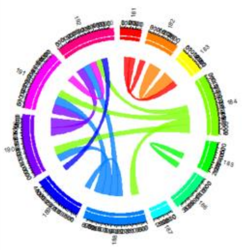
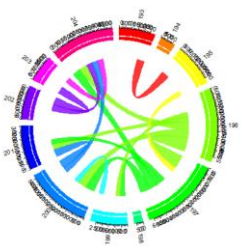
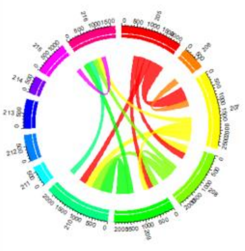
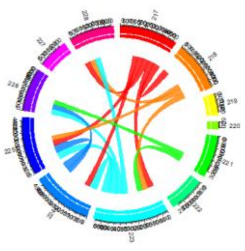
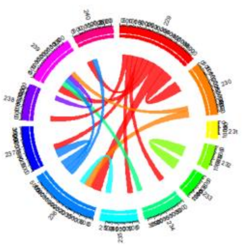
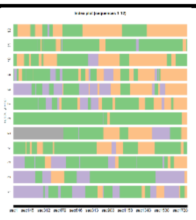
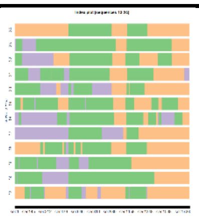
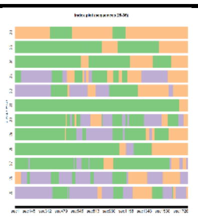
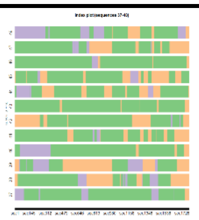
Mission 7				
Mission 8				
Mission 9				

Figure 73. Task dependency plots for ‘not withheld’ data. Blank missions and sessions are withheld data.

	Session 1	**Session 2**	**Session 3**	**Session 4**
Mission 2				
Mission 3				

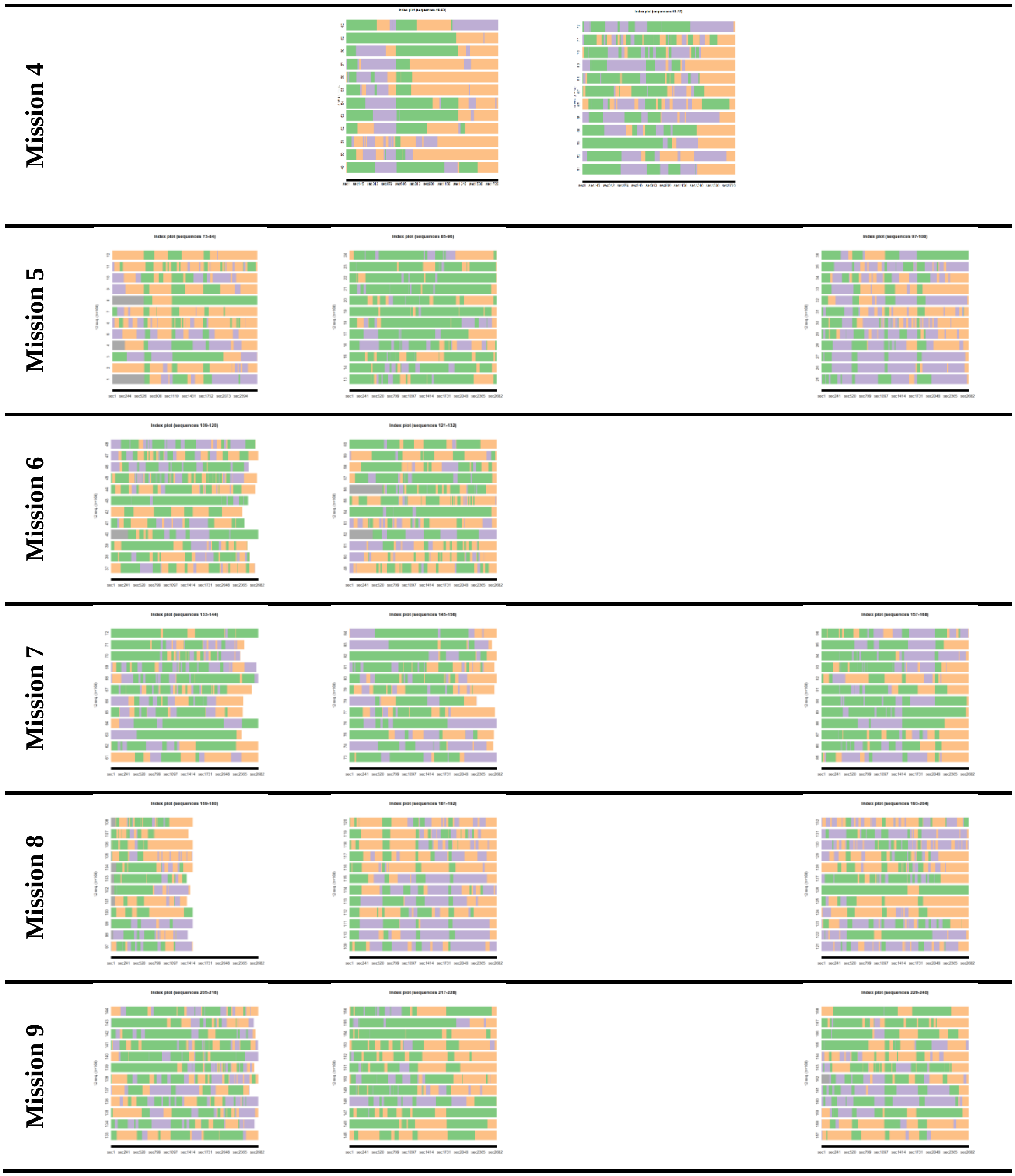


Figure 74. Sequence plots of individual (green), team (purple) and MTS (orange) tasks

Appendix B

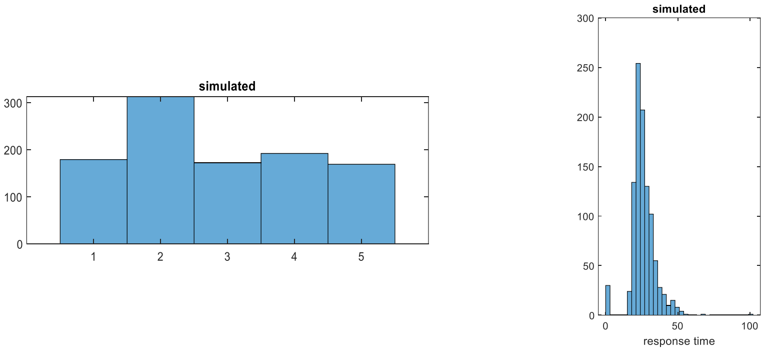


Figure 75. Example choice and response time distributions for simple simulated data

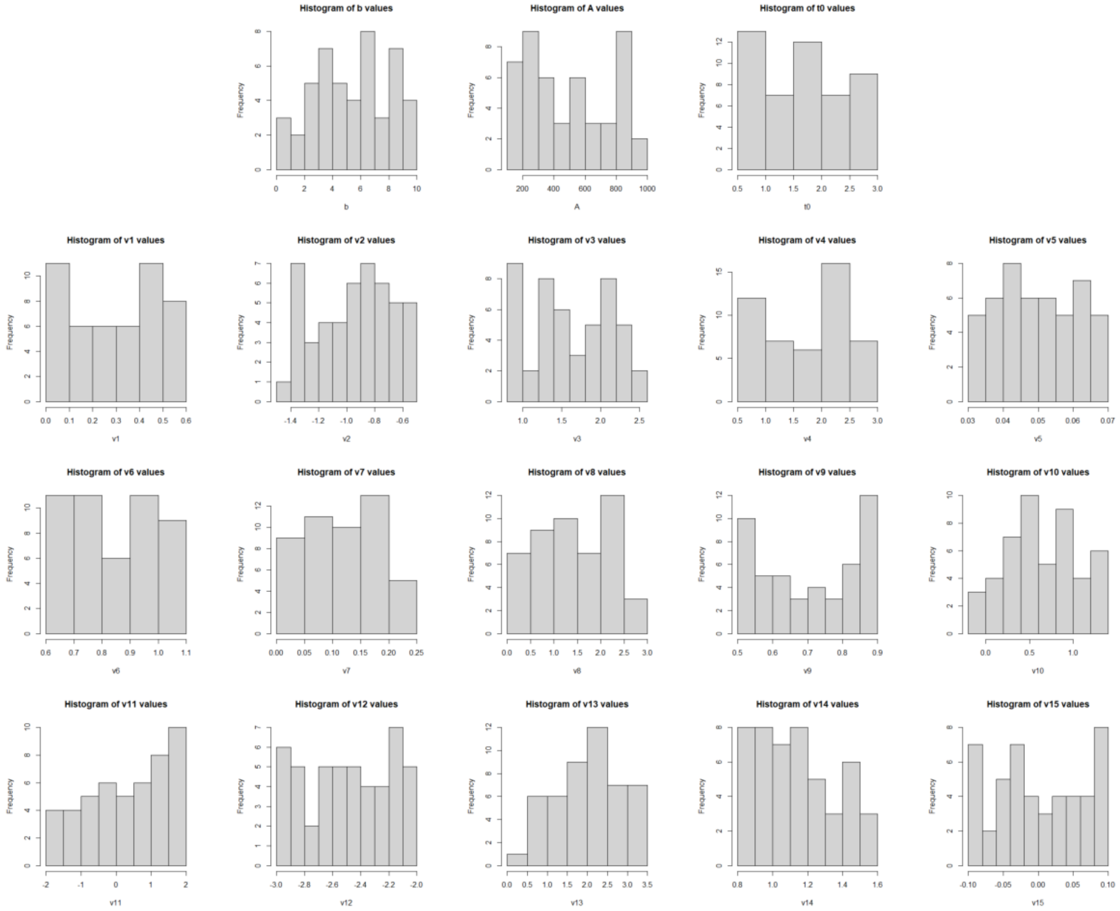


Figure 76. Histograms of actual parameter values in 48-subject simulated data

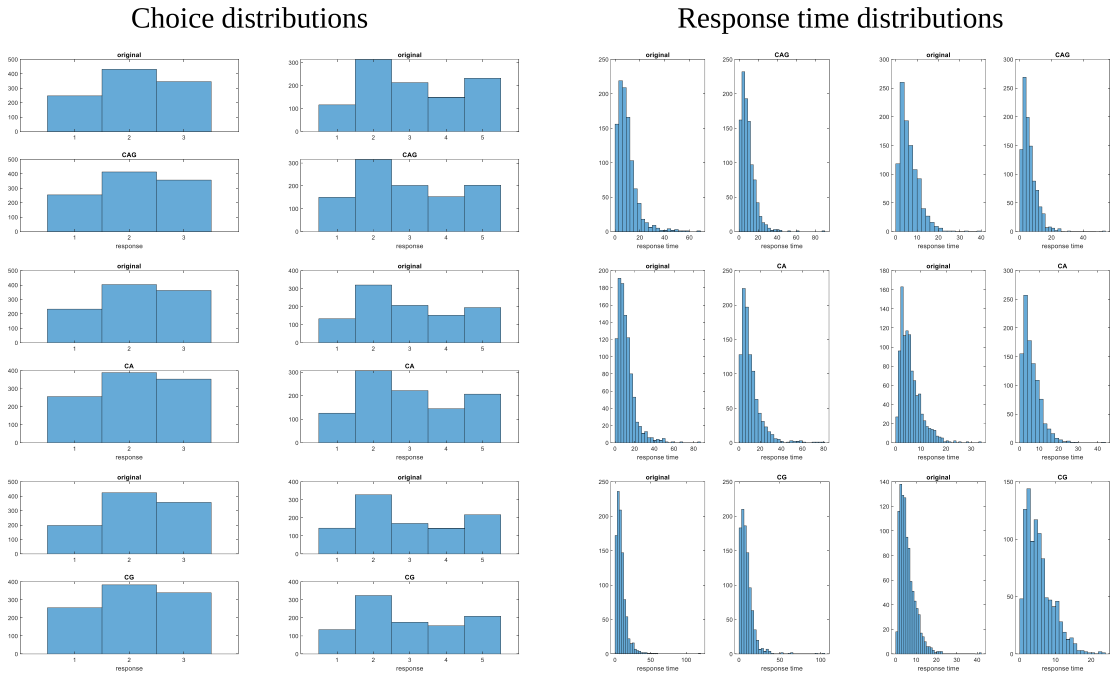


Figure 77. Sample of choice and response time distributions for CAG (top row), CA (middle row) and CG (bottom row) models.

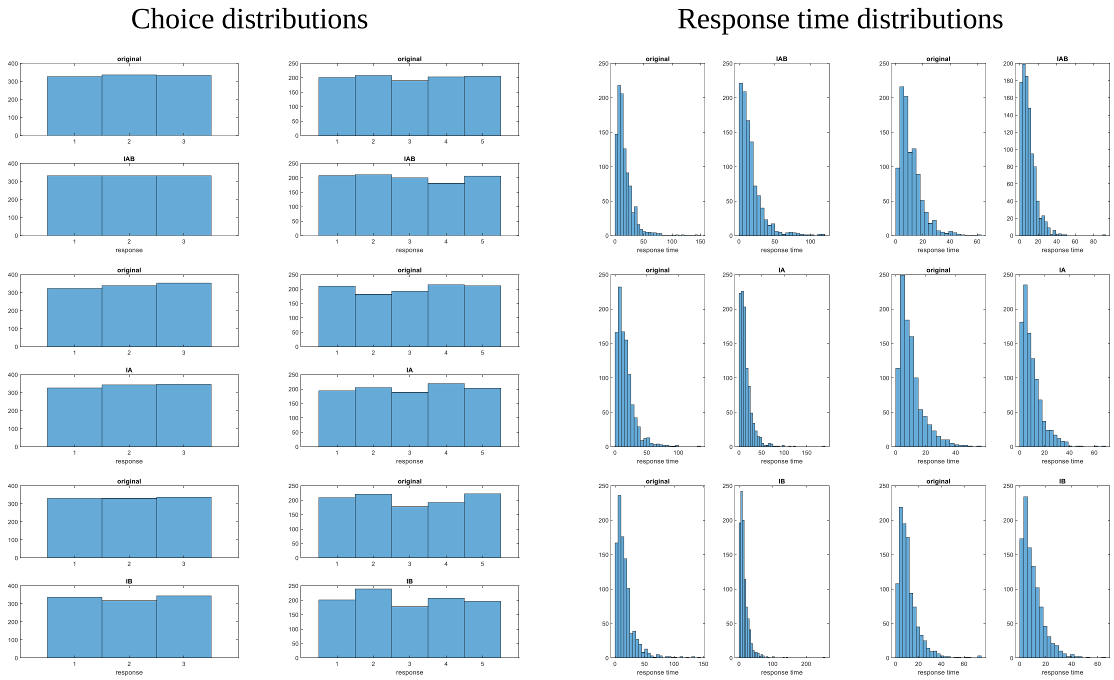


Figure 78. Sample of choice and response time distributions for IAB (top row), IA (middle row) and IB (bottom row) models.

