2018

Bridging the Gap Between Micro- and Macrocognition: Testing the Multifunction Mental Model Hypothesis

Brittany Nelson
Michigan Technological University, bnelson1@mtu.edu

BRIDGING THE GAP BETWEEN MICRO- AND MACROCOGNITION: TESTING THE MULTIFUNCTION MENTAL MODEL HYPOTHESIS

By

Brittany L. Nelson

A THESIS

Submitted in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

In Applied Cognitive Science and Human Factors

MICHIGAN TECHNOLOGICAL UNIVERSITY

This thesis has been approved in partial fulfillment of the requirements for the Degree of MASTER OF SCIENCE in Applied Cognitive Science and Human Factors.

Department of Cognitive and Learning Sciences

Thesis Advisor:	*Dr. Shane T. Mueller*
Committee Member:	*Dr. Elizabeth Veinott*
Committee Member:	*Dr. Robert Hoffman*
Department Chair:	*Susan Amato-Henderson*

Table of Contents

Acknowledgements

I first and foremost need to express a deep thank you to my advisor Shane T. Mueller. Thank you for being an advisor and mentor of the highest quality. Thank you for challenging me and for always having your door open when I needed help. The genuine care and concern you have for your students and their growth does not go unnoticed. Thank you for providing an outstanding example I try to model myself after.

I also need to thank my committee members Elizabeth Veinott and Robert Hoffman. Thank you for your years of experience and knowledge which you have generously lent to this research. “You great benefactors, sprinkle our society with thankfulness. For your own gifts, make yourselves praised.” (William Shakespeare, a long time ago).

Thank you Steven Landry for helping me understand one of the keys to the universe, Rstudio. Thank you to Steven Landry, Jason Sterkenburg, Maryam Fakhr, and Yuanjing Sun for your support, wisdom, kindness, conversations, and late nights in the office. Thank you for preventing me from being the only one working in the office past 2:00AM. Thank you for making work fun.

Thank you to Jinan Allan, Amy Oja, Kari Oja, and Brandon Martinez for keeping fun, laughter, and sanity in my life. “The smallest act of kindness is worth more than the grandest intention.” (Oscar Wilde) An additional thank you to Brandon Martinez for helping ensure this document is free from fragmented sentences. And a special thank you to Jinan Allan for always challenging me and helping me grow both as a student and as a person.

Thank you to my parents and David Schreifels for their unwavering belief in me and support. A special thank you to David Schreifels for introducing me to the craziness that is the Michigan Technological University pep band. Thank you for showing me that it is possible to both have hobbies and be a student.

Thank you to Susan Amato-Henderson, the head of our department, for her leadership, kindness, and for generally being an amazing and cool person. One of my regrets is not taking more classes from you.

Finally, thank you to Edward T. Cokely, for showing me the wonders of cognitive science. Without you I would have never stumbled onto the path.

If I wrote an acknowledgement section to the length these people deserved the printer would need to be stocked with paper twice.

Abstract

Unique benefits can be gained by combining advantages of both micro- and macrocognitive methods that would otherwise be impossible to gather from either of these methods separately. The proposed research examines several cognitive functions within one systematic study that combines some empirical investigation with post-hoc qualitative assessment to gather knowledge of strategies and computations. Thereby, analyzing a larger cognitive system in a standardized way. By analyzing several cognitive functions the multifunction mental model hypothesis (MMM) is explored. This hypothesis states that performance of one sensemaking operation is predictive of performance of other related sensemaking operations. Three additional hypotheses were also explored. (2) Through brief instruction and feedback, mental models are developed that involve understanding the relational structure between inter-correlated and independent feature(s). (3) Understanding of the relational structure of the features can be used to make error correction decisions. (4) The strategies that utilize the inter-correlated nature of the features can be recognized and verbalized by users. Four Experiments used a multi-cue probabilistic weather forecasting task. Evidence from Experiments 1-4 supported the MMM hypothesis. Systematic variability in probability estimation by using differentially weighted features and inter-correlated features were related to evacuation decisions, error detection, and error correction. Results also supported hypotheses 2-4. The present research provides evidence which supports the integration of micro- and macrocognitive methods for a richer understanding of cognitive function in complex sociotechnical systems.

Chapter 1: Introduction

Research on complex cognitive function has involved a tension between understanding cognitive function in naturalistic contexts and using laboratory methods to replicate and isolate that cognitive function for careful study (Gozli, 2017; Kingstone, Smilek, & Eastwood, 2008; McDermott, 2011; Newell, 1973). Experimental psychologists often use laboratory methods to often deal with micro- aspects of cognition in attention, decisions, memory, problem-solving, prediction, and judgments (Cacciabue & Hollnagel, 1995; Ebbinghaus, 1913; Fitts, 1946; Shipley, 1961). Analyzing micro aspects of cognition is valuable and necessary, but not sufficient. Frequently, methods used in the microcognitive paradigm analyze one or two cognitive functions often in the form of simple linear causal chains (Klein & Hoffman, 2008). Using this reductionist approach is valuable for gaining large amounts of information about micro aspects of cognition (Klein et al., 2003). However, while the microcognitive paradigm provides valuable information about isolated aspects, it may come at the cost of discovering emergent processes and abilities when analyzed together in context (Gozli, 2017).

Today, this tension characterizes research on cognitive function. At one end, applied researchers (including clinical psychologists, education research, Industrial/Organizational research, and Human Factors, and related fields) often focus on macro-level processes that emerge from the combination of many low-level processes. In particular, cognitive systems engineering uses naturalistic studies to analyze how subcomponents fit together in context (Crandall, Klein, & Hoffman, 2006; Klein & Hoffman, 2008). However, some have argued that naturalistic approaches can lose the

assurance of well-designed highly controlled experiments; namely, the ability to draw causal inferences (Kingstone et al., 2008; McDermott, 2011). They are also less equipped to establish the same level of fidelity achieved in laboratory settings. In this thesis, I argue that unique benefits can be gained by combining parts of both methods, resulting in advantages that would otherwise be impossible to gather from implementing either of these methods independently. This approach examines several cognitive functions within one systematic study that combines some empirical investigation with post-hoc qualitative assessment to gather knowledge of strategies and computations. Thereby, analyzing a larger cognitive system in a standardized way.

One integrative perspective for understanding higher-level cognition in context has been called sensemaking. Sensemaking has been defined as *internal and external function performed for the purpose of forming a deeper understanding, so that one can act effectively.* The sensemaking process is accomplished, in part, through the supporting process that is 'mental models' (Kaste, 2012; Klein & Hoffman, 2008; Klein, Moon, & Hoffman, 2006b)[1]. Sensemaking is responsible for a number of operations such as how people comprehend, explain, make inferences, detect anomalies, diagnose errors, make predictions, and learn (Klein & Hoffman, 2008; Klein, Moon, & Hoffman, 2006a; Starbuck & Milliken, 1988; Weick, Sutcliffe, & Obstfeld, 2005). Evaluating the role

[1] The concept of the 'mental model' has been distinguished from similar concepts such as a 'frame' (Klein, Phillips, Rall, & Peluso, 2007). This will be discussed in further detail later on.

mental models play in these sensemaking operations is valuable for making inferences about how performance of these operations could be improved.

This thesis focuses on evaluating several of these valuable sensemaking operations. I will report on the results of a series of experiments in which participants interacted with, and made predictions about, a simulated weather forecasting system. I hypothesize that a variety of different functions, supported by the process of sensemaking, in this task will draw on common aspects of a mental model of the simulated weather forecasting system. In other words, I hypothesize people have different mental models that range in their quality of the intelligent tool that they represent, in this case a simulated weather forecasting system, and that different chunks of those mental models might be valuable for certain relatable tasks. Consequently, I predict that, to the extent there are systematic individual differences in performance on some components of the task, those who perform better will also perform better on other operations of sensemaking. I term this prediction as the *Multifunction Mental Model Hypothesis* (MMM).

This is not to say that mental models are not dynamic and cannot change, nor that a different mental model could not be chosen entirely. In fact evidence suggests mental models *are* elaborated and refined (Johnson-Laird, 2005; Vosniadou & Brewer, 1992). However, simply because a mental model of an intelligent tool changes does not mean that functions used to operate the intelligent tool are not supported by the same parts or aspects of the mental model. It only implies that a new or refined mental model replaces the previous mental model.

To explore and test this hypothesis, I will present four experiments which were designed to evaluate several sensemaking operations, including: learning relations between variables (i.e. function learning), decision making, forecasting, system error detection, and system error correction. In contrast to traditional experimental psychology experiments that have studied these operations in isolation, I will examine how these operations are supported by a common knowledge base and how they are related. Analyzing sensemaking operations in weather forecasting is an ideal space for combining methods of both the micro- and macrocognitive paradigms. A number of micro- and macrocognitive studies have been conducted in weather forecasting (Gluck & Bower, 1988; Gluck, Shohamy, & Myers, 2002; Hoffman, LaDue, Trafton, Mogil, & Roebber, 2017). However, consistent with the microcognitive paradigm these microcognitive studies have not evaluated a larger cognitive system, merely micro aspects of that system such as category learning (Gluck et al., 2002). A number of macrocognitive studies have also been conducted in more naturalistic contexts (Hoffman et al., 2017). However, these studies are not able to achieve the same level of fidelity as can be achieved within the lab. Therefore, there is an opportunity to expand upon previous work.

In addition to the MMM hypothesis, three other hypotheses are also proposed. The second hypothesis is that through brief instruction and feedback, mental models are developed that involve understanding the relational structure between inter-correlated and independent feature(s). The third hypothesis is that understanding of the relational structure of the features can be used to make error correction decisions. The last

hypothesis is that the strategies that utilize the inter-correlated nature of the features can be recognized and verbalized by users.

Exploring error detection, diagnosis, and correction by using inter-correlated features is valuable. Across many naturalistic contexts—including weather forecasting—the features used for error detection, diagnosis, and correction are inter-correlated. In order to help explain error detection and diagnosis by using inter-correlated features I will provide an illustrative example. An easy way to do that is to demonstrate inter-correlation as a result of location. Imagine three weather sensors located at Northern Michigan University (NMU). Now imagine three weather sensors located 100 miles away at Michigan Technological University (MTU). Whether these sensors were reporting information on temperature, rain fall, or cloud coverage it is highly likely that the sensors located at the same university are reporting the same information. For example, all three sensors will likely all be reporting snowfall or all three sensors will be reporting rain. If one of the sensors reported sunshine and the other two sensors in the same location reported rainfall one might think there is an error in the sensor reporting the inconsistent information. However, if the three sensors at NMU reported sunshine and the three sensors at MTU reported snowfall you wouldn't necessarily think that there was an error with the sensors. This is because the information reported from the sensors in the same location are inter-correlated while the information reported from the sensors located at NMU versus the information being reported from MTU are relatively independent. In order to accurately detect, diagnose, and correct an error by using inter-correlated features it requires a different kind of strategy and mental computation compared to

diagnosing errors by using independent features alone. Therefore, it is valuable to determine (1) if operates can learn the relational structure of the inter-correlated features (2) if that understanding can be utilized for detecting, diagnosing, and correcting errors and (3) if this strategy can be recognized and verbalized.

In this thesis I argue that studying mental models in the lab utilizes advantages of both micro- and macrocognitive paradigms. Studying mental models by combining these methods provides an opportunity to analyze how different sub components of cognition fit together in a larger system that is sensemaking in a systematic way. The MMM hypothesis is also explored; those who perform better in one sensemaking operation will also perform better on other operations of sensemaking. Implications of this hypothesis, such as training, is also explored.

Chapter 2: Literature Review

The outline of this literature review is as follows. First, I briefly review the definitions of micro- and macrocognition and their distinguishing features. Reviewing the micro- and macrocognitive paradigms is valuable for demonstrating their strengths and weaknesses and for arguing why combining methods from each paradigm creates unique advantages. I then provide a review of the definition of the integrative process under investigation—sensemaking. This review of what sensemaking is and how the sensemaking process operates lays the foundation to understand the role mental models play during the sensemaking process. Finally, the definition of mental model and its role within error management is discussed. This review is valuable for providing support for the MMM hypothesis and demonstrating some of its potential implications in an applied context.

