DERIVING PILOTS' KNOWLEDGE STRUCTURES FOR WEATHER INFORMATION: AN EVALUATION OF ELICITATION TECHNIQUES

by

KIMBERLY R. RADDATZ

B.S., University of Nebraska at Kearney, 1997
M.S., Kansas State University, 2003

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Psychology
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2011

Abstract

Systems that support or require human interaction are generally easier to learn, use, and remember when their organization is consistent with the user's knowledge and experiences (Norman, 1983; Roske-Hofstrand & Paap, 1986). Thus, in order for interface designers to truly design for the user, they must first have a way of deriving a representation of what the user knows about the domain of interest. The current study evaluated three techniques for eliciting knowledge structures for how General Aviation pilots think about weather information. Weather was chosen because of its varying implications for pilots of different levels of experience. Two elicitation techniques (Relationship Judgment and Card Sort) asked pilots to explicitly consider the relationship between 15 weather-related information concepts. The third technique, Prime Recognition Task, used response times and priming to implicitly reflect the strength of relationship between concepts in semantic memory. Techniques were evaluated in terms of pilot performance, conceptual structure validity, and required resources for employment. Validity was assessed in terms of the extent to which each technique identified differences in organization of weather information among pilots of different experience levels. Multidimensional scaling was used to transform proximity data collected by each technique into conceptual structures representing the relationship between concepts.

Results indicated that Card Sort was the technique that most consistently tapped into knowledge structure affected by experience. Only conceptual structures based on Card Sort data were able to be used to both discriminate between pilots of different experience levels and accurately classify experienced pilots as "experienced". Additionally, Card Sort was the most efficient and effective technique to employ in terms of preparation time, time on task, flexibility, and face validity. The Card Sort provided opportunities for deliberation, revision, and visual feedback that allowed the pilots to engage in a deeper level of processing at which experience may play a stronger role. Relationship Judgment and Prime Recognition Task characteristics

(e.g., time pressure, independent judgments) may have motivated pilots to rely on a more shallow or text-based level of processing (i.e., general semantic meaning) that is less affected by experience. Implications for menu structure design and assessment are discussed.

DERIVING PILOTS' KNOWLEDGE STRUCTURES FOR WEATHER INFORMATION: AN EVALUATION OF ELICITATION TECHNIQUES

by

KIMBERLY R. RADDATZ

B.S., University of Nebraska, Kearney, 1997
M.S., Kansas State University, 2003

A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Psychology
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2011

Approved by:

Major Professor
Richard J. Harris

Abstract

Systems that support or require human interaction are generally easier to learn, use, and remember when their organization is consistent with the user's knowledge and experiences (Norman, 1983; Roske-Hofstrand & Paap, 1986). Thus, in order for interface designers to truly design for the user, they must first have a way of deriving a representation of what the user knows about the domain of interest. The current study evaluated three techniques for eliciting knowledge structures for how General Aviation pilots think about weather information. Weather was chosen because of its varying implications for pilots of different levels of experience. Two elicitation techniques (Relationship Judgment and Card Sort) asked pilots to explicitly consider the relationship between 15 weather-related information concepts. The third technique, Prime Recognition Task, used response times and priming to implicitly reflect the strength of relationship between concepts in semantic memory. Techniques were evaluated in terms of pilot performance, conceptual structure validity, and required resources for employment. Validity was assessed in terms of the extent to which each technique identified differences in organization of weather information among pilots of different experience levels. Multidimensional scaling was used to transform proximity data collected by each technique into conceptual structures representing the relationship between concepts.

Results indicated that Card Sort was the technique that most consistently tapped into knowledge structure affected by experience. Only conceptual structures based on Card Sort data were able to be used to both discriminate between pilots of different experience levels and accurately classify experienced pilots as "experienced". Additionally, Card Sort was the most efficient and effective technique to employ in terms of preparation time, time on task, flexibility, and face validity. The Card Sort provided opportunities for deliberation, revision, and visual feedback that allowed the pilots to engage in a deeper level of processing at which experience may play a stronger role. Relationship Judgment and Prime Recognition Task characteristics

(e.g., time pressure, independent judgments) may have motivated pilots to rely on a more shallow or text-based level of processing (i.e., general semantic meaning) that is less affected by experience. Implications for menu structure design and assessment are discussed.