Table 37. Parameter recovery – MLE fit of single-stage CPT models

Parameter	Initial Value		CAG		CA		CG	
	3-choice	5-choice	3-choice	5-choice	3-choice	5-choice	3-choice	5-choice
b	2		1.6624	1.9427	1.3905	1.7034	1.1806	1.5273
			1.5875	1.9035	1.3244	1.6660	1.7300	1.4668
			1.8699	2.0085	1.5726	1.7637	1.9332	1.6366
			1.3929	1.8440	1.1411	1.6325	1.0423	1.5538
			1.6998	1.8133	1.4163	1.5961	1.7580	1.5797
			1.7893	1.6572	1.4888	1.4679	1.7411	1.5486
			1.5347	1.9162	1.3123	1.6893	1.3969	1.5645
			1.5783	1.9301	1.2894	1.7215	1.2616	1.7698
			1.6620	1.8922	1.3624	1.6274	1.5498	1.4563
			1.5384	1.8231	1.2914	1.6012	1.4268	1.5844
			1.7937	1.8986	1.5220	1.6631	2.0138	1.9230
			1.6977	1.9700	1.3994	1.7411	1.9553	1.9597
			1.6460	1.7422	1.3546	1.5239	1.3613	1.9534
			1.7193	1.9228	1.4469	1.6759	1.8668	1.5874
			1.6624	1.8957	1.3916	1.6748	1.4002	2.0096
			1.6290	1.8745	1.3538	1.6431	1.6855	1.3240
			1.6372	1.8449	1.3696	1.6047	1.2339	1.6619
			1.4099	1.8749	1.1983	1.6220	1.2647	1.5879
			1.8243	1.8949	1.4993	1.6568	1.7859	1.7695
			1.6847	1.7887	1.4258	1.5545	1.9325	1.4545
α	1.2296	1.0804	1.2212	1.1928	1.0849	1.0819	NA	NA
	0.8246	1.1271	0.9313	1.2035	0.7709	1.0942		
	0.8771	1.0506	0.9745	1.1601	0.8231	1.0483		
	1.2146	1.0308	1.1960	1.1352	1.0584	1.0233		
	0.8755	1.0068	0.9748	1.1092	0.8173	0.9953		
	0.9258	0.9339	1.0211	1.0562	0.8668	0.9390		
	1.0249	1.0325	1.0663	1.1617	0.9270	1.0515		
	1.1286	0.9425	1.1594	1.0704	1.0161	0.9577		
	1.0339	1.1768	1.0524	1.2092	0.8943	1.0946		
	1.0154	0.9904	1.0532	1.1134	0.8967	0.9995		
	0.8150	0.8137	0.9065	0.9890	0.7515	0.8634		
	0.7560	0.8065	0.8864	1.0044	0.7122	0.8853		
	1.1471	0.7556	1.1313	0.8961	0.9855	0.7622		
	0.8328	1.0583	0.9344	1.1537	0.7767	1.0407		
	1.0770	0.7770	1.1181	0.9504	0.9783	0.8262		
	0.9753	1.1684	0.9736	1.2743	0.8140	1.1706		
	1.2067	0.9810	1.1958	1.0835	1.0573	0.9616		
	1.0192	1.0720	1.0780	1.1336	0.9378	1.0147		
	0.9713	0.9155	1.0155	1.0559	0.8567	0.9364		
	0.7523	1.1051	0.8892	1.1648	0.7284	1.0494		
γ			1.0e-04 *	1.0e-04 *	NA	NA	1.0e-04 *	
	1.5944	2.1760	0.1000	0.1001			0.1004	0.0000
	1.0150	1.5987	0.1000	0.1000			0.1000	0.0000
	2.1286	1.3599	0.1001	0.1002			0.1000	0.0000
	1.2000	1.4539	0.1000	0.1000			0.1001	0.0000
	1.7321	2.4245	0.1000	0.1000			0.1000	0.0000
	2.1617	1.0418	0.1001	0.1001			0.1000	0.0000
	2.3344	2.1110	0.1000	0.1002			0.1002	0.0000
	2.0075	2.4452	0.1000	0.1002			0.1002	0.0000
	0.6517	0.6183	0.1000	0.1000			0.1000	0.0000
	2.0583	1.3158	0.1000	0.1002			0.1000	0.0000
	1.6376	1.5144	0.1000	0.1000			0.1000	0.0000
	1.1742	2.1401	0.1000	0.1002			0.1000	0.0000
	1.1224	0.5614	0.1000	0.1002			0.1000	0.0000
	1.7040	1.2654	0.1001	0.1000			0.1000	0.0000
	1.8784	1.6790	0.1001	0.1000			0.1000	0.0000
	0.6676	2.1789	0.1000	0.1001			0.1000	0.2539
	0.8048	0.5181	0.1000	0.1000			0.1003	0.0000
	2.4923	0.8074	0.1000	0.1000			0.1002	0.0000
	0.7133	1.8735	0.1000	0.1002			0.1000	0.0000
	2.0498	1.2309	0.1001	0.1000			0.1000	0.0000

Table 38. Parameter recovery – MLE fit of single-stage IF model

Parameter	Initial Value		IAB		IA		IB	
	3-choice	5-choice	3-choice	5-choice	3-choice	5-choice	3-choice	5-choice
b	2		0.9833	1.1434	0.9834	1.1434	0.9840	1.1434
			1.1066	1.1094	1.1066	1.1090	1.1065	1.1091
			1.1467	0.9836	1.1467	0.9831	1.1467	0.9827
			1.0216	0.9657	1.0206	0.9657	1.0206	0.9657
			1.1687	1.0334	1.1682	1.0330	1.1686	1.0287
			0.9579	1.0733	0.9572	1.0730	0.9570	1.0736
			1.0778	1.0576	1.0778	1.0560	1.0778	1.0573
			1.0134	0.9924	1.0134	0.9922	1.0133	0.9908
			1.0310	1.0986	1.0313	1.0986	1.0308	1.0986
			0.9993	0.9722	0.9986	0.9720	0.9986	0.9722
			1.0655	1.1896	1.0655	1.1896	1.0655	1.1896
			1.2125	1.0797	1.2125	1.0795	1.2125	1.0795
			0.9254	1.1314	0.9254	1.1314	0.9254	1.1314
			1.0656	1.1390	1.0656	1.1375	1.0655	1.1369
			1.1338	1.0064	1.1331	1.0062	1.1324	1.0064
			1.1104	0.9602	1.1104	0.9602	1.1104	0.9602
			1.0073	1.1880	1.0066	1.1878	1.0080	1.1871
			1.1580	1.0751	1.1580	1.0751	1.1580	1.0751
			0.9626	0.9860	0.9626	0.9867	0.9626	0.9861
			0.9836	0.9971	0.9837	0.9971	0.9836	0.9971
α	0.8418	1.1152	10.8694	0.0927	3.1508	1.5059	NA	NA
	1.1401	0.8686	0.0256	267.0630	0.2203	4.4221		
	1.1379	1.0234	27.1561	267.9623	0.0000	4.8721		
	0.9734	0.9944	10.3527	252.0445	0.8666	0.0000		
	1.0054	0.9478	0.1504	268.4042	1.3793	268.3970		
	1.0722	0.7689	13.4878	266.9652	2.5632	268.1373		
	1.0164	1.1481	265.8533	0.1402	0.0000	2.4492		
	1.1880	0.9177	0.0379	29.5865	0.1178	16.0771		
	1.0435	1.1106	268.5585	266.4918	2.0632	0.0001		
	0.9855	0.9971	244.0127	0.2007	0.0000	3.3074		
	0.8474	1.2019	266.7907	267.9481	0.0007	0.0000		
	0.8638	1.0994	268.3310	21.9265	0.0000	6.4484		
	1.2117	1.1220	268.5020	268.1140	0.0000	0.0001		
	1.2024	1.2024	268.3709	265.8508	0.0014	2.1831		
	0.8056	1.1797	7.3002	0.1674	4.3390	3.4128		
	1.0474	0.8415	261.6736	268.3512	0.0000	0.0000		
	1.1056	0.7643	263.9928	6.8635	4.6193	5.0366		
	0.8983	1.2393	1.2647	0.0629	0.8785	1.6056		
	1.0039	0.9855	0.1002	0.1194	0.9568	2.3222		
	1.1505	0.7712	0.1737	267.8688	1.5529	0.0001		
β	1.2370	1.4772	1.6096	0.0007	NA	NA	0.7435	1.0862
	0.6623	1.4177	0.0146	1.9020			2.1421	0.8451
	1.4736	1.5423	7.1952	1.8534			2.3507e+151	1.4849
	1.1127	1.7481	2.1372	863.3434			1.2808	2.9281e+151
	2.1353	1.2349	0.0004	1.4716			0.9686	0.6226
	1.2572	2.2703	1.7228	1.6266			0.8282	1.0332
	1.2015	0.6974	463.4253	0.0003			431.0520	0.8770
	1.6003	1.8595	0.5662	1.5769			2.7578	0.5901
	0.9155	0.7135	1.8321	630.8570			0.9516	9.3403e+128
	0.9610	2.0581	2.0688	0.2259			10.0152	0.9156
	0.9518	2.2818	138.3185	220.6821			1.6975e+151	6.5171e+59
	1.3714	0.8956	422.8069	1.8362			5.2223e+99	1.0182
	1.3604	1.5000	666.0909	270.2281			4.4543e+151	2.6442e+17
	2.4595	1.7197	278.2250	1.9541			5473178.6948	1.3775
	1.0161	2.1110	1.3910	0.0004			0.6000	0.7834
	1.0244	0.9799	138.4273	638.8824			1.0031e+85	4.1289e+151
	0.9435	1.4798	1.6087	1.6376			0.7205	0.8797
	1.1376	1.9254	1.3763	0.0012			1.2635	1.2125
	0.6710	0.6192	0.0012	0.0006			1.1806	0.9407
	0.5584	0.6429	0.0003	628.3058			0.8943	8.2096