Distinguishing Microcognition from Macrocognition

Microcognition and macrocognition are complementary paradigms of research (Klein et al., 2003). However, to better understand this it is helpful to examine and define each perspective more clearly. Microcognition is the study of invariant processes often in the form of binary oppositions such as: massed vs. distributed practice, serial vs. parallel processing, exhaustive vs. self-terminating search, single code vs. multiple code, and so on (Cacciabue & Hollnagel, 1995; Klein et al., 2003; Newell, 1973). The study of microcognition often utilizes college students in controlled artificial laboratory settings (Smieszek & Rußwinkel, 2013). One of the advantages of microcognitive study is

internal validity, or the ability to draw causal inference (McDermott, 2011). Convenient and large samples are useful when analyzing the effects of several independent variables on a dependent variable, which requires much larger sample sizes in order to conduct more complex statistical analyses.

In comparison, macrocognition is the framework for describing cognition as it naturally occurs (Klein et al., 2003; Schraagen, Klein, & Hoffman, 2008). The study of macrocognition focuses on the performance of complex human-machine systems as a whole (Smieszek & Rußwinkel, 2013). To accomplish this goal, researchers often analyze subject matter experts within naturalistic contexts using cognitive task analysis methods (Crandall et al., 2006; Klein & Hoffman, 2008; Klein et al., 2003). Macrocognitive research includes topics such as naturalistic decision making, planning, problem detection, coordination, adaptation, and sensemaking (Klein, Pliske, Crandall, & Woods, 2005; Klein et al., 2003). Although many of these topics are also studied from a microcognitive perspective there are two typical differences. (1) Reliance on studying the functions in a natural context, and (2) examining how multiple microcognitive functions interact to produce emergent complex behavior.

Micro- and macrocognition are not antagonist paradigms of research (Smieszek & Rußwinkel, 2013). Rather, each can be used to inform and inspire the other (Klein et al., 2003). Some have suggested a bottom up approach; start with microcognitive phenomena to inspire research in macrocognitive function (Klein et al., 2003). While others have suggested that by first observing phenomena as it naturally occurs we are more likely to create universally valid theories (Kingstone et al., 2008). Theories derived from

phenomena observed in naturalistic contexts will likely be more robust compared to phenomena only analyzed within the lab. Effects discovered within the lab may be so sensitive to other variables within more naturalistic contexts that the same effects may never be observed in those more naturalistic contexts (Kingstone et al., 2008).

As mentioned above, there are two main distinctions between micro- and macrocognition: analyzing the system as a whole and analyzing the phenomena in context. The present research focuses on the integrated cognitive system, but does not focus on cognition in context. Creating naturalistic conditions in the lab is challenging (Schraagen et al., 2008). However, consistent with the recommendations provided by Kingstone et al., (2008) the emergent process under investigation, sensemaking, is based on the expansive research conducted in naturalistic settings (Hoffman et al., 2017; Kaste, 2012). Mental models used by experts in weather forecasting have been observed in complex naturalistic environments (Hoffman et al., 2017). These mental models are used for a number of sensemaking operations. However, testing the interactions and relations of several related sensemaking operations has yet to be explored in the lab. I'm attempting to analyze a larger cognitive system within the lab. This fills a valuable gap because the nature of the microcognitive paradigm is fundamentally reductionist. This reductionist approach misses the opportunity to analyze the fluctuations and interactions between the primary functions/behavior and their supporting functions (Klein & Hoffman, 2008). Such as the use of mental models for various related sensemaking operations.

What is sensemaking?

From military operations, to leadership, to weather forecasting, researchers are studying the role of sensemaking on vital operations (Alberts & Garstka, 2004; Ancona, 2012; Hoffman et al., 2017). Sensemaking has been studied from diverse disciplinary backgrounds. There is not one unified definition of sensemaking accepted across disciplines (Weick, 1995). It is beyond the scope of the present paper to review all definitions of sensemaking[2]. However, since the primary focus of the present paper is on the role sensemaking plays in complex human-machine systems, it is more valuable to review what some notable systems engineers' perspective is on sensemaking.

In their seminal paper, Klein, Moon, and Hoffman (2006a) provide a thorough investigation as to what is meant when researchers say 'sensemaking.' The authors distinguish their definition from previous definitions such as "how people make sense out of their experience in the world," indicating that this type of definition is too broad and could encompass years of previous research in concepts such as creativity, curiosity, comprehension, and situation awareness. Rather, the authors define sensemaking as "a motivated, continuous effort to understand connections (which can be among people, places, and events) in order to anticipate their trajectories and act effectively" (Klein et al., 2006a, p. 71).

---

[2] See (Dervin & Naumer, 2009) for review on approaches of sensemaking.

Based on previous experience, when an operator approaches a system they have a frame (which is related to yet distinct from the concept of mental model). Sensemaking is required when there is some kind of ambiguity, complexity, or anomaly (without which there would be nothing to make sense of). Within a socio-technical system this occurs when feedback from an intelligent tool is inconsistent with the operators' current frame. An operator could choose to ignore the anomaly. However, if the operator does not ignore the anomaly then both mental mechanisms and external behavior may be required to increase their understanding of the anomaly which can elaborate an existing frame or choose a new frame entirely. Across the literature of sensemaking, many support the notion that sensemaking is not limited to only internal mechanisms but rather sensemaking also consists of behaviors (Dervin, 1983; Pirolli & Russell, 2011; Weick et al., 2005). This could be in the form of communication. Some have gone as far as to say sensemaking involves any activity performed for the purpose of "collecting and organizing information for deeper understanding" (Pirolli & Russell, 2011, p. 1). Somehow an operator needs to go through a process of reconciling what he/she already knows about the system (which is in the form of a frame) with the new information that does not currently fit within their existing understanding (or frame). This is the process of sensemaking. This processes is depicted in the data/frame model (Klein et al., 2006b).

Mental Models

Without mental models many of the functions of sensemaking would not be possible (Fallon, Murphy, Zimmerman, & Mueller, 2010; Klein et al., 2006b)[3]. Mental models support vital functions such as reasoning, explaining, and predicting (Johnson-Laird, 2001, 2006; Jones, Ross, Lynam, Perez, & Leitch, 2011; Rouse & Morris, 1986). Some dispute the existence of mental models (Johnson-Laird, 1983), however, the continually growing empirical and theoretical evidence provides strong support for their existence (Gentner & Stevens, 2014; Johnson-Laird, 1983, 2005; Klein & Hoffman, 2008). Similarly to sensemaking, the definition of 'mental model' is controversial (Moray, 1999; Revell & Stanton, 2012; Richardson & Ball, 2009; Rouse & Morris, 1986).

Before reviewing what mental models are, it is valuable to review what they are not. Some scholars view mental models as only a store of knowledge. Defining mental models as collections of knowledge has been considered as whole a class of definitions for mental models (Schumacher & Czerwinski, 1992). However, I argue that mental models are not mere collections of knowledge. Specifically, defining mental models in this way neglects the relational structure of mental models (Craik, 1943). Defining mental models as *only* knowledge stores results in a loss of utility by neglecting many of their functions such as problem solving and prediction (Rouse & Morris, 1986). Both

[3] See Rouse and Morris (1986) for a review of the diverse definitions of mental models.

empirical and theoretical research has suggested mental models are used for these and other vital functions (Gentner & Stevens, 2014; Johnson-Laird, 1983, 2005; Rouse & Morris, 1986). Mental models are also distinct from a 'frame' (Klein et al., 2006a). A frame has been defined as "a structure for accounting for the data and guiding the search for more data." (Klein et al., 2007, p. 118). In other words, a mental model is similar to the concept of a frame but a frame has some distinct aspects to it. Such as, taking the form of a narrative or story (Klein et al., 2007).

The origins of the theory of mental models could lead back all the way to the work of Thomas Aquinas *Summa Theologica* (1267) (Klein & Hoffman, 2008). However, according to Johnson-Laird (1983), the history of the theory of mental models really begins with Kenneth Craik. According to Craik all thinking is a manipulation of internal representations of the external world (Craik, 1943). Craik laid the groundwork for future theoretical and empirical research on mental models. Many of the first principles of mental models identified by Johnson-Laid are attributed to the work of Craik, including: the principle of iconicity, the principle of possibilities, and the principle of truth. Some have suggested that these and additional principles are what distinguish mental models from other types of mental representations such as linguistic structures and semantic networks (Johnson-Laird, 2005). In order to unpack what mental models are and how they operate a couple of the Johnson-Lairds' proposed mental model principles will be briefly reviewed.

The first principle is the principle of iconicity. This principle simply states that a "mental model has a structure that corresponds to the known structure of what it

represents" (Johnson-Laird, 2005, p. 187). The iconic nature of mental models can include mental imagery in combination with organized knowledge of concepts and relationships (Forrester, 1971; Klein & Hoffman, 2008). The imagistic nature of mental models can be used to help explain how a dynamic system operates (Klein & Hoffman, 2008). If an operator's mental model did not correspond to the dynamic system, the operator would likely not be able to infer the cause and effect relations between the different elements of that system.

The second principle is the principle of strategic variation. This principle simply states that "given a class of problems, reasoners develop a variety of strategies from exploring manipulations of models" (Johnson-Laird, 2005, p. 191). If we define error diagnosis in terms of problem solving, then exploring manipulations to mental models is key to an accurate diagnosis. Variation is valuable for providing insight, learning, and creativity (Johnson-Laird, 2004). Strategic variation is similar to the elaboration cycle within the data/frame model (Kaste, 2012; Klein et al., 2006b; Klein et al., 2007). This principle implies that mental models can be dynamic. There is an interaction between the nature of the mental model and the task the mental model is supporting.

Mental models are challenging to define (Klein & Hoffman, 2008). Indeed, it is unclear what the *correct* definition is of mental models. The concept of a 'mental model' has been defined in a plurality of ways and is similar yet distinct from a number of other related concepts such as a frame. However, the principles listed above help distinguish mental models from other types of mental representations. It is unclear whether or not the principles above will ever be able to be truly falsifiable. However, these principles are

congruent with other definitions used by cognitive engineers. Particularly, in terms of mental models being: imagistic, dynamic, and mapping onto something in the world (Klein & Hoffman, 2008).

The principles also illustrate the usefulness and necessity of mental models for many vital sensemaking functions such as error detection, diagnoses, and correction in intelligent tools. Mental models represent the dynamic relationships and interactions between different elements of an intelligent tool, and are used to understand the causal relations necessary for error diagnosis (Klein et al., 2007).

To summarize, the definition of mental models are internal representations of the external world. The structure of the mental model corresponds to the spatial, temporal, and causal relations of the elements perceived in the external world by using a combination of mental imagery and organized knowledge of concepts and relations. Finally, mental models can test hypotheses by running variations of existing mental models.

Mental models play a role in all macrocognitive functions. Particularly, in the macrocognitive phenomena sensemaking. Many aspects of error management requires various sensemaking operations which rest on the use of mental models. Good error management is vital for the future of effective and enduring intelligent tools. Therefore, the role of mental models within error management is a valuable place to explore.

Error Management of Intelligent Tools

The focus of the present research is on the role mental models play in complex human-machine systems. The modern world is becoming increasingly technologically advanced. Utilizing intelligent tools is frequently cheaper, more accurate, and reliable compared to human performance alone. However, when intelligent tools are not accurate and reliable it can be necessary to detect, diagnose, and correct it. This process is known as error management (McBride, Rogers, & Fisk, 2014). Understanding how people detect, diagnose, and correct errors is valuable for designers to create more optimized and adaptive systems. Unfortunately, despite the vital role of error management within complex human-machine systems, error management is still poorly understood (McBride et al., 2014). To the extent that operators have an accurate mental model of the intelligent tool they are operating they are better equipped to detect, diagnosis, and correct errors in the intelligent tool.