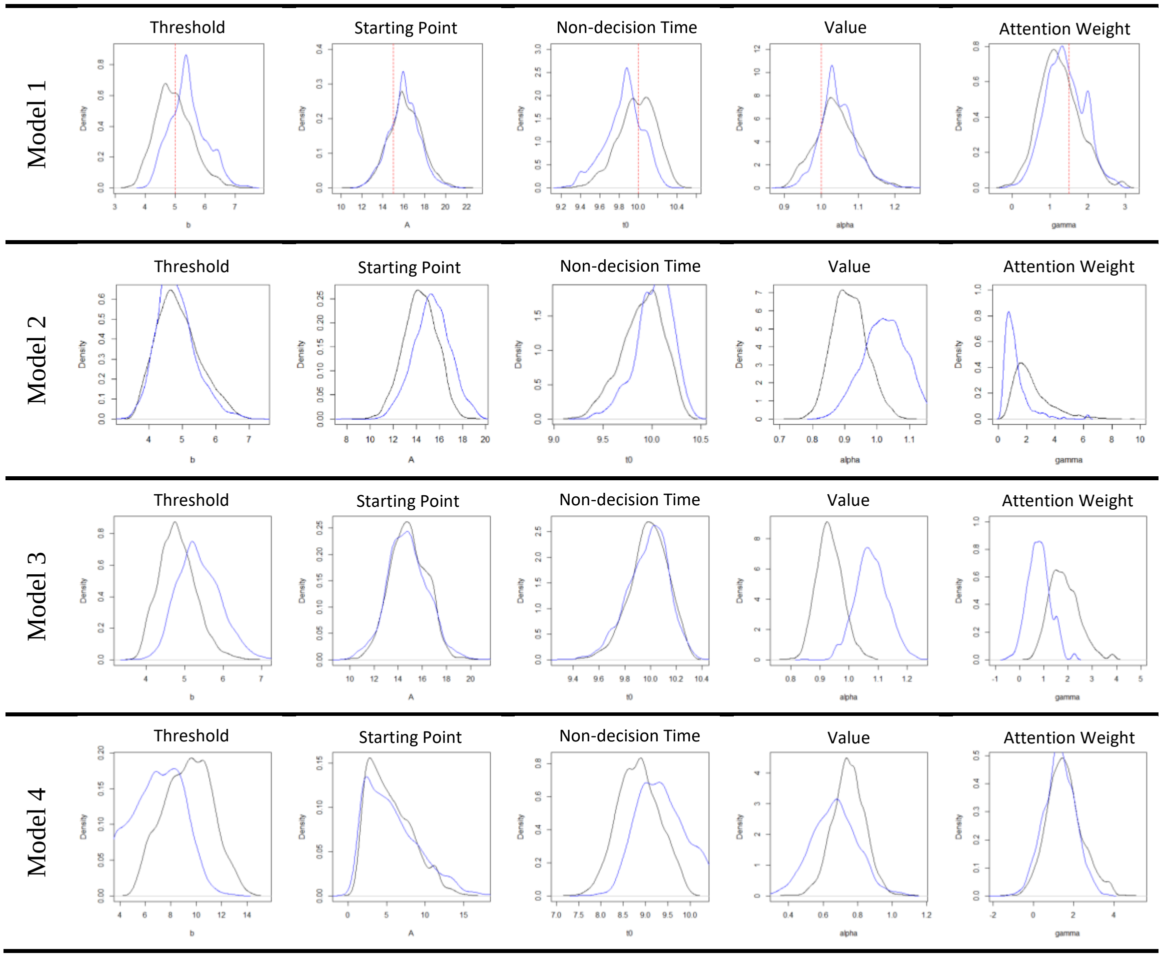
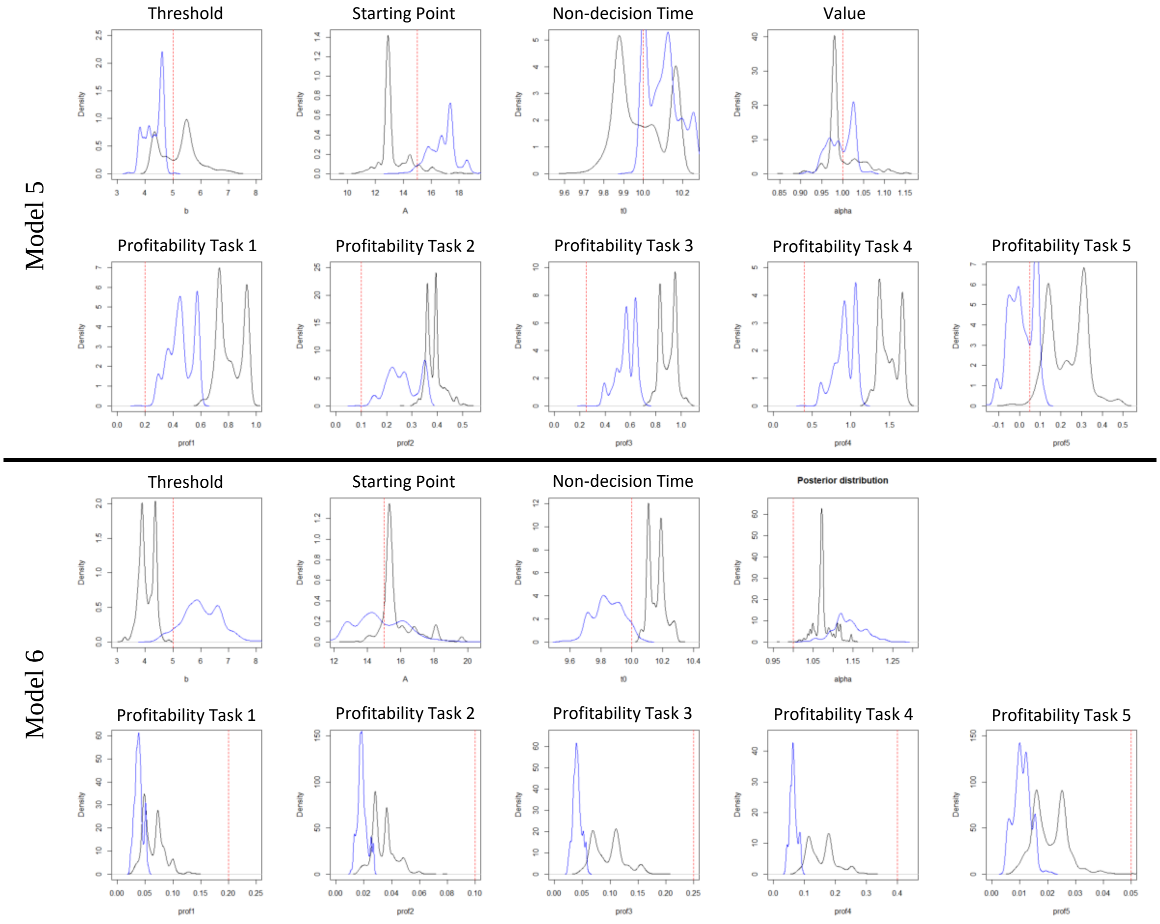


Figure 79. Model investigation - PMwG fit of single-stage CPT models. Red line indicates generating value.


Model 5
Threshold
Starting Point
Non-decision Time
Value
Density
b
A
t0
alpha
Profitability Task 1
Profitability Task 2
Profitability Task 3
Profitability Task 4
Profitability Task 5
prof1
prof2
prof3
prof4
prof5
Model 6
Threshold
Starting Point
Non-decision Time
Posterior distribution
Density
b
A
t0
alpha
Profitability Task 1
Profitability Task 2
Profitability Task 3
Profitability Task 4
Profitability Task 5
prof1
prof2
prof3
prof4
prof5

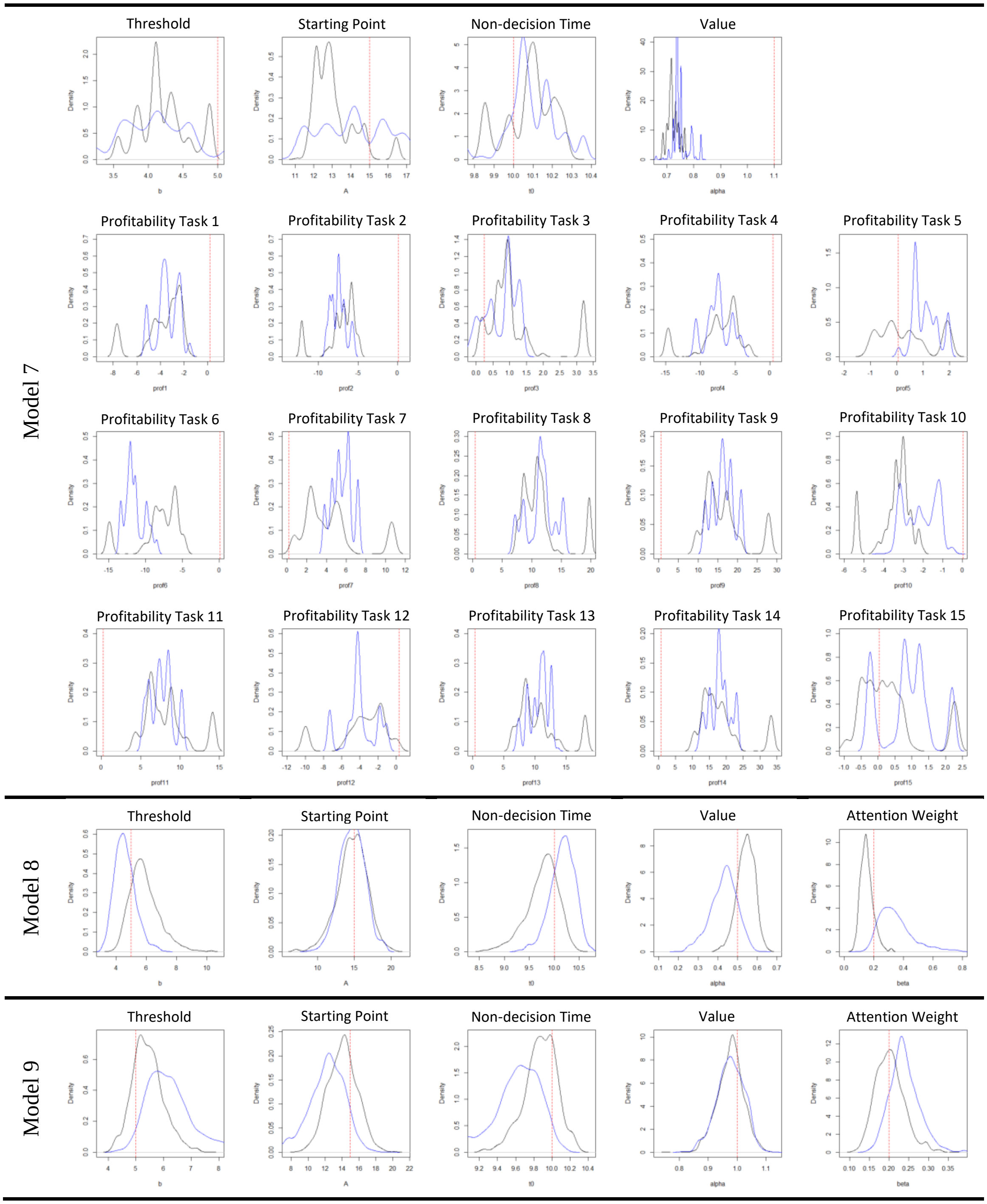
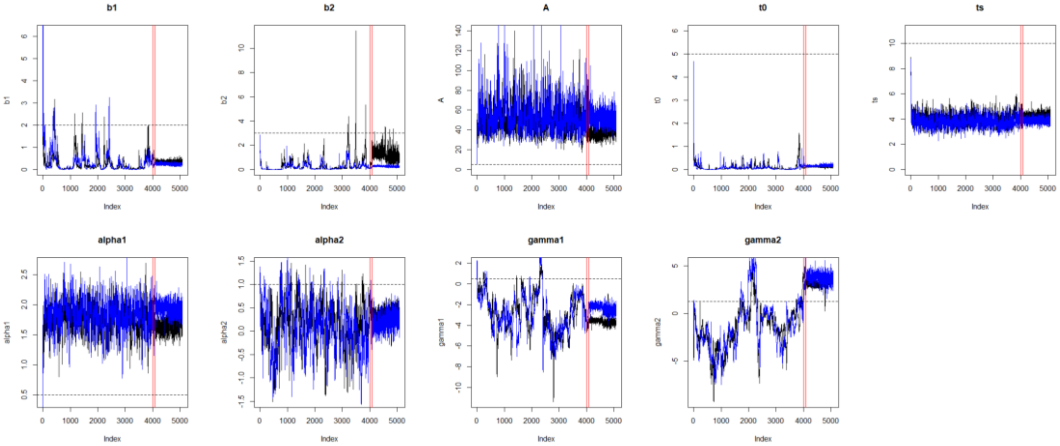
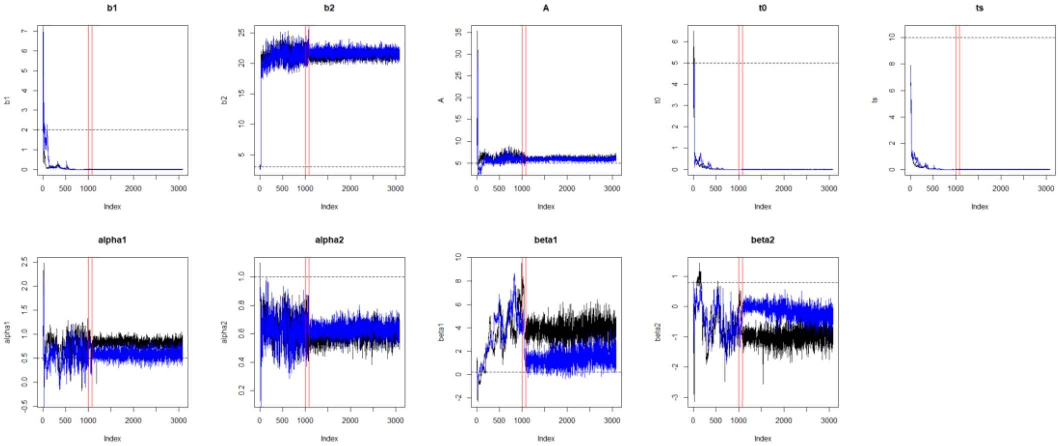


Figure 80. Model investigation - PMwG fit of single-stage IF models. Red line indicates generating value.