Quality mental models can be useful for error management. However, it should be noted that a mental model is not the representation of the intelligent tool itself, rather the internal representation that the user has created of that intelligent tool (Moray, 1999; Norman, 1983). Therefore, mental models often do not perfectly correspond to what it is representing. As a result, mental models are often not complete and inaccurate (Norman, 1983)[4]. However, through effective training mental models can be elaborated and

[4] See Norman (1983) for full discussion of system mental model challenges.

refined; increasing operator performance in detecting, diagnosing, and correcting errors. Therefore, analyzing how to effectively train operators to create more accurate mental models is valuable.

Summary

The vital role mental models play in performance in macrocognitive processes makes their evaluation necessary. Empirical evidence of mental models has been limited (Klein & Hoffman, 2008). I attempt to help fill this gap. Specifically, I attempt to gather evidence for four hypotheses. (1) MMM hypothesis; that performance of one sensemaking operation is predictive of performance of another sensemaking operation. (2) Through brief instruction and feedback, mental models are developed that involve understanding the relational structure between inter-correlated and independent feature(s). (3) Understanding of the relational structure of the features can be used to make error correction decisions. (4) The strategies that utilize the inter-correlated nature of the features can be recognized and verbalized by users.

Chapter 3: Experiment 1

Experiment 1 was designed to gather evidence for the first and second hypotheses. In order to gather evidence for the second hypothesis, Experiment 1 was designed to test learning of using independent differentially weighted features to make weather predictions (in the form of probability estimates). Analyzing accuracy of probability estimates was used to help infer the quality of participants' mental models of the simulated weather forecasting system. In order to gather evidence for the MMM hypothesis, Experiment 1 was also designed to test the relation between participants' probability estimates and evacuation decisions.

Methods

Participants. Twenty-four participants were recruited from the Michigan Technological University student subject pool. Students participated in the study for course credit.

Materials and Procedure. All experiments were programed and administered through the Psychology Experiment Building Language (PEBL) (Mueller & Piper, 2014).

Similar to the method used by Casteel (2016), participants were asked to imagine they were a plant manager during the task. Instructions: "In this task, you are a manager who is making decisions about whether to evacuate your facility, which is located on the eastern seaboard. There is a hurricane in the Atlantic, and you will need to decide, based on National Weather Service (NWS) information, the probability of whether the hurricane will come, and whether you should evacuate the facility."

On each trial participants were given 8 features of differentially weighted diagnostic information for the likelihood of a hurricane: wind speed above 74mph, rotating winds over the surface of the sea, rising sea level, relative humidity level of 850 hectopascals, falling pressure, temperature above 80°F, rough choppy sea, and overcast skies (see Table 1).[5] Each feature, reported by the simulated weather forecasting system, either increased or decreased the likelihood of a hurricane. Whether features increased or decreased the probability of a hurricane was indicated with the direction of an arrow (see Figure 1). For example, if wind was reported, it was either reported as a positive indicator with an up arrow and "wind speed above 74mph" or as a negative indicator with a down arrow and "wind speed below 74mph".

Indicator strength described the influence each feature had on the probability of the hurricane. For example, a "very good" indicator increased or decreased the probability of a hurricane much more than a feature with an indicator strength reported as "poor."

---

[5] How these features are related to forecasting in more naturalistic settings was not explained. The features were chosen based only on their face validity for being indicators of a hurricane. They are not representative of the complex dynamic nature of how hurricanes form. Weather forecasting in the real world is much more complex. The present research is limited by not analyzing sensemaking operations in naturalistic contexts.

Table 1 Materials used for Experiments: Features and their Weights

Feature Number	Features of Information	Indicator Strength
F1	Wind speed above 74mph	Very good
F2	Rotating winds over the surface of the sea	Very good
F3	Rising sea level	Good
F4	Relative humidity level of 850 Hectopascals	Fair
F5	Falling Pressures	Fair
F6	Temperature above 80°F	Poor
F7	Rough choppy sea	Very poor
F8	Overcast Skies	Not an indicator at all

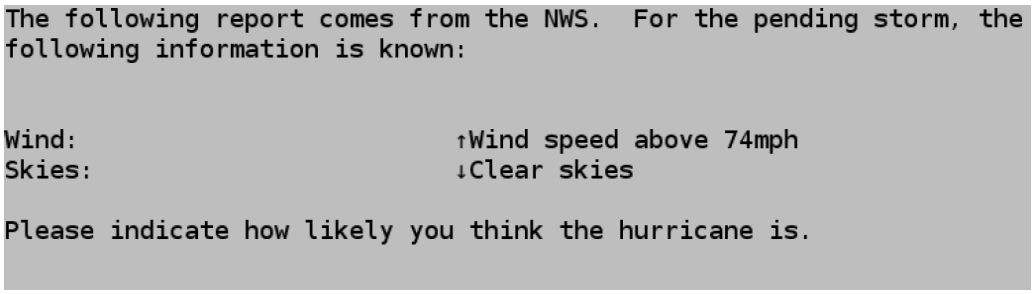

Figure 1: Example of typical message shown to participants on each trial.

Participants started with 10 practice trials. A $90(8^3)$ taguchi factorial design was used, meaning participants completed ninety trials with eight features that had three levels (positive, negative, absent) per feature. Using a taguchi design ensured that there was a unique feature set on each trial and that every possible combination of features was shown at least once.

Based on information provided by the weighted features, participants rated the likelihood of a hurricane from 0% to 100% on a thermometer in the upper right corner of their screen (see Figure 2). After estimating the probability of a hurricane, participants

indicated whether they should evacuate or stay. They were told they should evacuate if there was a high probability of a hurricane, however, that if a hurricane does not hit it would unnecessarily cost the company money and job performance would suffer.

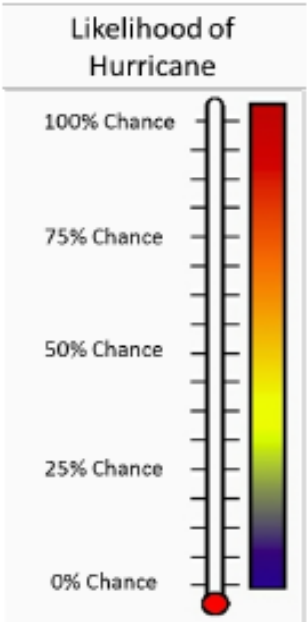

Figure 2: Likelihood assessment scale for the probability of a hurricane.

After participants made their two judgments (probability estimate and evacuation decision), both verbal and audio feedback was provided. A box was shown with a 15% range around the true estimate for the probability of a hurricane based on the weights of the features shown (see Table 2 for statistical model indicator strength)[6]. A beep would also sound if participants were within the range estimate. If correct, participants were

[6] The statistical model used is a simple linear model and is not representative of the real-world dynamic nature of weather. Models used in weather forecasting in more naturalistic settings are dynamic and therefore much more complex.

given a score of 100 points. After participants reviewed the feedback, they clicked either a "Good estimate. Click to continue" button or "Try to improve. Click to continue" button. The timing of each trial, which features were present, participants probability estimate of the likelihood of a hurricane, the model probability estimate of a hurricane (i.e. the true probability estimate), and participants evacuation decisions were recorded.

Table 2: Feature weight values used for statistical model.

Feature Number	Statistical model Indicator Strength
F1	.8
F2	.8
F3	.6
F4	.5
F5	.5
F6	.2
F7	.1
F8	.01

Results & Discussion

In order to help ensure data quality, each participants' accuracy was compared to what would be achieved only by chance. Accuracy was calculated by the number of times participants made probability estimates within the 15% band of the hidden statistical model probability estimate. No participant scored at or lower than chance. Accuracy ranged between 23% - 55% (M = .25, SD = .43), suggesting participants took the Experiment seriously and did not simply click through.

The features that were utilized by participants were analyzed. To reduce bias from non-linear data, a logodds transformation was conducted on participants' probability estimates. A multiple linear regression was performed to predict participants' probability estimates based on each of the features. Each feature was treated as an independent variable within the model. Results are shown in Table 3. Table 2 shows the weights of each feature for the hidden model probability estimate of a hurricane. Results suggest that almost all of the features significantly predict participants' probability estimates. Participants are likely not, therefore, utilizing the take the best heuristic (Gigerenzer & Goldstein, 1996). Indeed, of the eight differentially weighted features the only feature not a statistically significant predictor of participants' responses was the feature that participants were told was not an indicator (i.e. the only irrelevant feature). Consistent with hypothesis 2 this suggests participants created a mental model that contained the relational structure of all eight features when making probability estimates.

Table 3. Experiment 1 multiple linear regression

Variable	B	SE B	t	p
F1	.70	.03	23.78	<.01
F2	.56	.03	18.82	<.01
F3	.33	.03	11.31	<.01
F4	.27	.03	9.24	<.01
F5	.21	.03	7.05	.01
F6	.15	.03	4.99	<.01
F7	.07	.03	2.37	.02
F8	-.04	.03	-1.29	.20

In order to evaluate the relations of evacuation decision and probability estimates, a correlation was calculated, r =.81, N = 24, p <.01. The evacuation decision was aggregated according to the average evacuation decision for all participants on each trial. Consistent with hypothesis 1, results suggest that participants' evacuation decisions were closely related to their probability estimates. In other words if participants indicated a high probability of a hurricane they were more likely to make the decision to evacuate.

In summary, Experiment 1 suggested through brief instruction, visual feedback, and audio feedback that participants created a mental model of the system which related all eight differentially weighted features to the probability estimates consistent with the second hypothesis. This result provides some support for hypothesis 2. The strong correlation between participants' probability estimates and evacuation decisions suggests that the same knowledge (or *mental model*) used to make the probability estimates was also used make the evacuation decisions consistent with the MMM hypothesis. The result suggests participants are basing their evacuation decisions on their estimations by weighing a number of differentially weighted features into a probability estimate which suggests on a linear scale whether a hurricane is coming or not. However, how would an inaccurate and unreliable weather forecasting system influence participants' mental models? In more naturalistic settings of weather forecasting systems are not always entirely accurate (Berger, 2017). One notable example is when the Global Forecasting System (GFS) inaccurately predicted hurricane Sandys' day of landfall. An important aspect of error management is determining where an error could come from. Is the error a human error or a technological error? Accurately answering this question likely

influences how a user decides when to use or not use a system. Therefore, it is valuable to analyze error detection and error correction sensemaking operations.

Chapter 4: Experiment 2

Experiment 1 suggested that through feedback, participants learned to weigh independent and differentially weighted features to make probability estimates consistent with hypothesis 2. Some evidence was also gathered for the MMM hypothesis; evacuation decisions were closely related to probability estimates. Experiment 2 was designed to extend testing for other vital sensemaking operations. There were two vital operations experiment 2 was designed to gather data for. (1) Experiment 2 was designed to test for participants' ability to make a judgements about the source of an error (either themselves or the simulated forecasting system). (2) Experiment 2 was designed to test for participants' ability to correct an error in the feature report of the simulated weather forecasting system.

Methods

Participants. Data from seventeen participants were collected from Michigan Technological University student subject pool (N = 17). Students participated in the study for course credit.

Materials and Procedure. Materials and Procedures were similar to Experiment 1. However, the evacuation decision was removed and participants were told that there were some malfunctions in the "weather forecasting system's sensor report". They were told that the simulated weather forecasting system may report a false feature (i.e., a

positively indicated feature is actually a negatively indicated feature). Each trial presented all eight features. There were ten practice trials and seventy non-practice trials.

Half of the trials were error trials, meaning that inaccurate information was reported. If participants were *accurate* on a *non-error trial*, participants were given a score of 100 and moved on to the next trial. If participants were *inaccurate* on a *non-error trial* they were shown visual feedback with the 15% red band around the true estimate. Participants were asked to click either the "I was wrong" or "System malfunction" button. Regardless of the button they clicked they were told they were in fact wrong, and instructed to click the "ok" button to start a new trial.