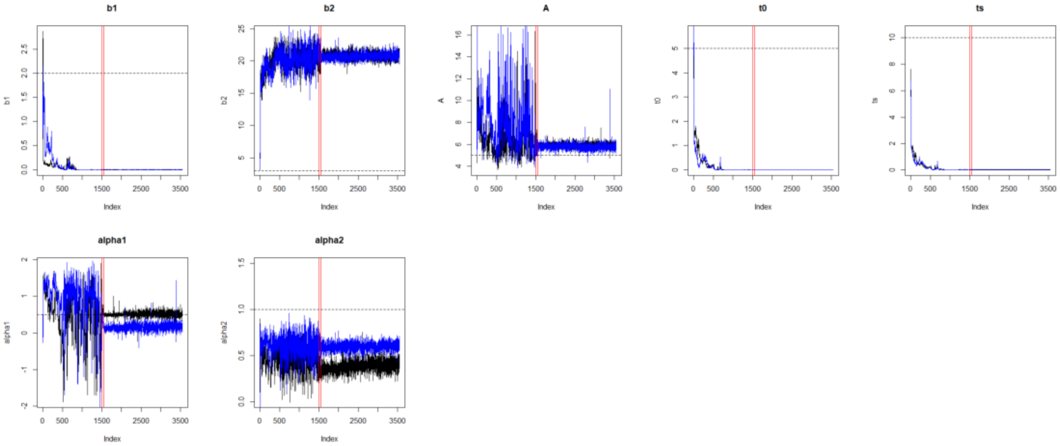
Model T1



Model T2



Model T3



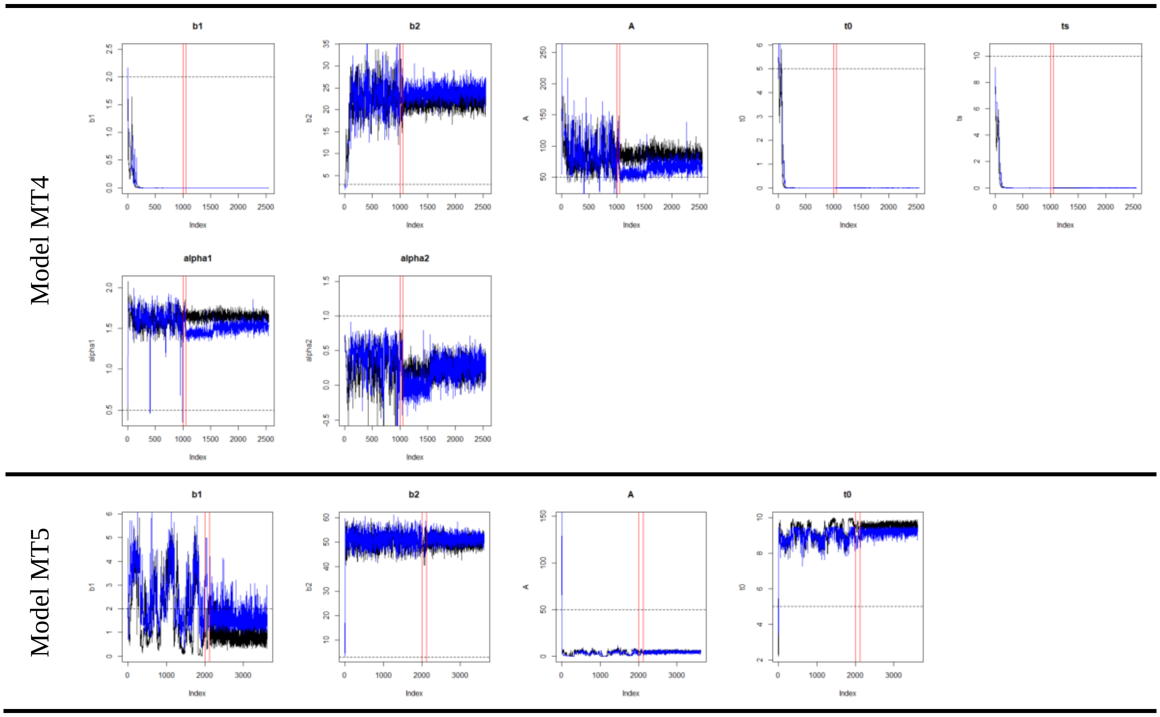


Figure 81. Model investigation - PMwG trace plots of initial structure two-stage models. Horizontal dotted line indicates generating values. Vertical red lines differentiate burn-in, adaptation, and sampling phases.

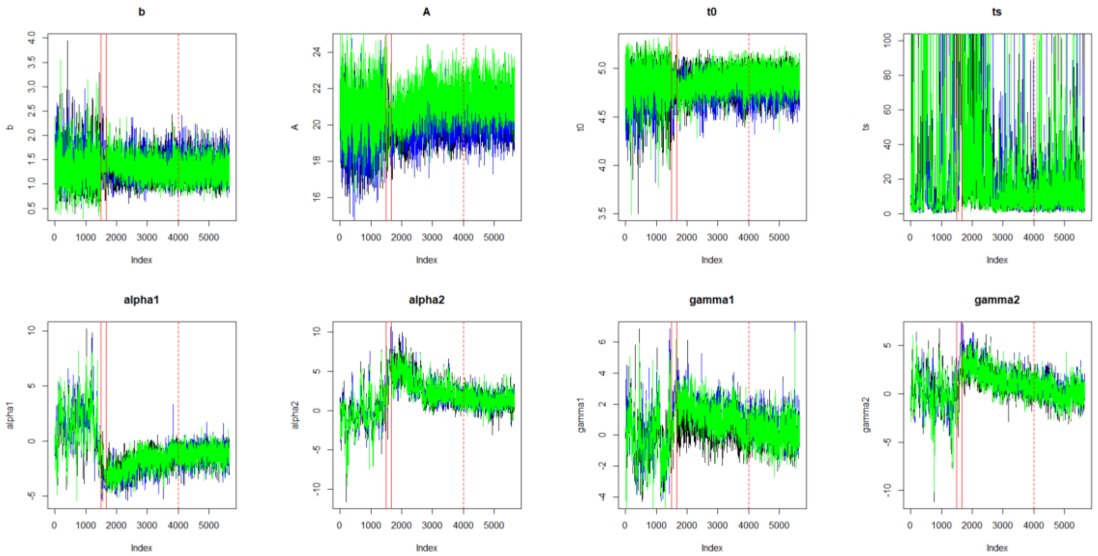
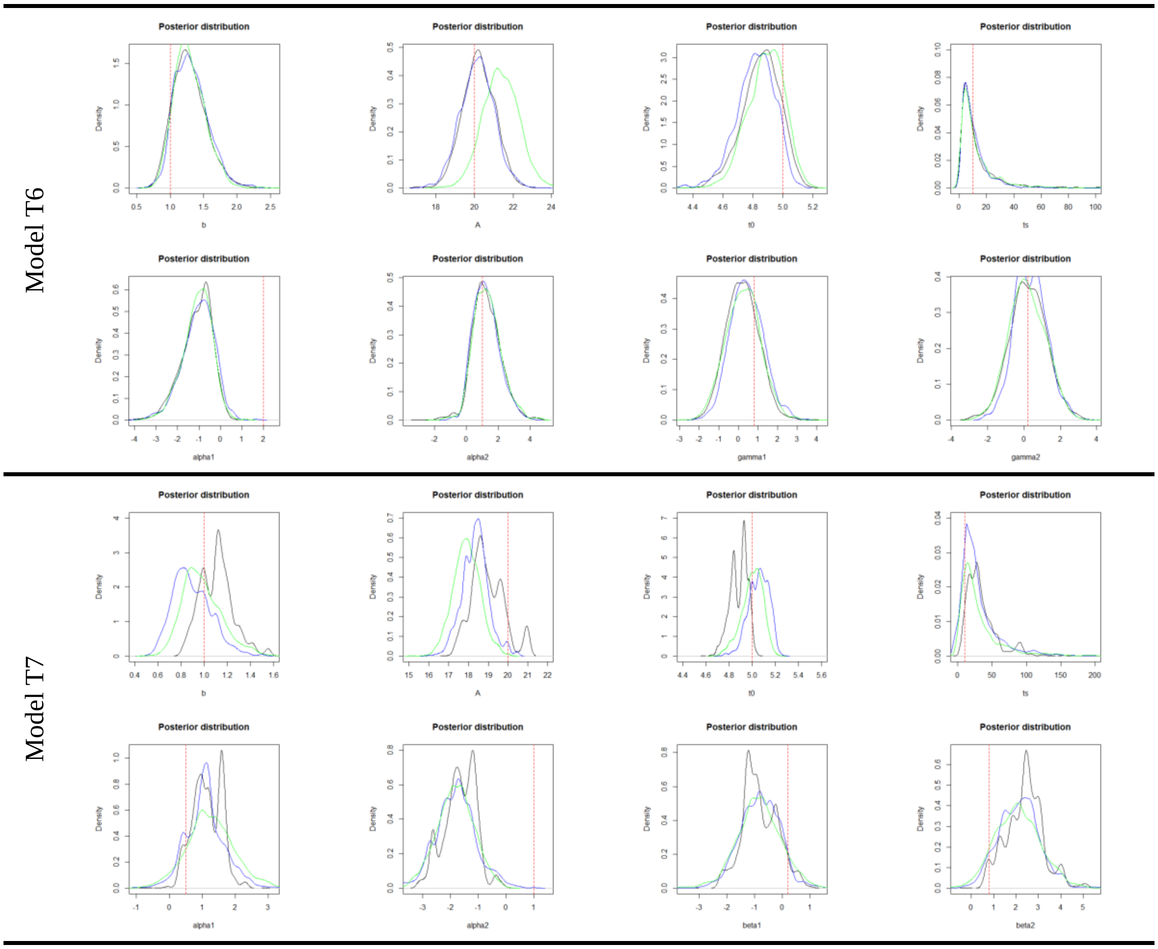


Figure 82. Model investigation - PMwG trace plots of revised structure two-stage model T6. Horizontal dotted line indicates generating values. Vertical red lines differentiate burn-in, adaptation, and sampling phases.



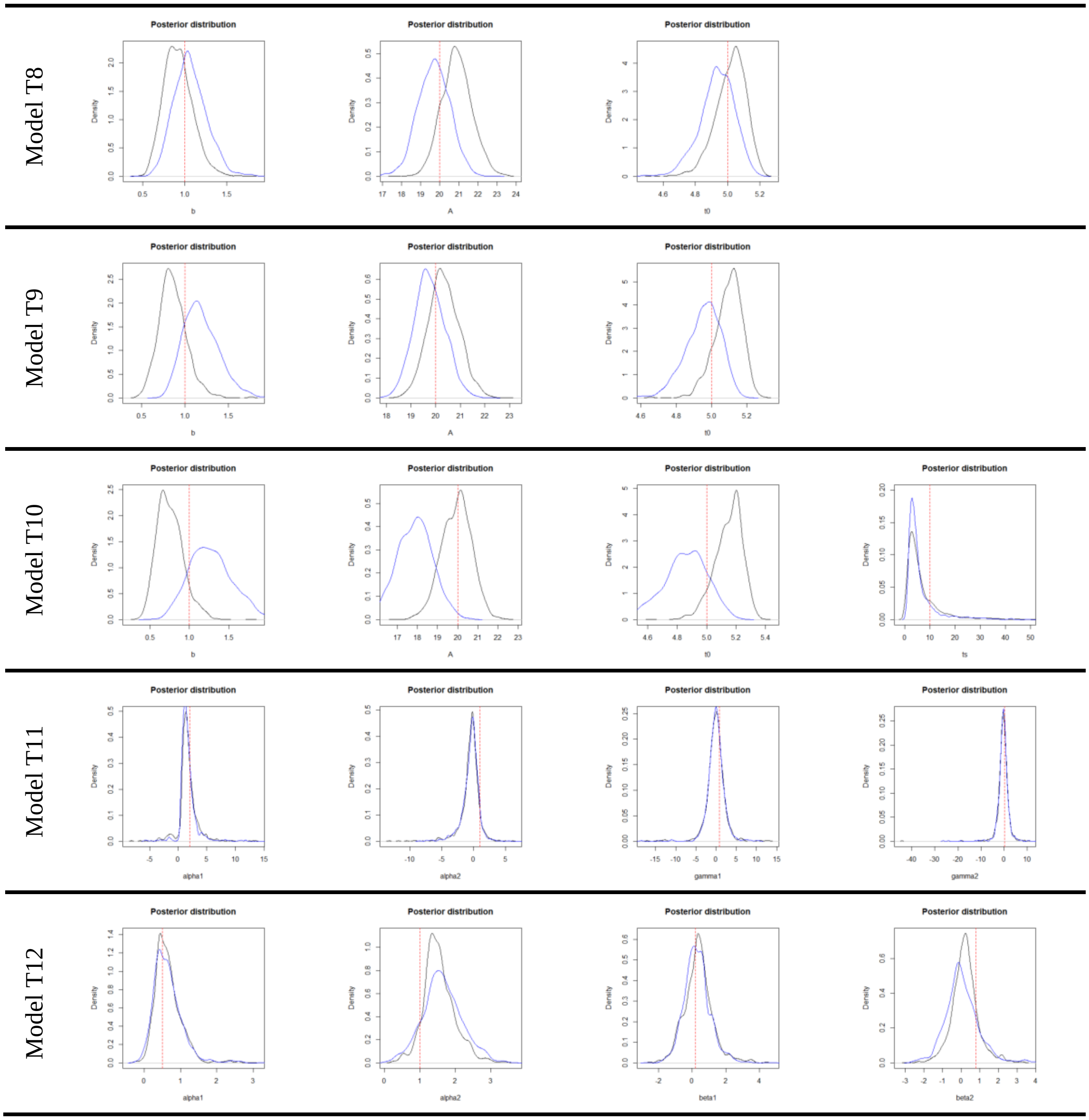


Figure 83. Model investigation - PMwG fit of revised structure two-stage models. Red line indicates generating values.

Table 39. Bayes Factors comparing CAG (best fit) model to other MLE fit models

Subject	CA	CG	IAB	IA	IB
1	125.9	2.2e+91	1.9e+50	6.1e+48	6.1e+48
2	10.0	5.2e+104	1.9e+42	5.9e+40	5.9e+40
3	26.8	1.5e+103	2.0e+43	3.1e+42	3.1e+42
4	4.3e+5	1.2e+51	2.6e+51	2.4e+50	6.2e+57
5	4.1e+7	1.1e+26	1.1e+60	1.3e+60	1.1e+74
6	3.0e+6	1.5e+37	6.7e+53	2.7e+52	1.4e+63
7	1.1e+6	9.0e+35	2.4e+60	1.3e+59	3.5e+68
8	5.8e+5	1.1e+25	2.0e+59	6.9e+59	1.3e+71
9	9.5e+5	6.9e+45	1.5e+65	1.6e+64	4.3e+67
10	1.3e+7	3.0e+33	9.6e+57	6.9e+56	2.6e+68
11	4.2e+7	5.6e+34	5.1e+67	1.6e+66	5.9e+73
12	0.6	1.0e+122	6.2e+38	2.0e+37	2.0e+37
13	9.2	1.3e+111	6.3e+44	2.0e+43	2.0e+43
14	1.0e+4	4.9e+72	1.7e+51	1.6e+51	2.0e+52
15	138.3	1.5e+92	1.4e+48	9.4e+46	9.4e+46
16	97.1	4.5e+90	1.8e+47	5.8e+45	5.8e+45
17	5.8	1.1e+108	3.3e+45	1.0e+44	1.0e+44
18	310.4	4.4e+88	4.2e+46	6.5e+46	6.5e+46
19	1.4e+5	1.6e+56	9.5e+59	1.3e+59	5.1e+61
20	4.7e+5	3.1e+49	6.1e+50	3.4e+49	4.9e+57

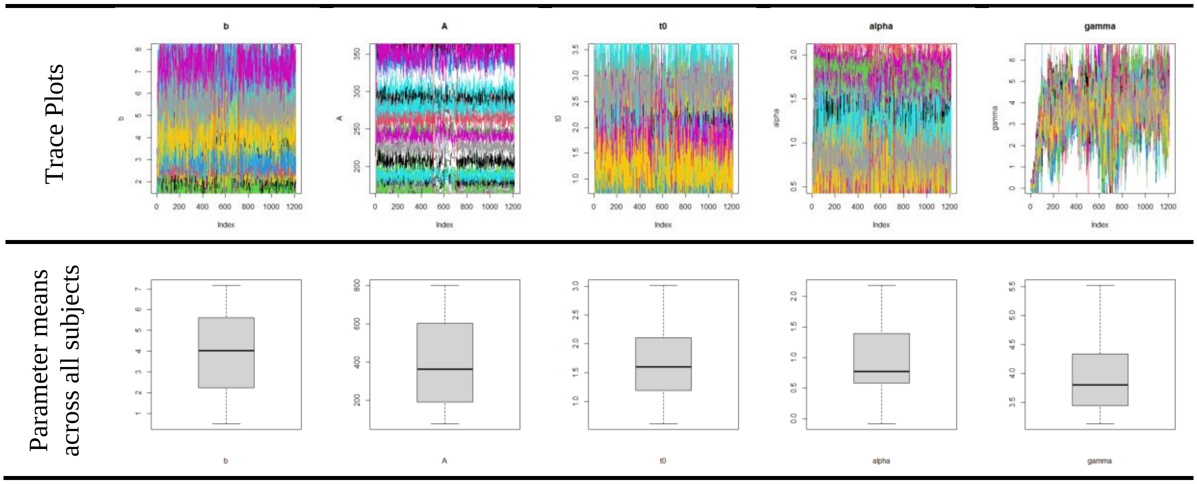


Figure 84. Simulated data parameter estimates – single-stage CPT model. Vertical red lines differentiate burn-in, adaptation, and sampling phases.