If participants were *accurate* on an *error trial*, they were told their estimate was wrong and shown the visual feedback with the 15% red band around the true estimate. Participants were asked to click either the "I was wrong" or "System malfunction" button. Participants were then informed that the "weather forecasting system" had malfunctioned. Participants were then asked to select where the error had occurred by choosing which features were incorrect. Participants could choose one or all features but had to choose at least one feature before clicking the "ok" button, and then move on to the next trial. The timing of each trial, which features presented, participants probability estimate, the model probability estimate, blame choice, and which features selected to correct the system fault were all recorded.

Results and Discussion

Similar to Experiment 1, accuracy of probability estimates were compared to accuracy that would be achieved only by chance. No participants scored at or lower than chance. Accuracy ranged between 19% - 66% (M = .34, SD = .47).

The relations between average accuracy by participant in probability estimates and blame attribution across three different conditions (system correct human error, system incorrect human error, and system incorrect human correct) was analyzed. First the relation between average accuracy by participant and blame attribution on trials when the system was correct but participants made incorrect probability estimates was analyzed $r = -.03$, $p > .05$ (see Figure 3). This result suggests that there are no differences in blame attributions on the system correct human error condition based on probability estimate performance. This suggests that regardless of ability levels (or how good participants' mental models are) participants on this condition accurately blamed themselves.

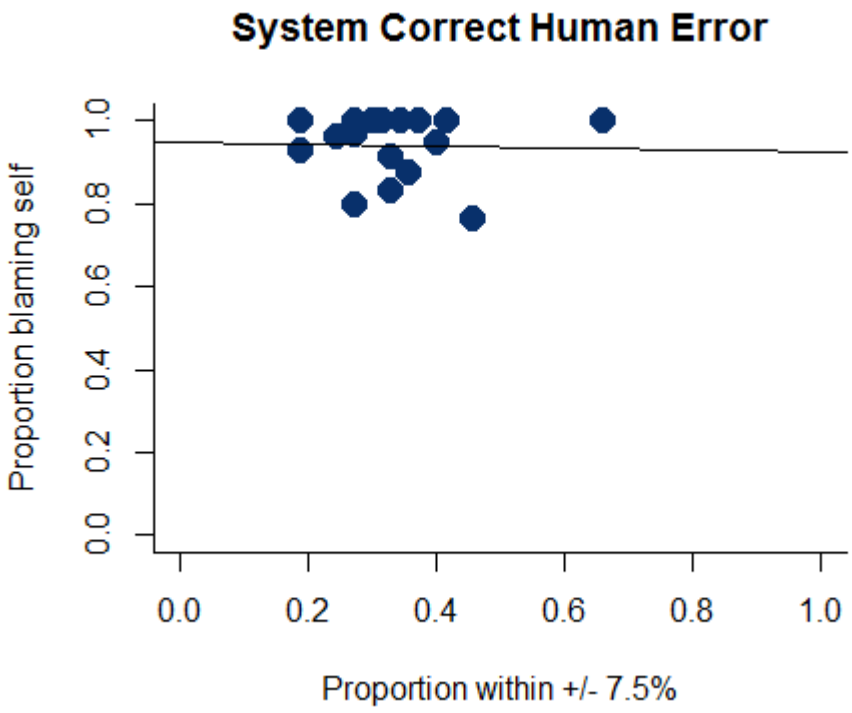

Figure 3: Relationship between average accuracy by participant and likelihood to blame themselves when the system was correct and participants made an error.

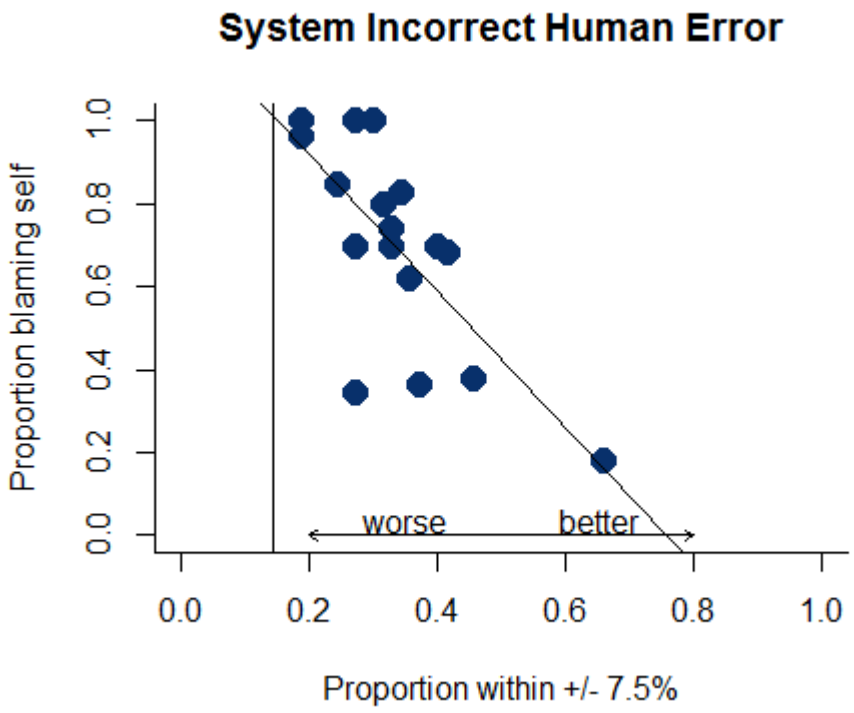

Figure 4: Relationship between average accuracy by participant and likelihood to blame themselves when the system was incorrect and participants made an error.

There was a large correlation between average accuracy by participant and blame attribution on trials when the system was incorrect and participants made an error $r = -.75$, $p < .01$ (see Figure 4). This result suggests that participants who generally perform better are less likely to blame themselves. In this condition there is no incorrect blame attribution, however, the result suggests that participants who have a better mental model of the system (as suggested by their probability estimate performance) are likely aware they have a good mental model of the system, compared to participants who have a poorer mental model.

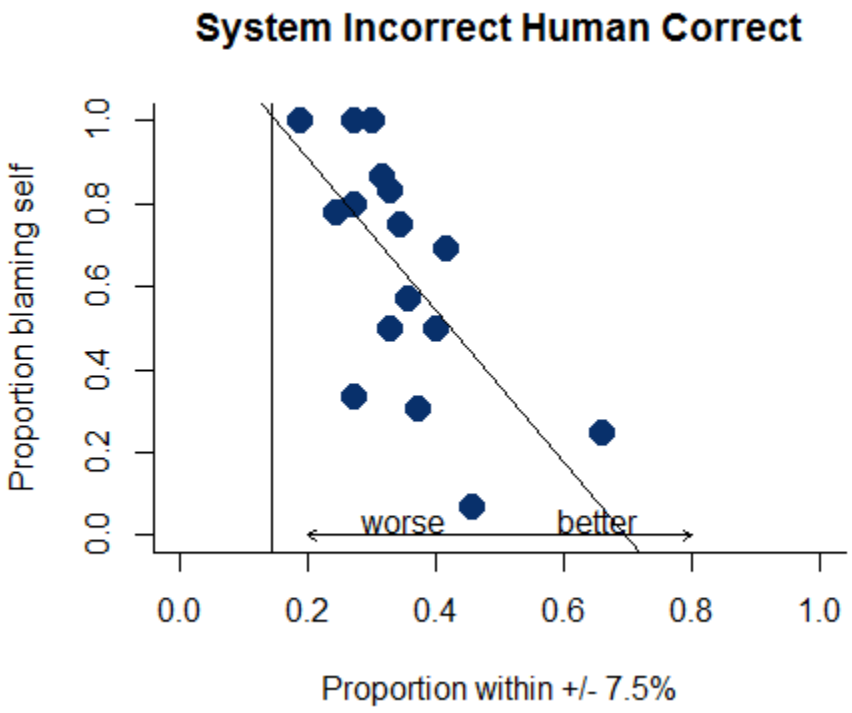

Figure 5: Relationship between average accuracy by participant and likelihood to blame themselves when the system was incorrect and participants made an error.

There was a large correlation between average accuracy by participant and blame attribution on trials when the system was incorrect and participants were correct in their probability estimates $r = -.70$, $p < .01$ (see Figure 5). Consistent with the previous result, this result suggests that participants who generally perform better are less likely to blame themselves. If participants generally perform worse this result suggests that they are likely aware of it, therefore, tend to blame themselves even when they are correct.

An overall correlation was performed between average accuracy by participant for detecting where the error occurred (i.e. blame attributions) and average accuracy by

participant for making probability estimates, which was a strong and statistically significant relation $r = .9$, $p < .01$. This result is consistent with the MMM hypothesis.

Average accuracy for correcting probability estimates by adjusting the direction of the sensor features ranged between 57% - 100% (M = .83, SD = .48). A correlation was performed between average accuracy by participant when making the initial probability estimates and average accuracy by participant for correcting the probability estimates $r = -.68$, $p < .01$. This results is not consistent with the MMM hypothesis.

Some of the results from Experiment 2 were consistent with the MMM hypothesis. Participants' accuracy at probability estimation was strongly associated with participants' ability to make accurate blame attributions. Results also suggested that participants who had a better mental model of the system (i.e. were more accurate in their probability estimates) made blame attributions that suggested they knew they had a good mental model of the system. While participants who had a poorer mental model of the system were more likely to blame themselves. However, other results did not support the MMM hypothesis. This suggests that the mental model created with initial learning through feedback while making weather predictions may not transfer to all sensemaking operations tested. Experiment 3 was designed to be a replication study of Experiment 2. Experiment 3 differed only by providing further error diagnosis feedback.

Chapter 5: Experiment 3

Consistent with the first hypothesis, Experiment 2 suggested that participants' accuracy at probability estimation was strongly associated with participants' ability to make accurate blame attributions. The goal of Experiment 3 was to replicate results from Experiment 2. The only difference between Experiment 2 and Experiment 3 was additional feature specific error diagnosis feedback.

Method

Participants. Data from twenty-seven participants were collected from the Michigan Technological University student subject pool. Students participated in the study for course credit. One participant's data were removed for not completing the experiment.

Materials and Procedure. Materials and Procedures were similar to that of Experiment 2. However, contrary to Experiment 2 after participants chose which features were inaccurately reported they were provided with visual feedback if they chose the correct features. The timing of each trial, which features presented, participants probability estimate, the model probability estimate, blame choice, and which features selected to correct the system fault were all recorded.

Results and Discussion

Similar to Experiments 1-2 average accuracy by participant was analyzed. No participant scored at or lower than chance. Accuracy ranged between 20% - 61% ($M = .34$, $SD = .48$).

Results were very similar to Experiment 2. The relation between average accuracy by participant and blame attribution on trials when the system was correct but participants made incorrect probability estimates was analyzed $r = .07$, $p > .05$. This result suggests that there are no differences in blame attributions on the system correct human error condition. There was a medium sized correlation between average accuracy by participant and blame attribution on trials when the system was incorrect and participants made an error $r = -.49$, $p < .01$. This result suggests that participants who generally perform better are less likely to blame themselves. There was a medium sized correlation between average accuracy by participant and blame attribution on trials when the system was incorrect and participants were correct in their probability estimates $r = -.51$, $p < .01$. This result also suggests that participants who generally perform better are less likely to blame themselves.

A correlation was performed across trials between average accuracy by participant for detecting where the error occurred and average accuracy by participant for making probability estimates was statistically significant $r = .92$, $p < .01$.

Average accuracy for correcting probability estimates by adjusting the direction of the sensor features ranged between 48% - 93% (M = .57, SD = .49). A correlation was performed between average accuracy by participant when making the initial probability estimates and average accuracy by participant for correcting the probability estimates $r = -.33$, n = 27, $p > .05$. Results are not consistent with the MMM hypothesis.

Results from Experiment 3 largely replicated results from Experiment 2. Experiments 4a-4b were designed to test for further valuable sensemaking operations.