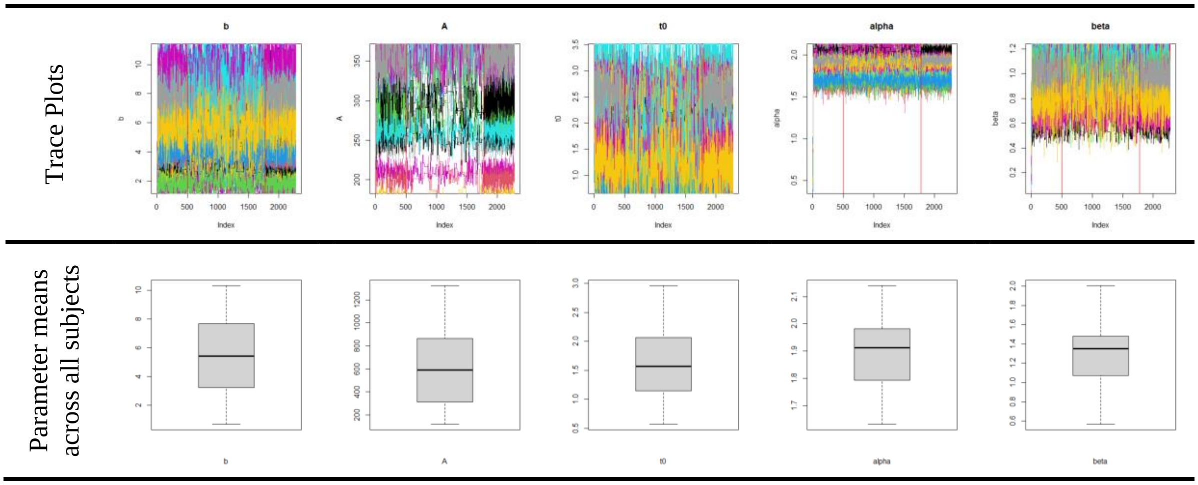


Figure 85. Simulated data parameter estimates – single-stage IF model. Vertical red lines differentiate burn-in, adaptation, and sampling phases.

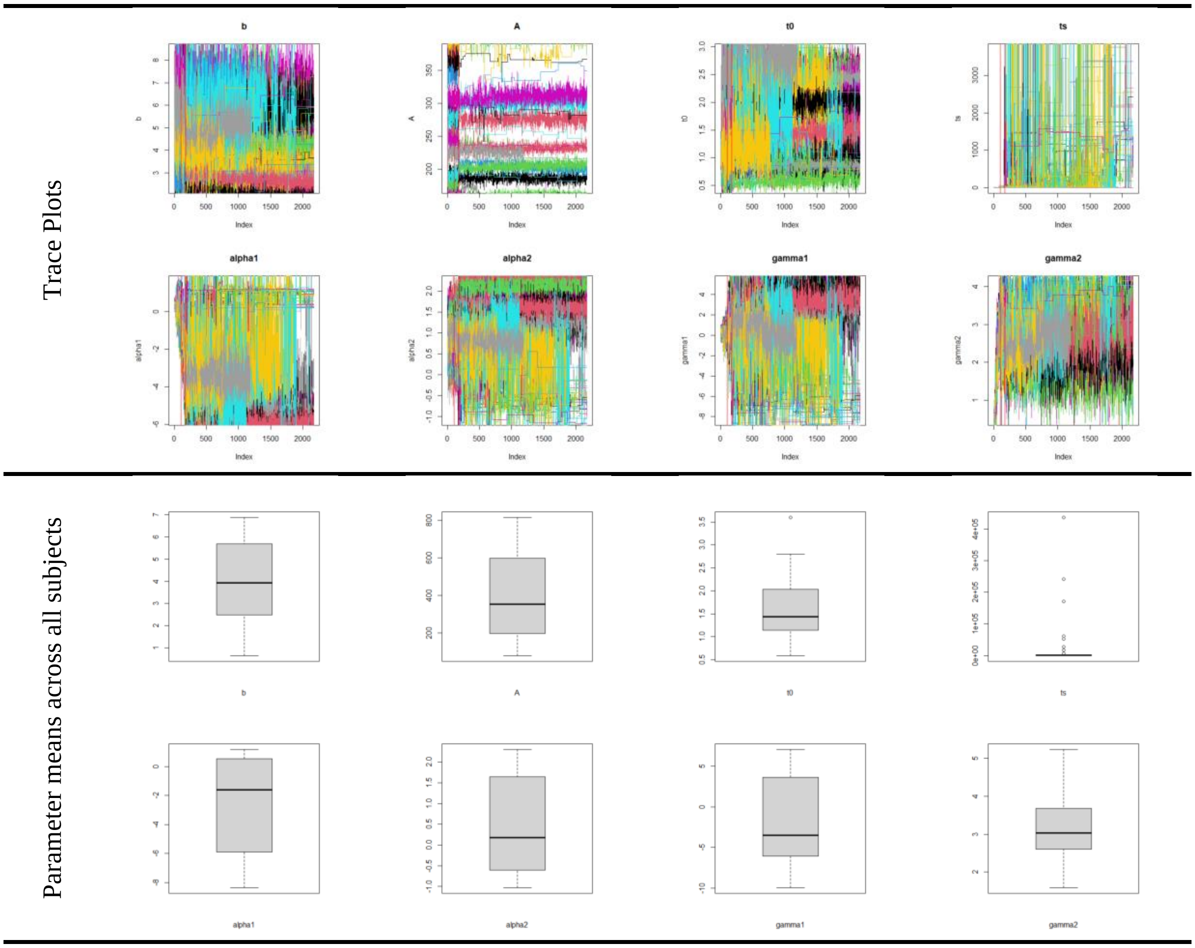


Figure 86. Simulated data parameter estimates – two-stage CPT model. Vertical red lines differentiate burn-in, adaptation, and sampling phases.

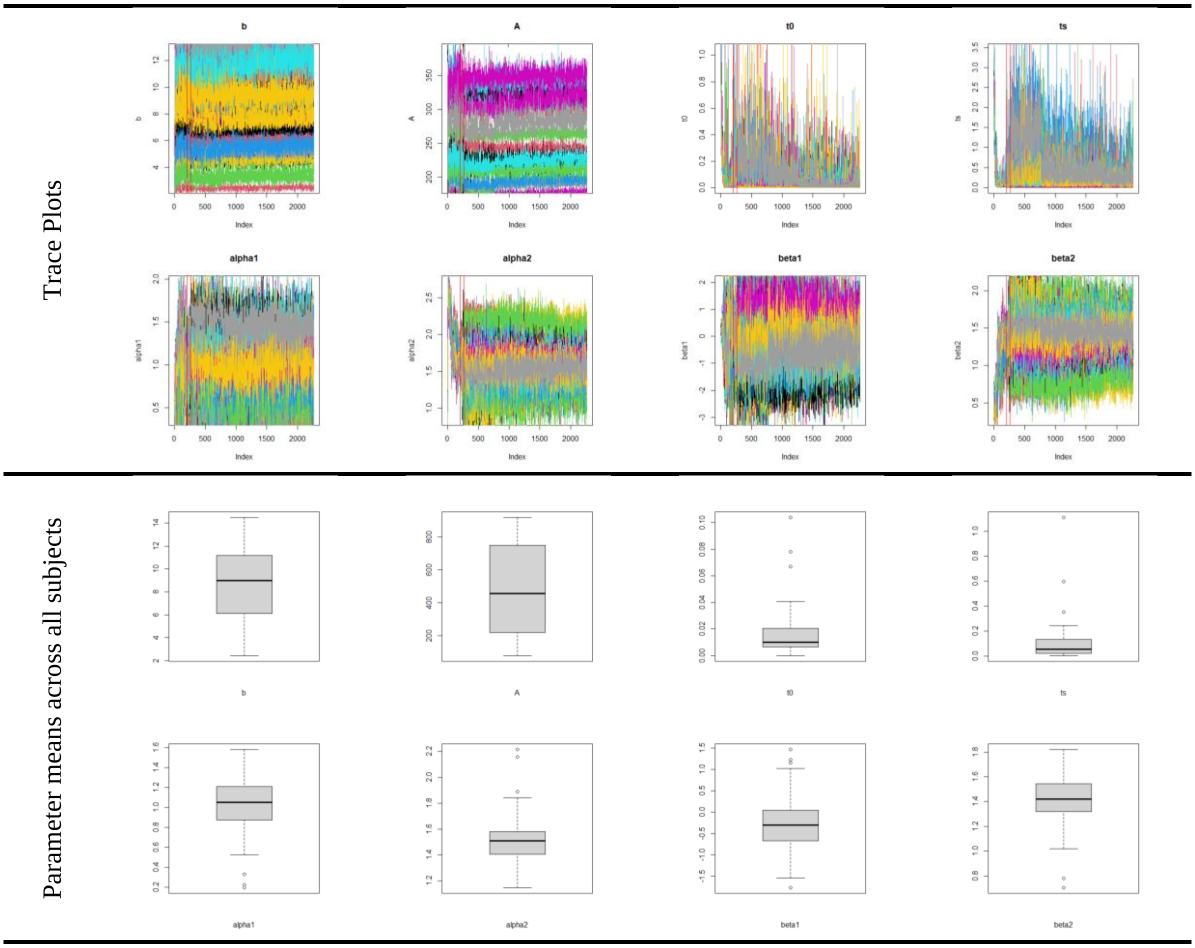


Figure 87. Simulated data parameter estimates – two-stage IF model. Vertical red lines differentiate burn-in, adaptation, and sampling phases.

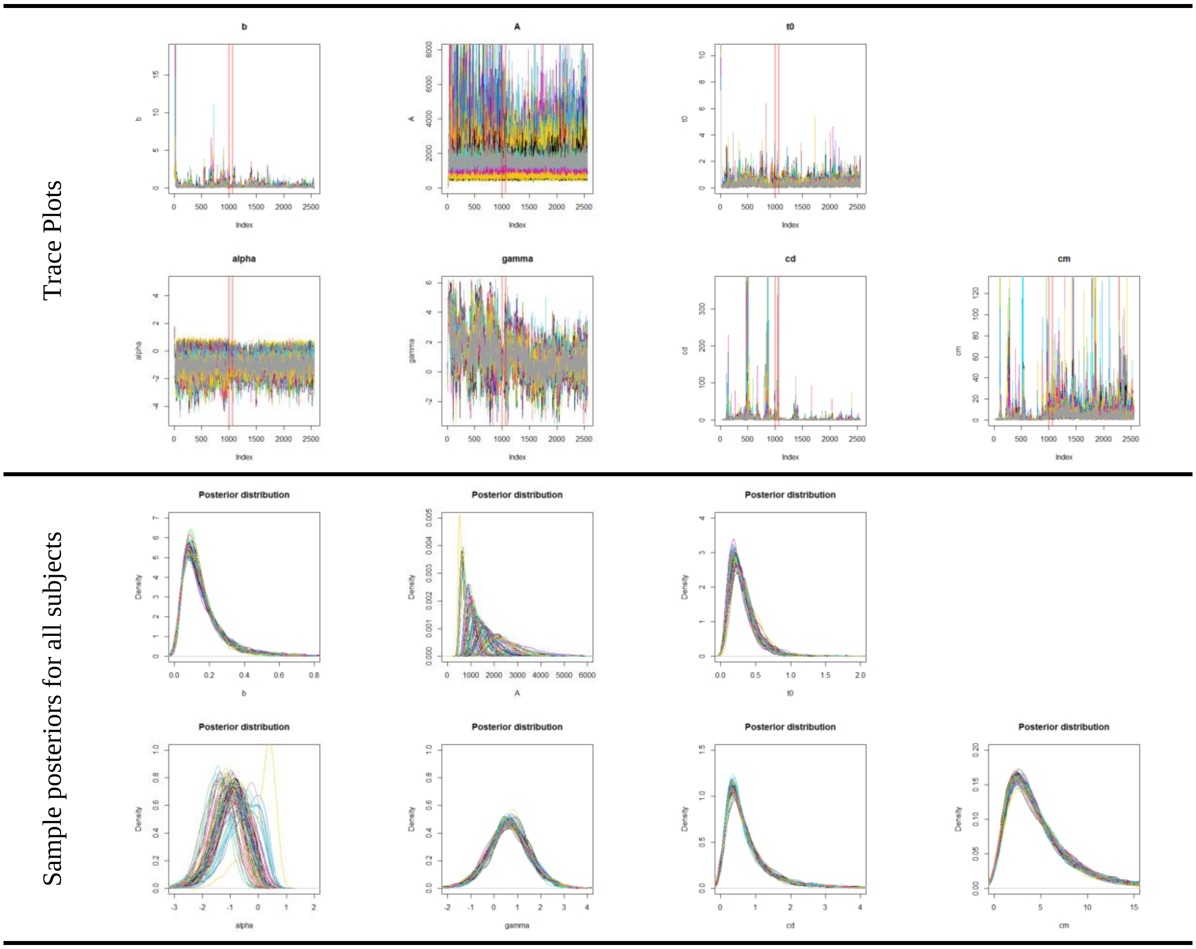


Figure 88. Campaign 3 data parameter estimates – single-stage 7-parameter CPT model. Vertical red lines differentiate burn-in, adaptation, and sampling phases.