Particularly, Experiments 4a-4b were designed to test for learning the relations between inter-correlated features as depicted in the example from Chapter 1. Research has been conducted on the use of negatively correlated features in decision making (Fasolo, McClelland, & Todd, 2007). However, *understanding* of inter-correlated features for the use of accomplishing complex goals has been relatively unexplored. Using inter-correlated features to make probability estimates are more closely related to how people make judgments and predictions across a variety of situations—including weather forecasting. Accurate use of inter-correlated features to make probability estimates, error detections, and error corrections requires unique mental arithmetic compared to independent features alone. Therefore, it is valuable to explore.

Chapter 6: Experiment 4a

Experiments 4a-4b are extensions of Experiments 1-3. Experiment 1 suggested that through feedback, participants learned to weigh independent and differentially weighted features to make weather predications. Some evidence was also gathered for the MMM hypothesis; evacuation decisions were closely related to probability estimates. Experiments 2-3 provided evidence that supports the notion that participants recognize how accurate their mental models are of the simulated weather forecasting system. Experiment 4a is designed to test whether participants can make probability estimates, detect errors, and correct errors by using inter-correlated features. Experiment 4a and 4b differ in two ways. The first difference was in how the hidden statistical model created the "true" probability estimates. In both Experiment 4a and 4b the first three features are inter-correlated with a correlation of .9. In Experiment 4a the weights of the first three features were added together to make the probability estimate. In Experiment 4b the value that appeared most often of the three inter-correlated feature was weighted and incorporated into the probability estimate. For example, if two of the three features were negative indicators and the third was a positive indicator (-1,-1, 1) the mode of the three features was incorporated into the model (-1). The second way the two Experiments differed was in Experiment 4b the shown probability estimate was removed from the screen when the error detection in the sensor report questions were asked.

Methods

Participants. Data from twenty-six participants were collected from the Michigan Technological University student subject pool. See Table 4 for basic demographic information.

Table 4: Demographics for Experiments 4a and 4b

		Experiment 4a	Experiment 4b
Age		*M* = 19.62 *SD* = 1.63 range 18-24	M = 20.16 *SD* = 177 range 18-23
Gender		54% (*N* = 14) Male	52% (*N* = 13) Male
Ethnicity			
	Caucasian	92% (*N* = 24)	80% (*N* = 20)
	Asian	8% (*N* = 2)	12% (*N* = 3)
	Black	0% (*N* = 0)	4%(N = 1)
	Hispanic	0% (*N* = 0)	4%(N = 1)

Materials and Procedure. Materials and procedure were similar to Experiments 1-3, with a few notable differences. In this task, simulated data was reported from a simulated weather forecasting system. In order to gather some ecologically validity, features utilized in this task were based on features from the actual Global Forecasting System (GFS) model ((NOAA), 2018). The real GFS is a model which combines data

from four separate sub-models: atmosphere, ocean, land/soil, and sea/ice. Five features were chosen from the GFS model: long waves, high waves, rotating winds over the surface of the sea, soil moisture 40-100 cm below the surface, and 50% cloud cover. Features were chosen based solely off of their face validity. The first three features were chosen from the ocean sub model, which would be the inter-correlated features. The independent feature (soil moisture) was chosen from the land/soil sub model. Finally, the last independent feature (50% cloud cover) was chosen from the atmospheric sub model.

The first three features were inter-correlated with a correlation of .9. Meaning during the practice trials the features reported information consistent with each other approximately 90% of the time. The two independent features had a correlation of less than .1 with the first three inter-correlated features and with each other.

Participants were given 40 practice trials with visual and audio feedback. Practice trials did not include error trials. Practice trials were designed for participants to gain a mental model of how the simulated weather forecasting system operated through instruction and feedback while making probability estimates by using inter-correlated and independent differentially weighted features. Participants' accuracy in making probability estimates was considered to be an empirical measure of how accurate participants' mental models were of the system.

Participants also responded to 96 error detection trials. During the error detection trials participants did not make probability estimates. In contrast, on each trial participants were shown a list of both inter-correlated and independent features (see

Table 5). The first three feature, because they were inter-correlated, were listed as having very good indicator strength when reported together.

Half of the 96 trials contained an error. An error occurred when one or more feature(s) were incorrectly reported, or when the simulated system calculation incorrectly calculated an estimate based on the features displayed, or both the features and the calculation was incorrect. On each trial participants were asked two questions. “Do you think there is an error in the sensors above?” “Suppose the sensors are correct, is there an error in the probability estimate?” Participants were asked to click on the ‘yes’ or ‘no’ buttons on each trial.

Participants were provided with feedback from a hypothetical forecasting system named after the actual “European Forecasting System” (EFS). The EFS reported if the simulated system did make an error in the sensor report. After clicking the ‘yes’ or ‘no’ button(s) participants were shown a new page with the features and probability estimates. Similar to previous experiments, participants were asked to choose any incorrectly reported features.

If the probability estimates were incorrect participants were asked to provide the accurate probability estimate. Similar to previous experiments visual feedback was displayed with a 15% red band around the true statistical model probability estimate.

Table 5: Materials for the proposed Experiment.

Cues of Information	Indicator Strength
Long waves	
High waves	Very Good
Rotating winds over the surface of the sea	
Soil moisture 40-100 cm below surface	Very Good
50% cloud cover	Fair

Immediately after completing the 70 error detection trials some post-hoc questions were asked (see Appendix A for post-hoc questions). Data from nineteen participants was gathered to assess their knowledge and computations for sensor error detection. Results were coded according to their sensitivity to the inter-correlated and independence of the features. After responding to the post-hoc questions, participants were asked to explain their thought process while responding to five different trials they had previously responded to. These examples consisted of both inter-correlated and independent features in the sensor report, in order to assess sensitivity to the independent and inter-correlated features.

Results & Discussion

For the practice trials, no participants scored at or lower than chance. Participants accuracy in their probability estimates made during the practice trials ranged between 33% - 68% (M = .49, SD = .5).

Based on data collected for empirical analyses alone, it is unclear how previous knowledge of a hurricane influences responses on these tasks. Therefore, the qualitative

data was analyzed to determine how previous knowledge of hurricanes and weather forecasting influenced responses on the tasks. Based on the assessment of the qualitative data, only three of the twenty three participants reported using their previous knowledge of hurricanes to make sensor error report decisions. The vast majority reported using feedback from the weather system to learn how to make correct error detection decisions and probability estimates.

Half of the ninety-six error trials had an error in the sensor report, as produced by the "Global Forecasting System." Accuracy for detecting an error in the sensor report ranged between 52% - 97% (M = .68, SD = .15). In order to accurately detect an error in the sensor report participants would need to incorporate the relational structure of the inter-correlated and independent features. Hypothesis 2 predicted that participants would incorporate the relational structure of the inter-correlated features. In order to test for sensitivity to feature correlations, first, a general linear model was conducted to compare the effect of (IV's) condition type on (DV) sensor error detection accuracy (See Table 6 for the different condition types). A chi-square analysis of deviance was performed on the model to test if the conditions explained variability in sensor error detection response more than chance $\chi^2 = 369.95$, df = 4, $p < .01$. Results suggest that condition type (as identified by Table 6) does influence sensor error detection responses. Based off of these results further analyses were conducted. A paired samples t-test was performed to test whether there was a statistically significant difference in error detection responses between independent feature inconsistent and inter-correlated feature inconsistent conditions. There was a significant difference in responses between when the

independent feature was inconsistent (M = .59, SD = .25) and when the inter-correlated feature was inconsistent (M = .37, SD = .21); $t(25) = 3.15$, $p < .01$. (A higher average indicates more "no error" responses to sensor error detection question.) Consistent with hypothesis 2, this result suggests that participants were sensitive to and understood the relational structure of the inter-correlated features.

Table 6: conditions to test for sensitivity to feature correlations

Condition	**Example**	**Number of Trials**
1 Features all the same	↑Long Waves, ↑ High Waves, ↑ Rotating Winds, ↑ Soil Moisture 40-100 cm below surface; ↓ Short Waves, ↓ Low Waves, ↓ Consistent Winds, ↓ Soil Moisture 10 cm below surface	24
Same plus missing feature	↑High Waves, ↑ Rotating Winds, ↑Soil Moisture 40-100 cm below surface	
2 Independent feature is inconsistent	↑ Long Waves, ↑ High Waves, ↑Rotating Winds, ↓ Soil Moisture 10 cm below surface; ↓ Short Waves, ↓ Low Waves, ↓ Consistent Winds, ↑Soil Moisture 40-100 cm below surface	24
Independent feature inconsistent plus correlated feature missing	↑ High Waves, ↑Rotating Winds, ↓ Soil Moisture 10 cm below surface; ↓ Low Waves, ↓ Consistent Winds, ↑Soil Moisture 40-100 cm below surface	
3 1 inter correlated feature is inconsistent	↓ Short Waves, ↑ High Waves, ↑ Rotating Winds, ↑ Soil Moisture 40-100 cm below surface	18
one correlated feature inconsistent plus independent variable missing	↓ Short Waves, ↑ High Waves, ↑ Rotating Winds	
4 1 independent and 1 correlated feature the same but 1 correlated feature inconsistent	↓ Short Waves, ↑ High Waves, ↑ Rotating Winds, ↓ Soil Moisture 10 cm below surface	30
Independent and 1 correlated feature the same but one correlated feature missing	↑ High Waves, ↓ Consistent Winds, ↑ Soil Moisture 40-100 cm below surface	

To the extent that participants responded differently to the independent feature inconsistent condition and the inter-correlated feature inconsistent condition then they are correctly incorporating the inter-correlated feature into their error detection decisions. In order to help test hypothesis 1, the difference between responses to condition 2 and condition 3 for each participant and the relation on performance when making probability estimates during the practice trials was calculated, $r = .1$, $p > .05$. Results were inconsistent with MMM hypothesis.

In condition 1 there is no error in the sensor report; all of the features are consistent. In condition 2 there is also no error in the sensor report; only the independent feature is inconsistent. Therefore, there should be no difference in error detection responses between conditions 1 and 2. If there is a difference in responses it suggests participants are incorrectly incorporating the irrelevant piece of information to make their error detection decisions. In order to test whether participants were incorporating the inconsistency of the independent feature into their error detection decision responses to condition 1 and condition 2 were compared. A paired sample t-test showed a statistically significant difference between condition 1 responses (M = .75, SD = .12) and condition 2 responses (M = .59, SD = .25) $t(25)=3.28$, $p < .01$.

In condition 3 there is an error in the sensor report; one of the inter-correlated features is inconsistent with another inter-correlated feature(s). In condition 4 there is also an error in the sensor report, however, in addition to one of the inter-correlated features being inconsistent the independent feature is also inconsistent. If participants are not inaccurately incorporating irrelevant information into their error detection decisions

then there should be little to no difference between responses to condition 3 and condition 4. In order to test whether participants inaccurately incorporated the irrelevant piece of information into their error detection decisions responses to condition 3 and condition 4 were compared. A paired sample t-test showed a statistically significant difference between responses to condition 3 (M = .37, SD = .21) and condition 4 (M = .28, SD = .16); $t(25) = 3.5$, $p < .01$. This suggests once again that participants are incorporating the irrelevant piece of information and that their mental models of the "GFS" system does not accurately represent the independence of the feature.

The relationship between performance during the practice trials and participants' ability to detect an error in the sensor report was analyzed. There was a medium sized correlation between the average for each participant during the practice trials and the average for each participant for detecting an error in the sensor report, $r = .4$, $p < .05$. Suggesting the knowledge of the system during the practice trials was related to the knowledge used to detect an error in the sensor report.

To the extent that there is a difference in responses between conditions 1 and 2 then participants are inaccurately incorporating the independent feature into their error detection decisions. A correlation was performed to determine if performance during the practice trials while making probability estimates predicted participants' likelihood to not incorporate the irrelevant piece of information. Inconsistent with hypothesis 1, participants performance when making probability estimates did not predict likelihood of incorporating the irrelevant piece of information, $r = -.1$, $p > .05$.