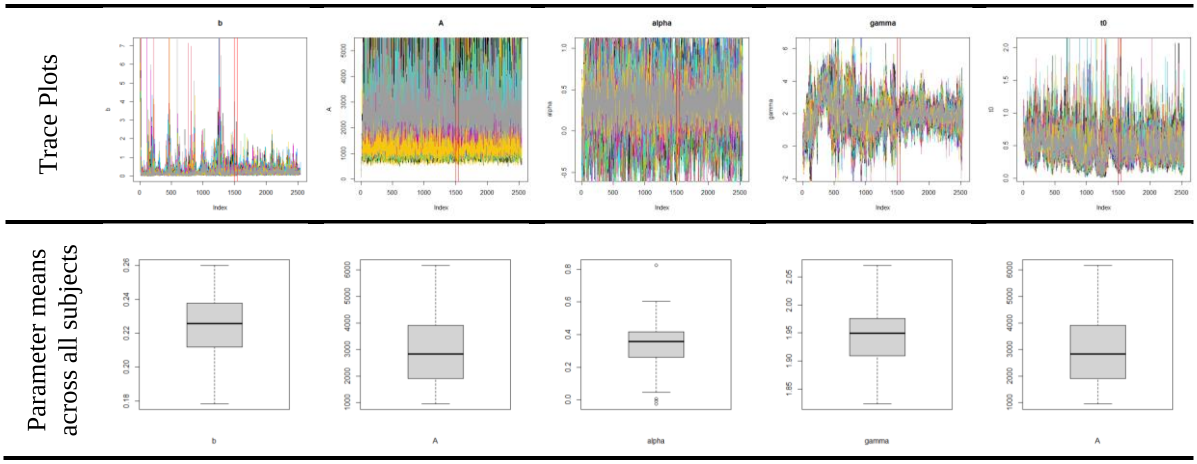


Figure 89. Campaign 3 data parameter estimates – single-stage 5-parameter CPT model. Vertical red lines differentiate burn-in, adaptation, and sampling phases.

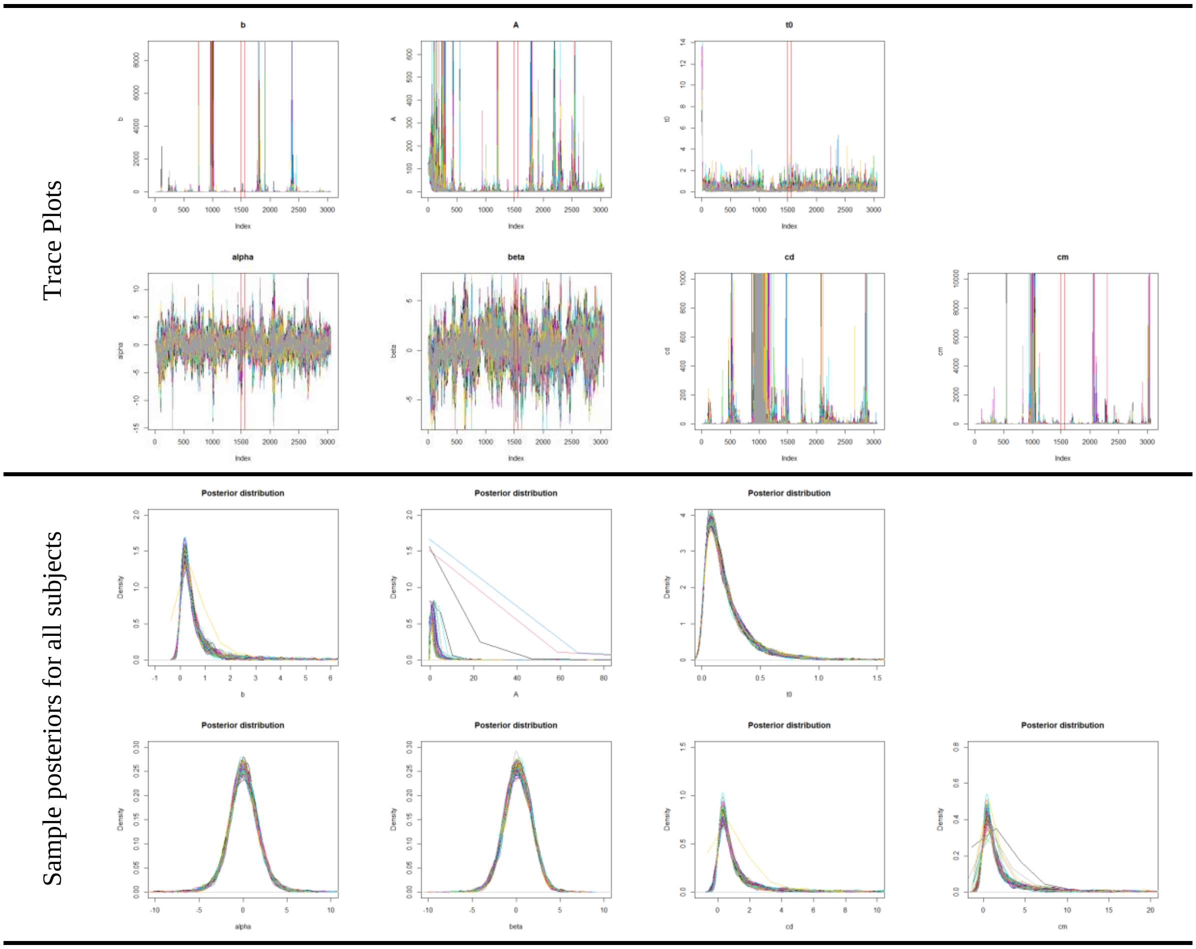


Figure 90. Campaign 3 data parameter estimates – single-stage 7-parameter IF model. Vertical red lines differentiate burn-in, adaptation, and sampling phases.

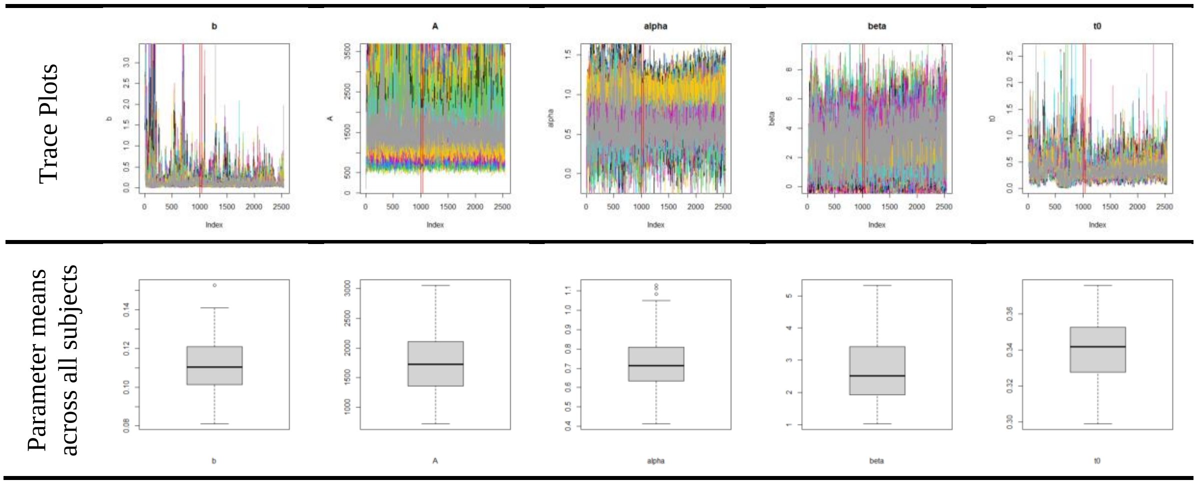


Figure 91. Campaign 3 data parameter estimates – single-stage 5-parameter IF model. Vertical red lines differentiate burn-in, adaptation, and sampling phases.

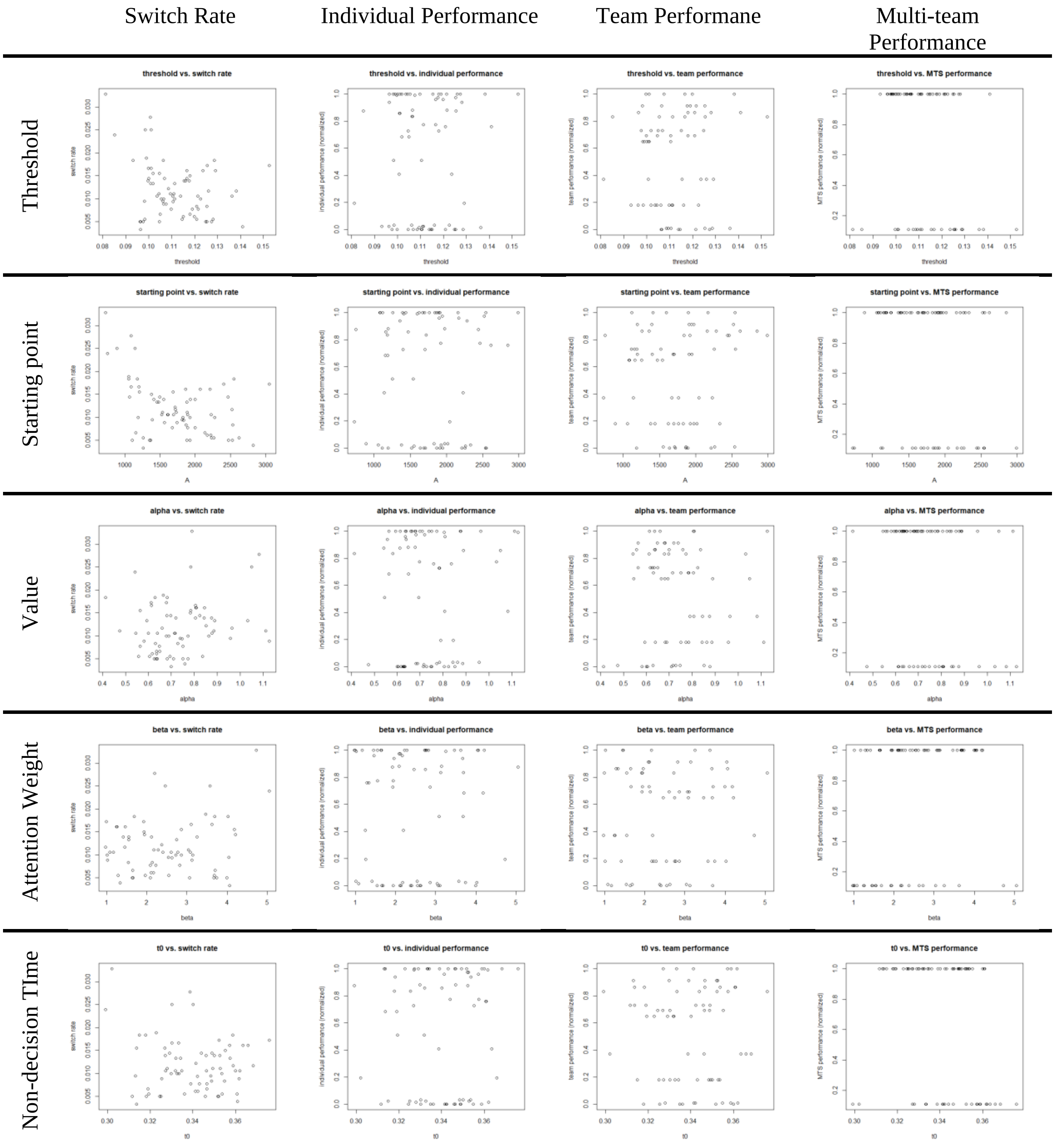


Figure 92. Parameter vs. outcome using single-stage IF model estimated parameters for campaign 3 data

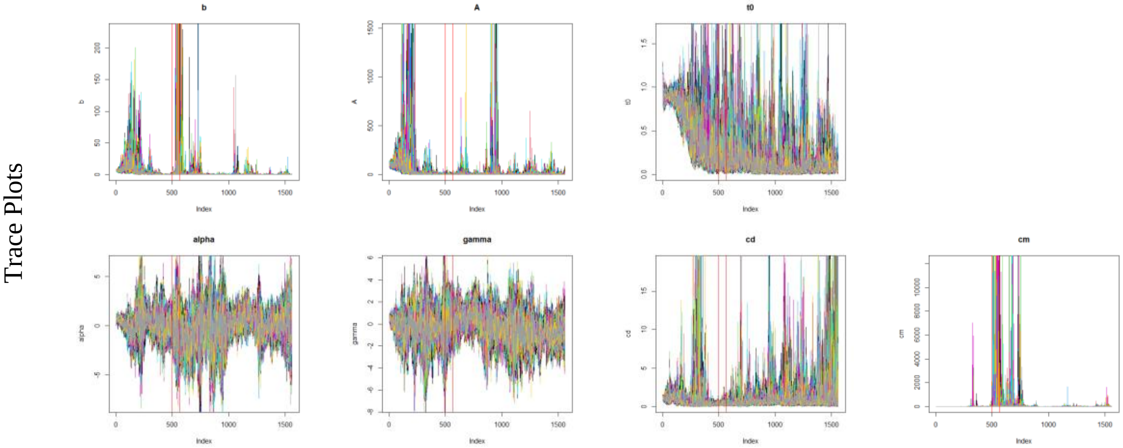


Figure 93. Campaign 4 data parameter estimates – single-stage 7-parameter CPT model. Vertical red lines differentiate burn-in, adaptation, and sampling phases.