Hypothesis 4 suggested that participants would be able to recognize their strategies for detecting errors. Data from nineteen participants was gathered to assess their strategies at detecting an error in the sensor report. Across the seven post-hoc interview questions approximately half of the participants did not report using the inter-correlated nature of the features to detect an error in the sensor report (N = 10) (see Appendix B for example responses). In contrast, participants reported using inconsistencies between the feature weights and the GFS systems' probability estimates. In order to test hypothesis 4 (that strategies could be recognized) a correlation was performed between the total score from each qualitative question and average sensor error detection accuracy by participant. There was a statistically significant relation between error detection accuracy and qualitative score $r = .53$, $p < .05$. This result may be from an insufficient amount of qualitative data. It could be the case that participants' used two strategies to detect an error in the system report and only reported one. As will be described in later results participants who made incorrect error detection judgments still made corrections consistent with the inter-correlation. This may suggest that participants used different knowledge to detect an error compared to correcting the error, or that they may have used diverse strategies. However, the only strategies reported in the present qualitative assessment was using either the comparison of the feature weights to the probability estimates or using the feature inter-correlation. No participants reported using both.

Further analyses were conducted to test for the sensitivity to feature correlation by analyzing which features were chosen to correct the error(s) in the sensor report. In other

words, which features participants chose to correct. A feature was reported incorrectly when it reported incongruent information with another inter-correlated feature. For example, when one inter-correlated feature was a positive indicator of a hurricane and another inter-correlated feature was a negative indicator of a hurricane.

To help test hypothesis 2 and 3, two binomial tests were performed. The first compared the proportion of correct changes consistent with the feature correlation when there was no error in the sensor report and participants incorrectly detected an error. The second, tested the comparison on trials where an error was present (i.e. feature(s) were inconsistent). The first binomial test for the *no* error present trials, indicated that the proportion of instances where participants made changes that were consistent with the feature correlation of .58 was higher than the expected .5, $p < .01$. The second binomial test for the error present trials, indicated that the proportion of instances where participants made changes that were consistent with the feature correlation of .59 was higher than the expected .5, $p < .01$. Results suggest that participants made changes consistent with the feature correlation even when participants inaccurately indicated there was an error in the sensor report. This result is consistent with hypothesis 2 and 3.

Once again hypothesis 1 was tested to determine if performance when making probability estimates predicted performance at correcting errors in the sensor report. A correlation was performed between the average accuracy of sensor error correction by subject and performance during the practice trials $r = -.12$, $p > .05$. There was not a statistically significant correlation between performance for sensor error correction and performance during the practice trials.

Half of the ninety-six error trials had an error in the shown probability estimate, as produced by the "Global Forecasting System." Participants accuracy in *detecting* the error in the probability estimate ranged between 61% - 95% (M = .8, SD = .4). A Pearson correlation was performed between the average accuracy by participant during the practice trials and the average accuracy for detecting an error in the probability estimate during the error detection trials, $r = .4$, $p < .05$. Results suggest that participants' understanding (or mental model) of the simulated weather forecasting system, gained during the practice trials, was associated with their ability to detect an error in the probability estimates during the error detection trials.

The accuracy for participants *correcting* the probability estimat*e* ranged between 60% - 92% (M = .77, SD = .42). A Pearson correlation was also performed between the average accuracy by participant during the practice trials and the average accuracy for correcting the probability estimate during the error detection trials, $r = .65$, $p < .01$. Results are consistent with the multi-function mental model hypothesis.

There was a statistically significant relationship between participants ability to accurately detect an error in the probability estimate and their ability to correct the error in the probability estimate, $r = .59$, $p < .01$. Suggesting that if participants are able to accurately detect the error they will also be likely to correct the error. However, even when some participants some of the time inaccurately detect an error they still may be able to correct it.

Results from Experiment 4a supported the MMM hypothesis. Many of the error detection and correction decisions and judgements had a statistically significant relation

with participants' performance when making probability estimates. Results also supported the second hypothesis; error detection decisions were consistent with understanding the relational structure of the inter-correlated features. Results also supported the third hypothesis; error correction decisions were consistent with the inter-correlated nature of the features. Finally, results also supported the fourth hypothesis; reported strategies had a statistically significant relation to accurate error detection decisions in the sensor report. Experiment 4b was designed to replicate results from Experiment 4a. Experiment 4b differed in two ways. (1) The first was the way the hidden statistical model weighed the inter-correlated features. (2) Because a number of participants' reported using the inconsistencies between feature weights and the GFS systems reported probability estimates as their strategy for detecting an error in the sensor report the reported probability estimate was not shown on the same screen when participants were asked sensor error detection questions.

Chapter 7: Experiment 4b

Experiment 4b was designed to replicate results from Experiment 4a. There were two differences between Experiment 4a and 4b. The first was in how the hidden statistical model weighed the inter-correlated features. In Experiment 4a the weights of the first three features were added together to make the probability estimate. In Experiment 4b the value that appeared most often of the three inter-correlated feature was weighted and incorporated into the probability estimate. For example, if two of the three features were negative indicators (-1,-1) and the third was a positive indicator (1) the mode of the three features was incorporated into the model (-1). Weighing the inter-correlated features in this way is more consistent with how probability estimates work with inter-correlated features in more naturalistic environments. The second way Experiment 4b differed was in how participants' were presented with the error detection question in the sensor report. In Experiment 4a participants' were shown the probability estimate and the sensor error detection question on the same screen. Qualitative analyses suggested a number of participants' used inconsistencies between feature weights and the GFS systems reported probability estimates as their strategy for detecting an error in the sensor report. Therefore, error detection questions were not displayed on the same screen. Participants would first respond to the sensor report error detection question and only after responding were they shown the probability estimate.

Methods

Participants. Twenty-six participants were collected from the Michigan Technological University student subject pool. Demographics are reported in Table 4.

Materials and Procedure. Materials and procedure were the same as those reported in Experiment 4a.

Results and Discussion

For the practice trials, no participants scored at or lower than chance. Accuracy ranged between 22% - 67% (M = .43, SD = .49).

Accuracy for detecting an error in the sensor report ranged between 52% - 94% (M = .71, SD = .45). In order to test for sensitivity to feature correlations a general linear model was performed to compare the effect of (IV's) condition type on (DV) sensor error detection accuracy (See Table 6 for the different condition types). A chi-square analysis of deviance was performed on the model to test if the conditions explained variability in sensor error detection response more than chance $\chi^2 = 60.60$, df = 4, $p < .01$. Results suggest that condition type (as identified by Table 6) does influence sensor error detection responses. Therefore, further analyses were conducted to test for sensitivity to the inter-correlated nature of the features when making error detection decisions. A paired samples t-test was conducted to test whether there was a statistically significant difference in error detection responses between independent feature inconsistent and inter-correlated feature inconsistent conditions. Results were similar to Experiment 4a, there was a significant difference in responses between when the

independent feature was inconsistent (M = .62, SD = .34) and when the inter-correlated feature was inconsistent (M = .34, SD = .33); $t(24) = 2.94$, $p < .01$. Consistent with Hypothesis 2, this result suggests that participants were sensitive to and understood the relational structure of the inter-correlated features.

To the extent that participants responded differently to the independent feature inconsistent (condition 2) and the inter-correlated feature inconsistent (condition 3) then they are correctly incorporating the inter-correlated feature into their error detection decisions. In order to help test hypothesis 1, the difference between responses to condition 2 and condition 3 for each participant and the relation on performance when making probability estimates during the practice trials was calculated, $r = .44$, $p < .05$. Results were inconsistent with hypothesis 1.

In condition 1 there is no error in the sensor report; all of the features are consistent. In condition 2 there is also no error in the sensor report; only the independent feature is inconsistent. Therefore, there should be no difference in error detection responses between conditions 1 and 2. If there is a difference in responses it suggests participants are incorrectly incorporating the irrelevant piece of information to make their error detection decisions. In order to test whether participants were incorporating the inconsistency of the independent feature into their error detection decision responses to condition 1 and condition 2 were compared. A paired sample t-test showed a statistically significant difference between condition 1 responses (M = .92, SD = .11) and condition 2 responses (M = .62, SD = .34) $t(24)=4.43$, $p < .01$.

In condition 3 there is an error in the sensor report; one of the inter-correlated features is inconsistent with another inter-correlated feature(s). In condition 4 there is also an error in the sensor report, however, in addition to one of the inter-correlated features being inconsistent the independent feature is also inconsistent. If participants are not inaccurately incorporating irrelevant information into their error detection decisions then there should be little to no difference between responses to condition 3 and condition 4. In order to test whether participants inaccurately incorporated the irrelevant piece of information into their error detection decisions responses to condition 3 and condition 4 were compared. A paired sample t-test showed a statistically significant difference between responses to condition 3 (M = .35, SD = .33) and condition 4 (M = .37, SD = .30); $t(24) = -0.59$, $p > .01$. This suggests that the independent feature does not influence error detection responses when an error is present.

To the extent that there is a difference in responses between conditions 1 and 2 then participants are inaccurately incorporating the independent feature into their error detection decisions. A correlation was performed to determine if performance during the practice trials while making probability estimates predicted participants' likelihood to not incorporate the irrelevant piece of information. Inconsistent with hypothesis 1, participants performance when making probability estimates did not predict likelihood of incorporating the irrelevant piece of information, $r = -.22$, $p > .05$.

The relationship between performance during the practice trials and participants' ability to detect an error in the sensor report was analyzed. There was a medium sized

correlation between the average for each participant during the practice trials and the average for each participant for detecting an error in the sensor report, $r = .49$, $p < .05$.

Results suggest participants are sensitive to the inter-correlated nature of the features. In order to help test hypothesis 4 which suggests participants are able to recognize their strategies for detecting errors qualitative data was analyzed. Seven post-hoc interview questions were asked (see appendix A for post-hoc questions). All of the participants interviewed (N = 11) provided responses consistent with understanding the inter-correlation between the features. Since all participants reported using the inter-correlated nature of the features to make their error detection responses the relation between strategies used to detect an error in the sensor report and performance when making probability estimates during the practice trials could not be analyzed. However, overall performance at detecting error in the sensor report is higher for experiment 4a compared to 4b which is consistent with the results reflected in the qualitative assessment. In contrast to Experiment 4a, more participants also reported using only the inconsistency between the first three inter-correlated features, which is consistent with the hidden statistical model. In Experiment 4a many participants reported using a majority rules strategy, regardless of whether the feature was independent or not. However, across both experiments participants reported using the independent feature as a tie breaker for choosing which inter-correlated feature was reported incorrectly.

Further analyses were conducted to test for the sensitivity to feature correlation by analyzing which features were chosen to correct the error(s) in the sensor report. In other words which features participants chose to correct. Two binomial tests were performed,

one for comparing the proportion of correct changes consistent with the feature correlation when there was no error in the sensor report and participants incorrectly detected an error and the second one for testing the comparison of the same proportion but on trials where an error was present (i.e. feature(s) were inconsistent). The first binomial test for the *no* error present trials, indicated that the proportion of instances where participants made changes that were consistent with the feature correlation of .81 was higher than the expected .5, $p < .01$. The second binomial test for the error present trials, indicated that the proportion of instances where participants made changes that were consistent with the feature correlation of .83 was higher than the expected .5, $p < .01$. Results suggest that participants made changes consistent with the feature correlation even when participants inaccurately indicated there was an error in the sensor report when there was none.

A correlation was performed between the average accuracy of sensor error correction by subject and performance during the practice trials $r = .26$, $p > .05$. There was not a statistically significant correlation between performance for sensor error correction and performance during the practice trials.

Participants accuracy in detecting the error in the probability estimate ranged between 44% - 99% (M = .71, SD = .45). A Pearson correlation was performed between the average accuracy by participant during the practice trials and the average accuracy for detecting an error in the probability estimate during the error detection trials, $r = .67$, $p < .01$. Results suggest that participants' understanding of the system, gained during the

practice trials, was associated with their ability to detect an error in probability estimates during the error detection trials.