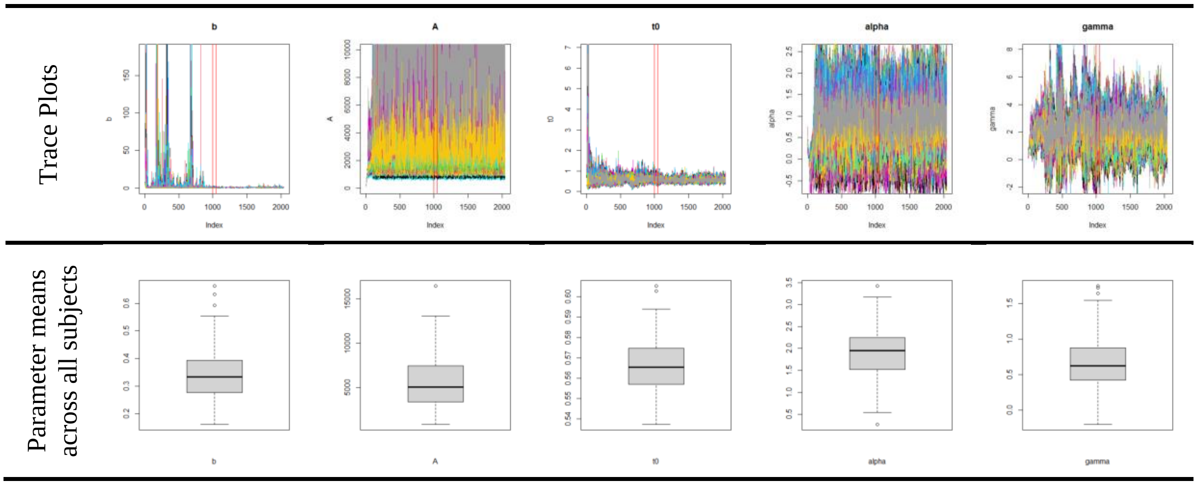


Figure 94. Campaign 4 data parameter estimates – single-stage 5-parameter CPT model. Vertical red lines differentiate burn-in, adaptation, and sampling phases.

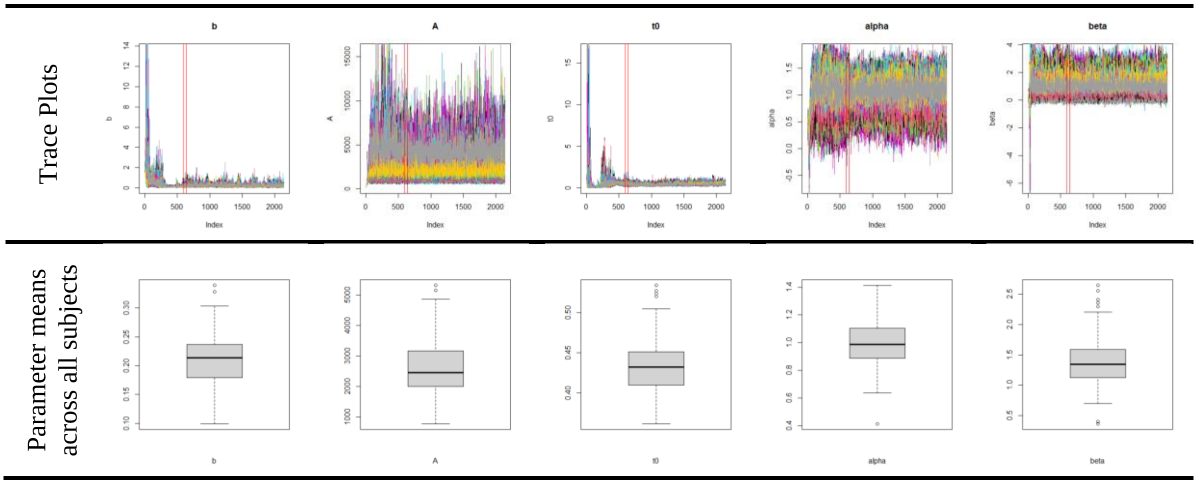


Figure 95. Campaign 4 data parameter estimates – single-stage 5-parameter IF model. Vertical red lines differentiate burn-in, adaptation, and sampling phases.

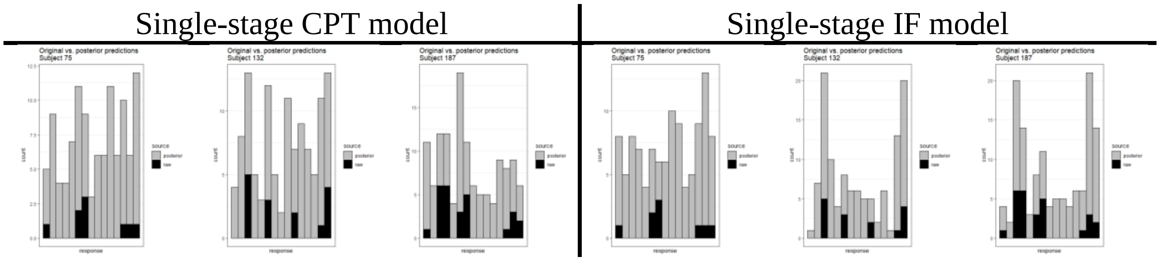


Figure 96. Choice distributions for subjects 75, 132, and 187 from campaign 4 showing the difference between the original (black) and posterior (grey) data for single-stage models.

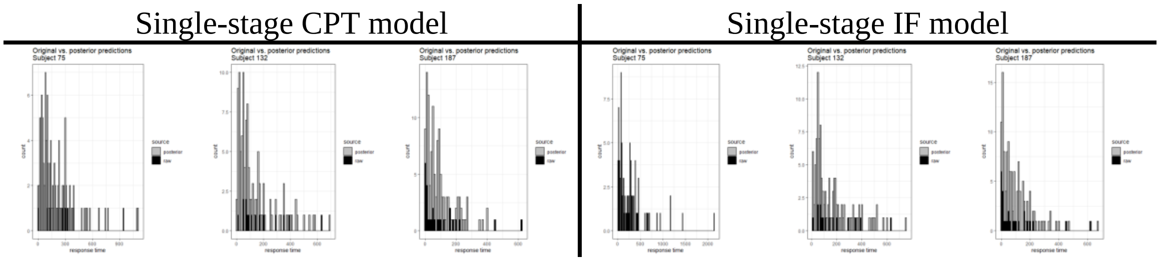


Figure 97. Response time distributions for subjects 75, 132, and 187 from campaign 4 showing the difference between the original (black) and posterior (grey) data for single-stage models.

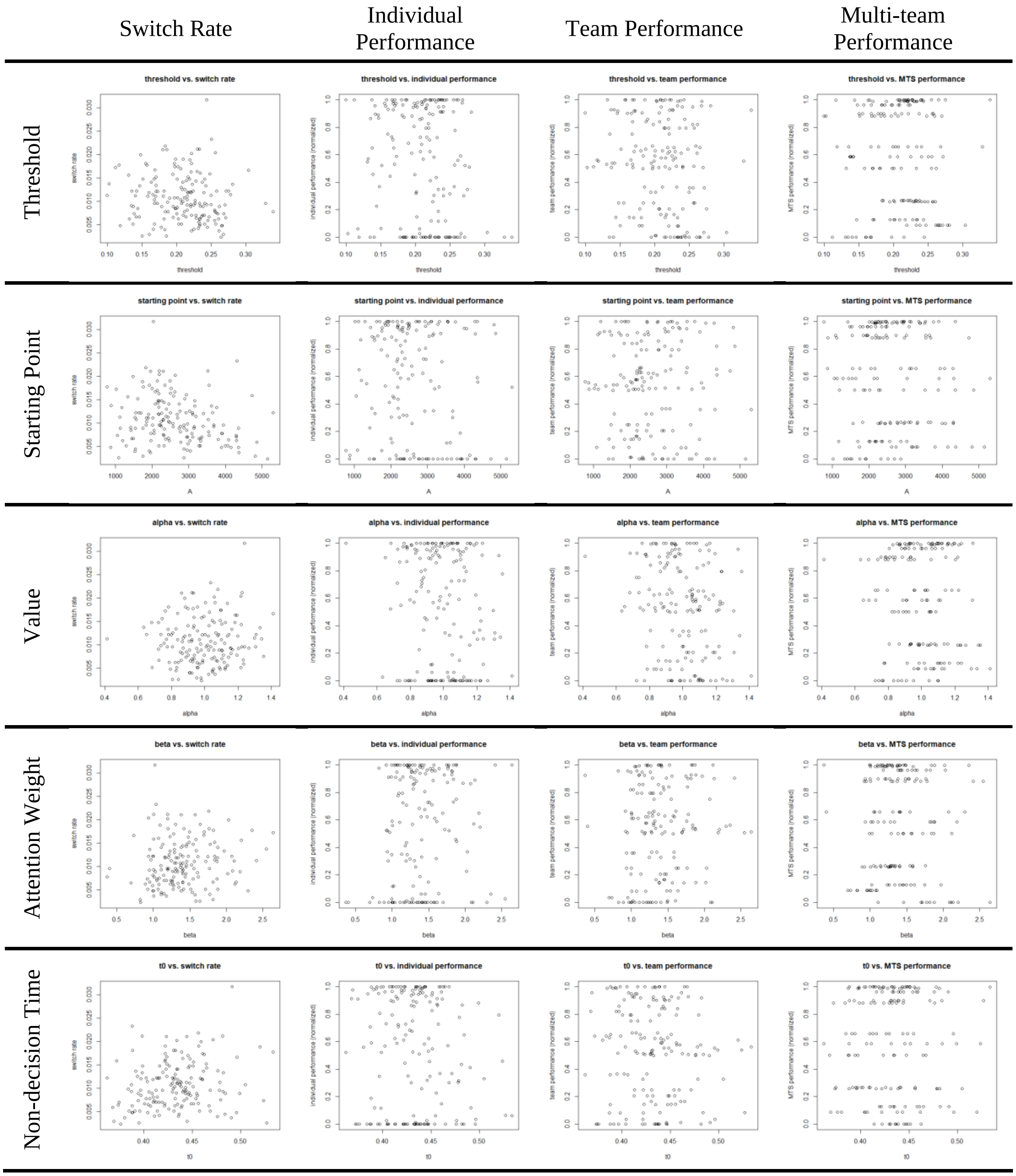


Figure 98. Parameter vs. outcome using single-stage IF model estimated parameters for campaign 4 data

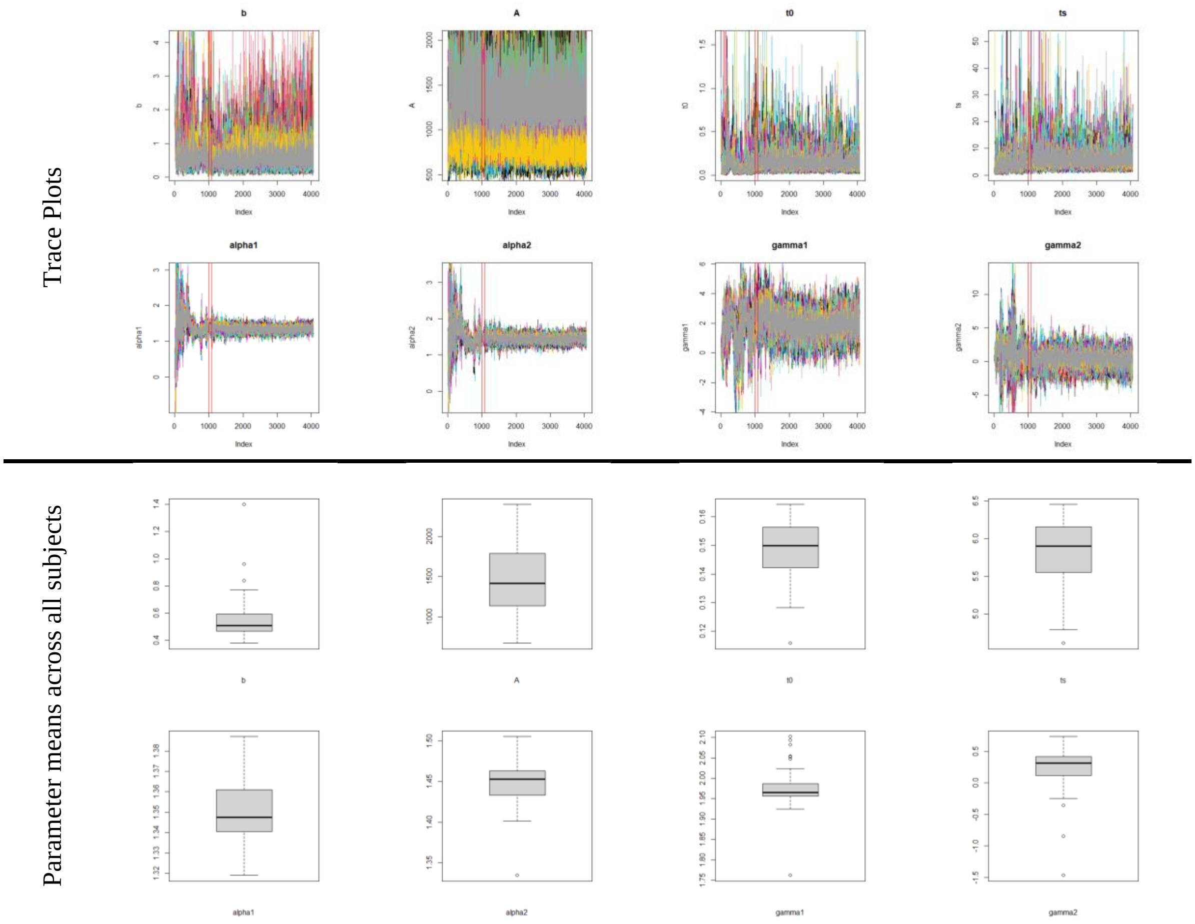


Figure 99. Campaign 3 data parameter estimates – two-stage CPT model. Vertical red lines differentiate burn-in, adaptation, and sampling phases.

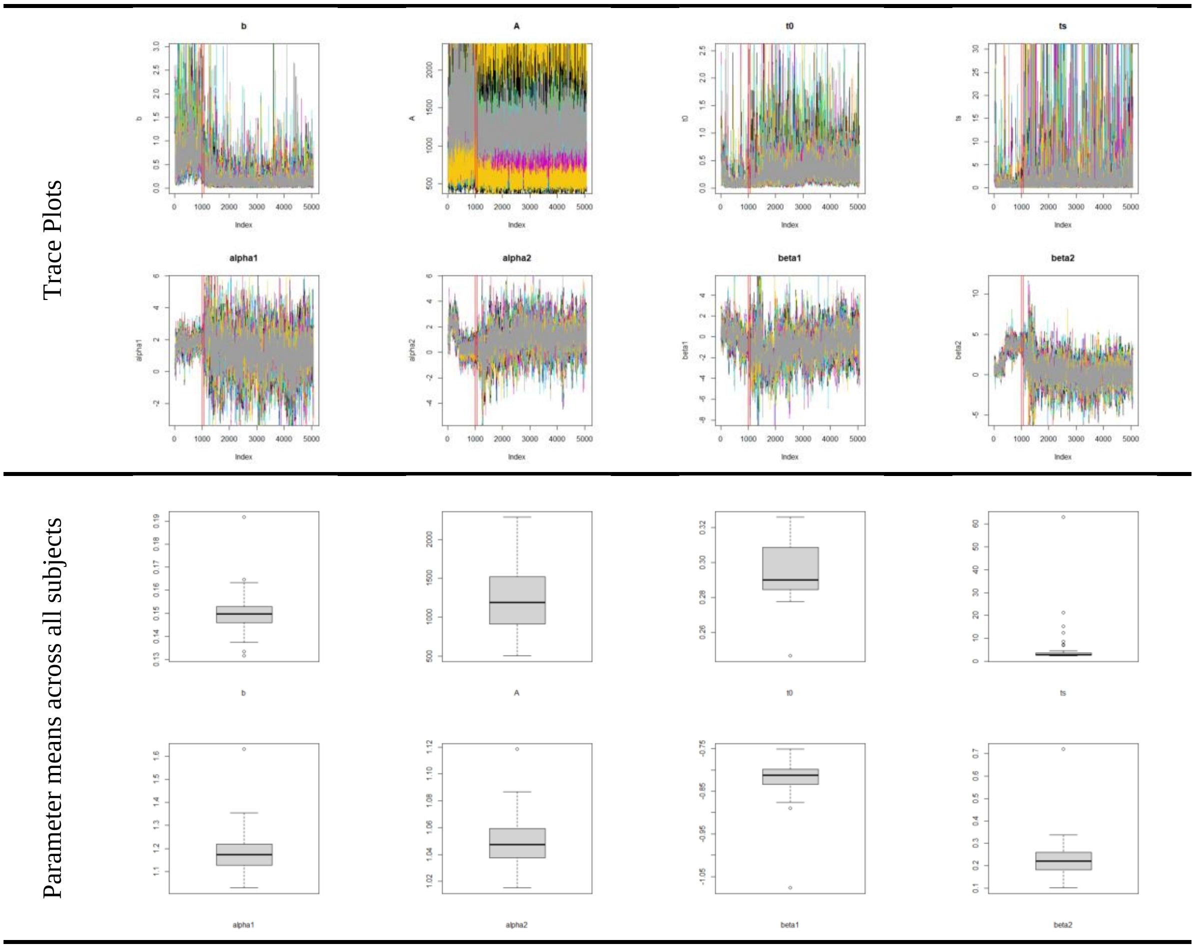


Figure 100. Campaign 3 data parameter estimates – two-stage IF model. Vertical red lines differentiate burn-in, adaptation, and sampling phases.

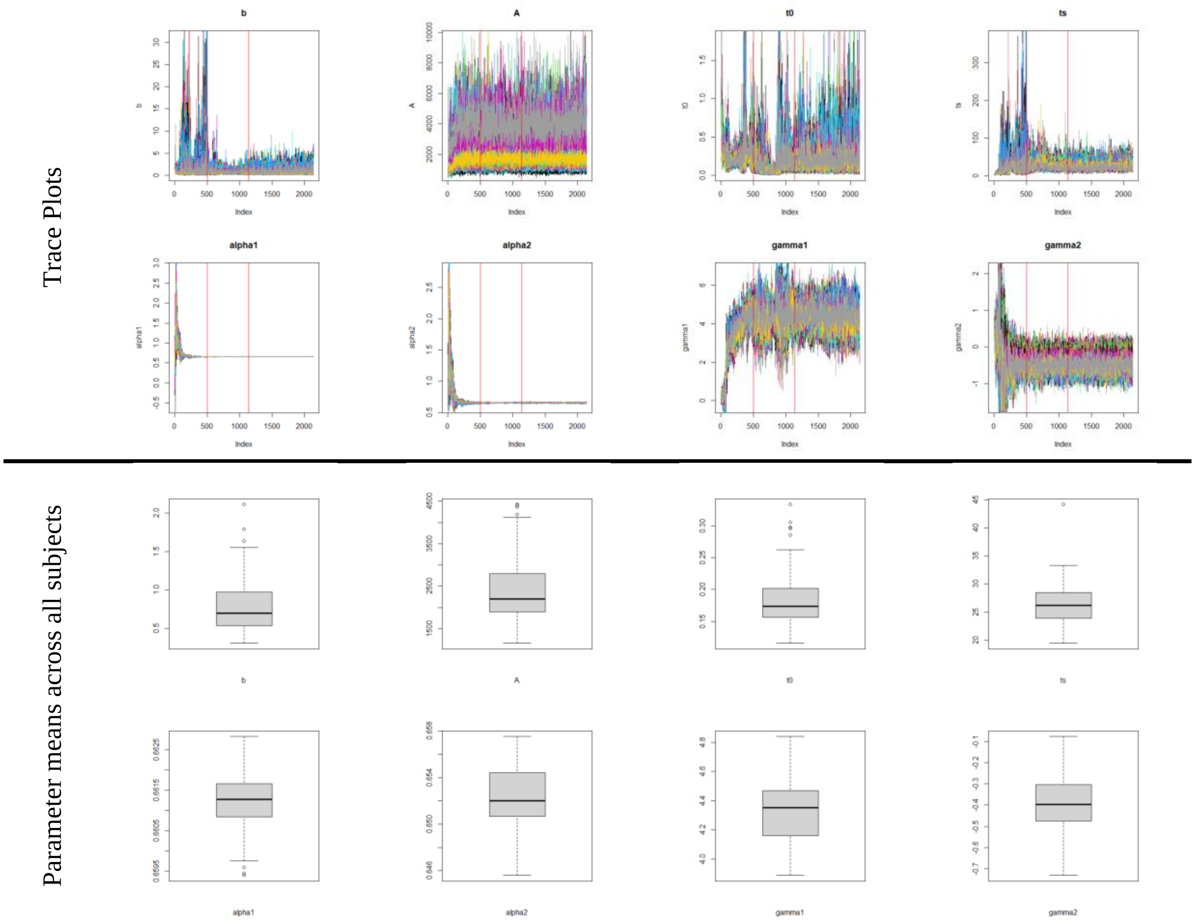


Figure 101. Campaign 4 data parameter estimates – two-stage CPT model. Vertical red lines differentiate burn-in, adaptation, and sampling phases.

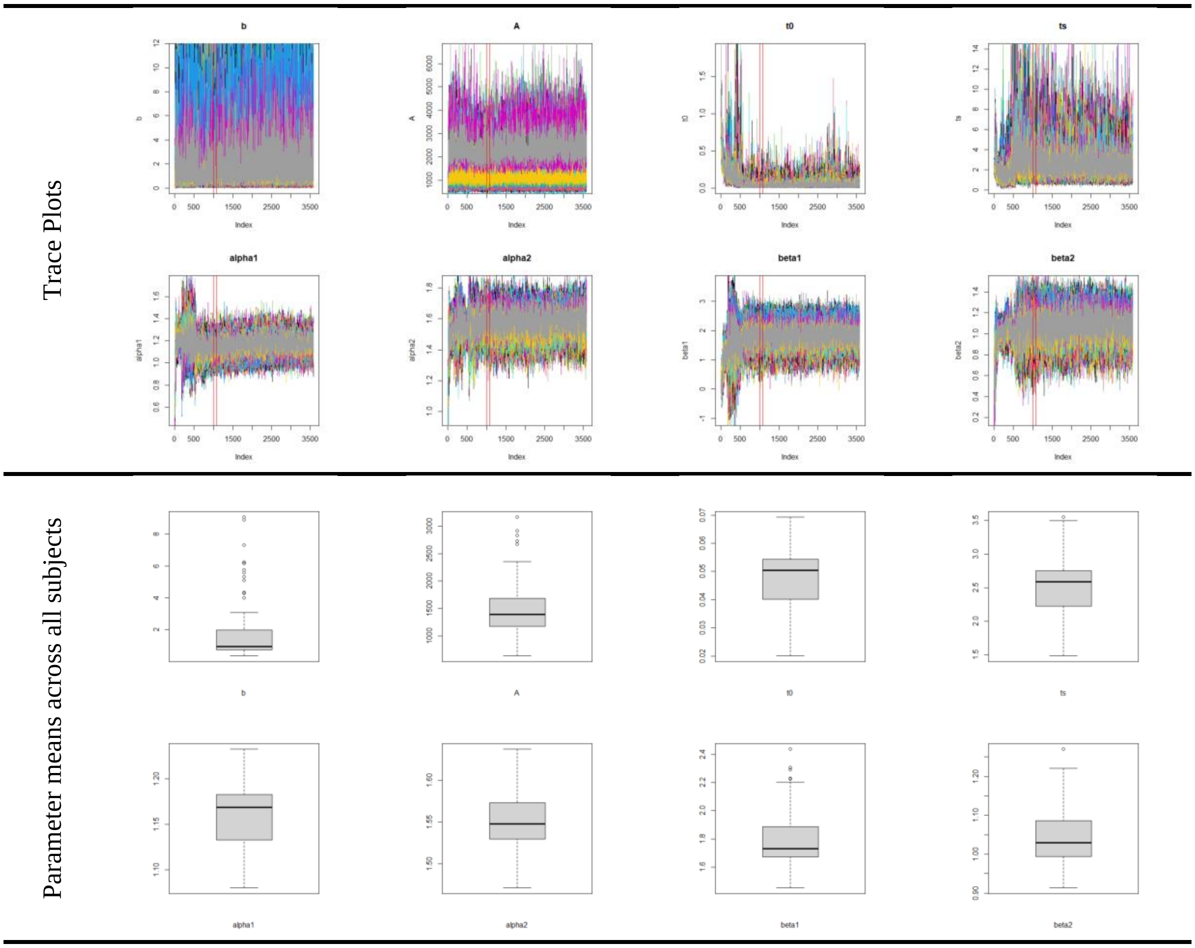


Figure 102. Campaign 4 data parameter estimates – two-stage IF model. Vertical red lines differentiate burn-in, adaptation, and sampling phases.

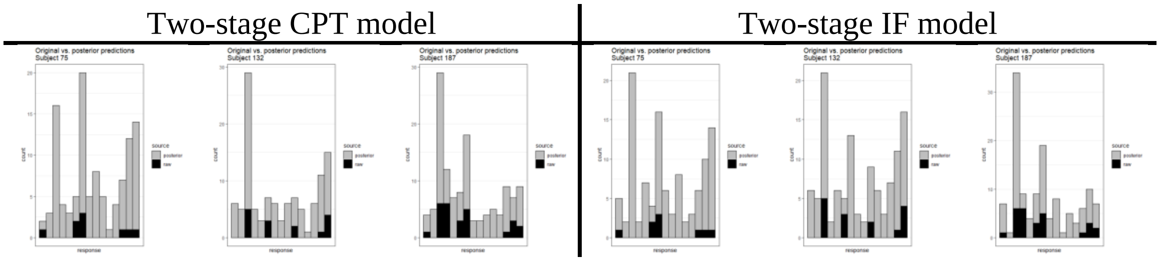


Figure 103. Choice distributions for subjects 75, 132, and 187 from campaign 4 showing the difference between the original (black) and posterior (grey) data for two-stage models.

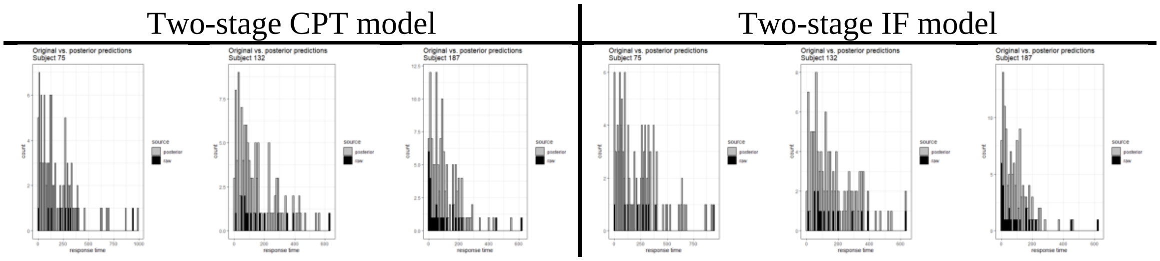


Figure 104. Response time distributions for subjects 75, 132, and 187 from campaign 4 showing the difference between the original (black) and posterior (grey) data for two-stage models.