The accuracy for participants correcting the probability estimat*e* ranged between 54% - 92% (M = .75, SD = .43). A Pearson correlation was performed between the average accuracy by participant during the practice trials and the average accuracy for correcting the probability estimate during the error detection trials, $r = .3$, $p > .05$. Results are not consistent with the multifunction mental model hypothesis.

There was a statistically significant relationship between participants ability to accurately detect an error in the probability estimate and their ability to correct the error in the probability estimate, $r = .73$, $p < .01$.

Based on quantitative results from novices alone it is unclear how specialized knowledge would influence performance on many of these sensemaking operations. Therefore, an attempt was made to collect data from experts in weather forecasting. Only one participant was recruited. The participant was a coursework completed PhD student in atmospheric sciences. In addition to asking this participant to run asking this participant the seven post-hoc interview questions and asking them to explain their thought processes while detecting error in the sensor report they were also asked specific questions about their knowledge of weather forecasting and how it related to the tasks. The participant indicated that their specialized knowledge did not influence how they responded to the task. The participant was also asked about the ecological validity of the features to how weather operated in the real world. They reported that at least some of the features and their weights was consistent with what they knew about how weather

operated, particularly, for the area where the task was held. Cloud cover would be a poor indicator because in the area where the task was given it is frequently cloudy, therefore, not a good indicator of a storm.

Results in Experiment 4b closely follow results found in Experiment 4a. Table 9, shows the analyses for testing for the multifunction mental model hypothesis across all experiment. Evidence was found across many of the sensemaking operations. However, not all analyses supported the MMM hypothesis. For example, average accuracy in probability estimates and correcting probability estimates in Experiment 4b. Explanations for the mixed results will be discussed in greater detail in Chapter 8.

Chapter 8: General Discussion

The present research investigated combining methods from both micro- and macrocognitive paradigms in order to create unique advantages. The primary process under investigation, sensemaking, and its supporting function mental models, are not traditionally evaluated using empirical methods. One of the goals of this research was to take another step closer to complimenting existing methods used to study mental models, such as Cognitive Task Analysis (Klein & Hoffman, 2008). In this thesis quantitative and qualitative data was used to infer how people think about the system they were operating. The structure of participants' mental models was inferred based on their performance while operating the simulated weather forecasting system and responses to qualitative assessment.

There were four hypotheses of the present research. (1) MMM hypothesis; performance of one sensemaking operation is predictive of performance of other related sensemaking operations. (2) Through brief instruction and feedback, mental models are developed that involve understanding the relational structure between inter-correlated and independent feature(s). (3) Understanding of the relational structure of the features can be used to make error correction decisions. (4) The strategies that utilize the inter-correlated nature of the features can be recognized and verbalized by users.

Experiment 1 suggested through brief instruction, visual feedback, and audio feedback that participants created a mental model of the system which related all eight differentially weighted features to the probability estimates. This evidence was consistent with hypothesis 2. Some evidence was also gathered for the MMM hypothesis;

evacuation decisions were closely related to probability estimates. This suggests that initial learning through feedback from making probability estimates was also used to make evacuation decisions. Experiment 2 was designed to extend testing for other vital sensemaking operations. Consistent with the MMM hypothesis Experiment 2 suggested that participants' mental model of the simulated system while making probability estimates was strongly associated with participants' ability to make correct blame attributions. Results from Experiment 3 largely replicated results from Experiment 2.

Experiments 4a-4b were extensions of Experiments 1-3. There were two differences between Experiments 4a and 4b. (1) How the hidden statistical model created the "true" probability estimates. (2) In Experiment 4b the probability estimate was removed from the screen when the error detection in the sensor report questions were asked. Experiments 4a and 4b were designed to test whether participants can make probability estimates, detect, diagnose, and correct errors by using inter-correlated features. Making accurate error detection, diagnosis, and correction determinations requires a different mental strategy and computation compared to independent features alone, as illustrated in the example in Chapter 1. Participants' accuracy in making their weather predictions (in the form of probability estimates) was used as an empirical measure of the quality of participants' mental models of the simulated weather forecasting system in addition to responses from the qualitative assessment.

Results from both Experiments 4a & 4b supported the second hypothesis; performance when making error detection decisions was consistent with understanding the relational structure between the inter-correlated and independent features. However,

surprisingly results also suggested participants incorrectly incorporated the independent and irrelevant piece of information when making their error detection decisions. Results from both Experiments also supported hypothesis 3, error correction decisions were made by incorporating the relational structure of the inter-correlated features. Assessment of qualitative data suggested that as a result of manipulations between Experiment 4a and 4b participants were more likely to learn the accurate relations between features (i.e. have a more accurate mental model of the simulated system) in Experiment 4b compared to 4a. However, results from Experiment 4a was still consistent with the fourth hypothesis; there was a statistically significant relation between verbalized strategies and performance when making error detection decisions. While results from both Experiments supported hypotheses 2-4 the results for the first hypothesis is somewhat mixed.

Across experiments results generally supported the MMM hypothesis; performance on one task was predictive of performance on other related tasks (see Table 9). However, this was not consistent for each related task. Each of these relations will be briefly reviewed. Do to the nature of how the hidden statistical model produced the "true" probability estimate in Experiment 4a, participants did not need to understand the inter-correlated nature of the features to make accurate probability estimates. Therefore, a participant could be fairly accurate while making probability estimates and be entirely unaware of the inter-correlated nature of the features. This likely explains the lack of a statistically significant relation between performance when making probability estimates and sensitivity to the inter-correlated nature of the features when making error detection

decisions for Experiment 4a. This result is also consistent with operating an intelligent tool in more naturalistic settings. Operators could have aspects of their mental model of the intelligent tool that are entirely inaccurate and still be able to operate the tool for certain sensemaking operations. The second relation that did not support the MMM hypothesis was between performance when making probability estimates and performance when making sensor error corrections. The lack of a statistically significant relation could suggest that some functions transfer during initial learning of the task while others do not. Performance when making sensor error corrections could require further knowledge or a better mental model of the system not achieved through feedback while making probability estimates. Future studies could test if initial learning while making error corrections transfers to making probability estimates. Other possible explanation could be some participants misunderstood the task. However, the average at making accurate error corrections was higher than chance. Finally, the lack of a statistically significant relation between performance when making probability estimates and correcting probability estimates in Experiment 4b could be from an outlier, insufficient sample size, randomness, or once again suggest that the knowledge required to make the probability estimates is unique from the knowledge required to correct the probability estimates. Future studies should be done to help test for the relations analyzed in this thesis.

Table 9: Results across experiments testing for evidence for the MMM hypothesis.

Function	Exp1	Exp2	Exp3	Exp4a	Exp4b
Probability Estimates and Evacuation Decisions	$r = .81^{**}$				
Probability Estimates and blame attributions		$r = .9^{**}$	$r = .92^{**}$		
Probability Estimates and using sensor feature weights to correct probability estimates		$r = -.68^{**}$	$r = -.34$		
Probability Estimates and sensitivity to inter-correlation				$r = .1$	$r = .44^{*}$
Probability Estimates and sensor error detection				$r = .4^{*}$	$r = .49^{*}$
Probability Estimates and sensor error correction				$r = -.12$	$r = .26$
Probability Estimates and error detection in probability estimates				$r = .4^{*}$	$r = .67^{**}$
Probability Estimates and error correction in probability estimates				$r = .65^{**}$	$r = .3$
Qualitative data on sensitivity to feature correlation and sensor error detection accuracy				$r = .53^{*}$	*NA*

$^{**}p < .01$;* $p < .05$.

Since the 1970's there has been a concern about all the different microcognitive processes fitting together (Newell, 1973). Specifically, by becoming increasingly narrow in investigation, there is concern that there will be little transfer or generalizability (Gozli, 2017). Unfortunately, not much has changed (Hommel & Colzato, 2015). Experimental researchers often only study microcognition outside of the larger process that they are supporting (Hommel & Colzato, 2015). However, the present research provides some evidence that suggests a larger cognitive system can be analyzed within the lab; combing advantages of empirical analysis, systematic analysis, and qualitative assessment.

Error detection, diagnosis, and correction are a part of good decision making within human-machine systems. Mental models are necessary for system error detection, diagnosis, correction, and many other vital functions. To the extent that we have a better understanding of how people form and use mental models, we can more adequately enable people to perform more efficiently and effectively in changing and unexpected environments. Across experiments results generally supported the MMM hypothesis. If the MMM hypothesis is true then there may be implications for training and learning transfer. Performance for one cognitive operation while operating an intelligent tool was predictive of many other cognitive processes when operating the same intelligent tool. This is consistent with research being conducted on Experiential User Guides (EUG); which suggests that training in some sensemaking operations (such as error detection and diagnosis) helps refine operators mental models and therefore improve performance for other sensemaking operations while operating the same intelligent tool (Mueller & Klein,

2011). Consistent with this research the present results indicate there is a certain subset of organized knowledge (or mental model) of the intelligent tool users are interacting with that can be built and refined and adapted to different related tasks when interacting with an intelligent tool. Therefore, training in multiple sensemaking operations may be useful for refining the mental model of the intelligent tool and that refinement will likely increase performance of other tasks when operating the intelligent tool.

This research also expands upon previous research conducted using inter-correlated features. Previous research has suggested people do incorporate negatively correlated features in their decision making as measured by an increase in deliberation time (Fasolo et al., 2007). However, learning and the use of inter-correlated features for accomplishing complex goals has been relatively unexplored. This use of inter-correlated features impacts all of the sensemaking operations. Results from Experiments 4a-4b suggest that participants can learn associations between weighted inter-correlated features and therefore incorporate this understanding into their mental models of the intelligent tool they are operating.

In addition to learning the inter-correlated nature of the features, results also suggested that participants use some irrelevant information to make error detection and correction decisions, despite never being provided with feedback or instruction to do so. This could be the result of the initial mental model and frame participants come into the lab with before they even start the task. The mental model within the data/frame theory contains background knowledge which is valuable for explaining how the system operates. It could be the case that participants already have a frame and mental model that

related all of the features together and not enough feedback was provided that challenged their frame and mental model to be refined (Klein et al., 2006b). Many laboratory studies do not consider learning and prediction by using inter-correlated vs. independent features (e.g. Gluck et al. (2002)). Future microcognitive studies may need to control for interpretation of features being inter-correlated. Future research should could also test whether it is generally adaptive to have an initial frame or mental model that contains a structure of inter-correlated features.

Across experiments there was variability in participants' performance when making probability estimates and no participant was completely accurate. In verbal reports some participants described having different baselines for creating there probability estimates. Some participants started at zero before incorporating information from the sensor report, some started at .5, while others started at .75. This is consistent with previous research on the use of improper linear models; people generally perform poorly when making predictions from integrating information (Dawes, 1979). Therefore, results help support the notion of using proper linear models.

Limitations

There are some notable limitations to the present research. First, the sample was taken from an undergraduate college population. It is unclear whether results will generalize to other populations. It is also unclear how results would generalize to other more naturalistic settings. Future research should validate these finding in more naturalistic settings with experts. Experts may be better equipped to ignore the irrelevant piece of information when making error detection and correction decisions.

Conclusion

Sensemaking is a vital process for a number of diverse operations, but little research has been conducted on sensemaking within the lab. This thesis describes research on studying a macrocognition process within a microcognition world. By combing methods from both micro- and macrocognitive paradigms future research will provide useful insight into how to create trainings and interventions to make sociotechnical systems more efficient and enduring. Based on data across four experiments results generally supported the MMM hypothesis. This has implications for training; training in multiple sensemaking operations may be useful for refining the mental model of the intelligent tool and that refinement will likely increase performance of other tasks when operating the intelligent tool.

References

(NOAA), N. C. F. E. I. (2018). Global Forecast System (GFS) Retrieved from https://www.ncdc.noaa.gov/data-access/model-data/model-datasets/global-forcast-system-gfs. https://www.ncdc.noaa.gov/data-access/model-data/model-datasets/global-forcast-system-gfs

Alberts, D., & Garstka, J. (2004). Network centric operations conceptual framework version 2.0. *US Office of Force Transformation and Office of the Assistant Secretary of Defense for Networks and Information Integration, US Department of Defense, Tech. Rep*.

Ancona, D. (2012). Framing and Acting in the Unknown. *S. Snook, N. Nohria, & R. Khurana, The Handbook for Teaching Leadership*, 3-19.

Berger, E. (2017). US forecast models have been pretty terrible during Hurricane Irma. Retrieved from https://arstechnica.com/science/2017/09/us-forecast-models-have-been-pretty-terrible-during-hurricane-irma/

Cacciabue, P. C., & Hollnagel, E. (1995). Simulation of cognition: Applications. *Expertise and technology: Cognition and human-computer cooperation*, 55-73.

Casteel, M. A. (2016). Communicating Increased Risk: An Empirical Investigation of the National Weather Service's Impact-Based Warnings. *Weather Climate and Society, 8*(3), 219-232. doi:10.1175/Wcas-D-15-0044.1

Craik, K. J. W. (1943). *The nature of explanation*. Cambridge Eng.: University Press.

Crandall, B., Klein, G. A., & Hoffman, R. R. (2006). *Working minds : a practitioner's guide to cognitive task analysis*. Cambridge, Mass.: MIT Press.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34*(7), 571.

Dervin, B. (1983). *An overview of sense-making research: Concepts, methods, and results to date.* Paper presented at the International Communications Association Dallas. https://www.ideals.illinois.edu/bitstream/handle/2142/2281/Dervin83a.htm

Dervin, B., & Naumer, C. (2009). Sense-making. *Encyclopedia of communication theory* (2 ed., pp. 876-880).

Ebbinghaus, H. (1913). *Memory: A contribution to experimental psychology*: Genesis Publishing Pvt Ltd.

Fallon, C. K., Murphy, A. K., Zimmerman, L., & Mueller, S. T. (2010). *The calibration of trust in an automated system: A sensemaking process.* Paper presented at the Collaborative Technologies and Systems (CTS), 2010 International Symposium.

Fasolo, B., McClelland, G. H., & Todd, P. M. (2007). Escaping the tyranny of choice: When fewer attributes make choice easier. *Marketing Theory, 7*(1), 13-26.

Fitts, P. M. (1946). German applied psychology during World War II. *American Psychologist, 1*(5), 151-161.

Forrester, J. W. (1971). Counterintuitive Behavior of Social Systems. *Theory and Decision, 2*(2), 109-140. doi:10.1177/003754977101600202

Gentner, D., & Stevens, A. L. (2014). *Mental models*. New York, NY: Psychology Press.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychological review, 103*(4), 650-669.

Gluck, M. A., & Bower, G. H. (1988). From Conditioning to Category Learning - an Adaptive Network Model. *Journal of Experimental Psychology-General, 117*(3), 227-247. doi:Doi 10.1037//0096-3445.117.3.227

Gluck, M. A., Shohamy, D., & Myers, C. (2002). How do people solve the "weather prediction" task?: Individual variability in strategies for probabilistic category learning. *Learning & Memory, 9*(6), 408-418. doi:10.1101/lm.45202

Gozli, D. G. (2017). Building Blocks of Psychology: on Remaking the Unkept Promises of Early Schools. *Integrative Psychological and Behavioral Science*, 1-24.

Hoffman, R. R., LaDue, D. S., Trafton, J. G., Mogil, H. M., & Roebber, P. J. (2017). *Minding the weather: How expert forecasters think*. Cambridge, Massachusetts: MIT Press.

Hommel, B., & Colzato, L. S. (2015). Learning from history: the need for a synthetic approach to human cognition. *Frontiers in psychology, 6*, 1435.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness*: Harvard University Press.

Johnson-Laird, P. N. (2001). Mental models and deduction. *Trends in cognitive sciences, 5*(10), 434-442.

Johnson-Laird, P. N. (2004). Mental models and reasoning. In R. Sternberg & P. L. Jacqueline (Eds.), *The nature of reasoning* (pp. 169-204). Cambridge: Cambridge University Press.

Johnson-Laird, P. N. (2005). Mental models and thought. In K. J. Holyoak & R. G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 185-208). Cambridge: Cambridge University Press.

Johnson-Laird, P. N. (2006). *How we reason*. New York: Oxford University Press.

Jones, N. A., Ross, H., Lynam, T., Perez, P., & Leitch, A. (2011). Mental Models: An Interdisciplinary Synthesis of Theory and Methods. *Ecology and Society, 16*(1), 46. Retrieved from <Go to ISI>://WOS:000289081200012

Kaste, K. P. (2012). *Naturalistic Study Examining the Data/Frame Model of Sensemaking by Assessing Experts in Complex, Time-Pressured Aviation Domains.* (Master's thesis), Embry-Riddle Aeronautical University.

Kingstone, A., Smilek, D., & Eastwood, J. D. (2008). Cognitive Ethology: A new approach for studying human cognition. *British Journal of Psychology, 99*, 317-340. doi:10.1348/000712607x251243

Klein, G., & Hoffman, R. R. (2008). Macrocognition, mental models, and cognitive task analysis methodology. In J. M. Schraagen (Ed.), *Naturalistic decision making and macrocognition* (pp. 57-80). Burlington, VT: Ashgate Publishing Company.

Klein, G., Moon, B., & Hoffman, R. R. (2006a). Making sense of sensemaking 1: Alternative persepectives. *IEEE Intelligent Systems, 21*(4), 70-73. doi:10.1109/Mis.2006.75

Klein, G., Moon, B., & Hoffman, R. R. (2006b). Making sense of sensemaking 2: A macrocognitive model. *IEEE Intelligent Systems, 21*(5), 88-92. Retrieved from <Go to ISI>://WOS:000241163500018

Klein, G., Phillips, J. K., Rall, E. L., & Peluso, D. A. (2007). *A data-frame theory of sensemaking.* Paper presented at the Expertise out of context: Proceedings of the sixth international conference on naturalistic decision making, New York, NY.

Klein, G., Pliske, R., Crandall, B., & Woods, D. D. (2005). Problem detection. *Cognition, Technology & Work, 7*(1), 14-28.

Klein, G., Ross, K. G., Moon, B. M., Klein, D. E., Hoffman, R. R., & Hollnagel, E. (2003). Macrocognition. *IEEE Intelligent Systems, 18*(3), 81-85. doi:Doi 10.1109/Mis.2003.1200735

McBride, S. E., Rogers, W. A., & Fisk, A. D. (2014). Understanding human management of automation errors. *Theor Issues Ergon Sci, 15*(6), 545-577. doi:10.1080/1463922X.2013.817625

McDermott, R. (2011). Internal and external validity. In J. N. Druckman, D. P. Green, J. H. Kuklinski, & A. Lupia (Eds.), *Cambridge handbook of experimental political science* (pp. 27-40). Cambridge: Cambridge University Press.

Moray, N. (1999). Mental models in theory and practice. In D. Gopher & A. Koriat (Eds.), *Attention and Performance XVII: Cognitive regulation of performance Interaction of theory and application* (Vol. 17, pp. 223-258). Cambridge, MA: MIT Press.

Mueller, S. T., & Klein, G. (2011). Improving users' mental models of intelligent software tools. *IEEE Intelligent Systems, 26*(2), 77-83.

Mueller, S. T., & Piper, B. J. (2014). The Psychology Experiment Building Language (PEBL) and PEBL Test Battery. *Journal of Neuroscience Methods, 222*, 250-259. doi:10.1016/j.jneumeth.2013.10.024

Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In W. G. Chase (Ed.), *Visual information processing* (pp. 283-310). New York: Academic Press.

Norman, D. A. (1983). Some observations on mental models. In D. Gentner & A. Stevens (Eds.), *Mental models* (Vol. 7, pp. 7-14). New York, NY: Psychology Press.

Pirolli, P., & Russell, D. M. (2011). Introduction to this Special Issue on Sensemaking. *Human-Computer Interaction, 26*(1-2), 1-8. doi:10.1080/07370024.2011.556557

Revell, K. M., & Stanton, N. A. (2012). Models of models: filtering and bias rings in depiction of knowledge structures and their implications for design. *Ergonomics, 55*(9), 1073-1092.

Richardson, M., & Ball, L. J. (2009). Internal representations, external representations and ergonomics: towards a theoretical integration. *Theoretical Issues in Ergonomics Science, 10*(4), 335-376.

Rouse, W. B., & Morris, N. M. (1986). On Looking into the Black-Box - Prospects and Limits in the Search for Mental Models. *Psychological Bulletin, 100*(3), 349-363. doi:Doi 10.1037//0033-2909.100.3.349

Schraagen, J. M., Klein, G., & Hoffman, R. R. (2008). The macrocognition framework of naturalistic decision making. In J. M. Schraagen, L. Militello, T. Ormerod, & R.

Lipshitz (Eds.), *Naturalistic decision making and macrocognition*. Burlington, VT: Ashgate Publishing Limited.

Schumacher, R. M., & Czerwinski, M. P. (1992). Mental models and the acquisition of expert knowledge. In R. Hoffman (Ed.), *The psychology of expertise* (pp. 61-79). New York, NY: Psychology Press.

Shipley, T. (1961). *Classics in psychology*. New York: Philosophical Library.

Smieszek, H., & Rußwinkel, N. (2013). *Micro-cognition and macro-cognition: trying to bridge the gap.* Paper presented at the Proceedings of the 10th Berlin Workshop on Human-Machine Systems: Foundations and Applications of Human-Machine Interaction, Berlin.

Starbuck, W. H., & Milliken, F. J. (1988). Executives' perceptual filters: What they notice and how they make sense. In D. Hambrick (Ed.), *The Executive Effect: Concepts and Methods for Studying Top Managers* (pp. 35-65). Greenwich: JAI Press.

Vosniadou, S., & Brewer, W. F. (1992). Mental models of the earth: A study of conceptual change in childhood. *Cognitive psychology, 24*(4), 535-585.

Weick, K. E. (1995). *Sensemaking in organizations*. Thousand Oaks: Sage Publications.

Weick, K. E., Sutcliffe, K. M., & Obstfeld, D. (2005). Organizing and the process of sensemaking. *Organization Science, 16*(4), 409-421. doi:10.1287/orsc.1050.0133

Appendix A – Post-hoc Interview Questions

Participants were instructed to consider the task of detecting an error in the sensor report and shown an example.

1. What information did you have available to you when making these decisions?

2. What did you look for when you made this decision?

3. How did you know that what you were paying attention to was the correct information?

4. Did you do anything to confirm what you were paying attention to was correct?

5. Have you had any previous experience with this kind of task that helped you determine the correct response?

6. What specific parts of the training or experience was helpful when making these decisions?

7. What short cuts or strategies did you use when solving these problems?

Appendix B – Interview Response Examples

When asked "what did you look for when you made this decision" participant 9 responded "I was just trying to see if it correlated with the graph (therm) on the side. Also, if they didn't go together if high waves and soil being under 40."

When asked about thought process on specific examples participant 9 responded "I was using both report and therm (probability estimate) I would say the chance would be a little bit higher. Using therm to determine if there was an error in the report because I didn't understand when just looking at the report. I would say therm should be lower just because the soil being below 40 and waves being longer."

When asked "what short cuts or strategies did you use when solving these problems" participant 11 responded "my strategy was to look at how many arrows there were to low compare to the temp and see if I thought they lined up."

When asked about thought process on specific examples participant 11 responded "This one I'm looking at how many low arrows there are to high and I'm looking at the bar has changed. I feel like in this one since there are more low arrows, then high, the prediction is wrong. There's something wrong in the features because of what's shown on the temp."

When asked "What factors need to be considered before fixing the error participant 19 responded "whether or not all three of these were the same and then figure out if which one was incorrect then you look at the soil moisture to help fix the problem."

When asked about thought process on specific examples participant 19 responded “the soil moisture is there and the wave’s length and winds are the same direction. No error.”