Table of Contents

List of Figures

List of Tables

List of Acronyms

AFSS	Automated Flight Service Stations
ANOVA	Analysis of Variance
CS	Card Sort
D	Dimension
FAA	Federal Aviation Administration
FISDL	Flight Information Service Data Link
FMS	Flight Management Systems
HCI	Human-Computer Interaction
KETs	Knowledge Elicitation Techniques
KSU	Kansas State University
LTM	Long-Term Memory
METAR	Aviation Routine Weather Report (French translation)
MDS	Multidimensional Scaling
MFD	Multifunction Display
OASIS	Operational and Supportability Implementation System
PIREP	Pilot Report
PRT	Prime Recognition Task
RJ	Relationship Judgment
SIGMET	Significant Meteorological Conditions
SME	Subject Matter Expert
TAF	Terminal Aerodrome Forecast
WM	Working Memory
WMDS	Weighted Multidimensional Scaling

Acknowledgements

The completion of this dissertation would not have been possible without the help and support of many talented, insightful, and inspirational people. I am deeply grateful to all of the following people for their contributions, both intellectual and emotional, as each have left an indelible mark on this dissertation and my professional development.

First and foremost I owe a great debt of gratitude to my major professor, Dr. Richard J. Harris, for his willingness to supervise my dissertation under less than ideal circumstances. Not only were his insights and perspectives integral to shaping the focus of the research, but also without his unwavering patience I would have never been able to complete this dissertation while working full-time.

I also wish to thank my committee members, Dr. Kevin Jordan, Dr. Lester Loschky, Dr. James Shanteau, and Dr. Kyle Douglas-Mankin for their contributions to the research and for their continued belief and support in my ability to complete this research even as the timeline kept getting extended.

This dissertation evolved from several years of FAA-funded grant work (*Contract DTFA-02-02-R-03491).* I want to thank Pete Elgin, my fellow researcher on the FAA grant, for his insights and feedback in the planning stages of this study. I also want to extend a very sincere thank you to Colleen Donovan and Tom McCloy for their financial support and their help in directing the focus of the study. I am also extremely grateful for the personal and professional support they have continued to provide me even though our FAA contract has ended.

I want to also acknowledge the help of all of the pilots and FAA personnel who functioned as subject matter experts in every capacity for which they were asked to contribute. These outstanding individuals include Wes Ryan, Greg Shetterly, Lowell Foster, David Sizoo, and Troy Zwicke from the FAA and Bill Gross from the Kansas State University – Salina Aviation Program. A very special thanks goes to Brad Amstutz who was involved in this *entire*

project (from inception to conclusion) and functioned as my personal flight instructor. He made me believe that there were no stupid questions even though I am sure I came close at least once or twice!

Of course, this study could not have been completed without the 53 General Aviation pilots who volunteered a significant amount of their time to participate in this study. I am forever grateful to them and for the passion they showed for aviation.

I also sincerely thank Abby Werth for all of her help with coding and recoding data for this research as well as her friendship and emotional support in helping me finish. She exceeded expectations as a research assistant and I am grateful we had the chance to work together.

I am also deeply grateful for the help of my colleague and friend, Gary Rao. Gary volunteered several hours of his time to code card sort data and I lost count of all of the ways he helped “bend” technology to my needs – there was not a technology problem or challenge he could not conquer. But I am most grateful for his motivational pep-talks (i.e., “nagging”) that helped keep me on task to finish the dissertation, especially when it was difficult to see the light at the end of the tunnel.

I also want to acknowledge the support of my coworkers at Sprint, in particular my manager, Donnelle Weller and my director, Clyde Heppner, for their personal and professional support and for their willingness to allow me to work a flexible schedule. Without their support, I never would have had the necessary time to complete my degree.

I have been blessed with a phenomenal group of friends, especially Beth Cady, Jerry Deehan, and Stacie Hoesly, who have been there for me in any number of different circumstances. Their faith and support saw me through the difficult times when I thought the end would never be in sight. I am truly grateful.

Finally, I truly owe the completion of this study to Tuan Tran. His insights and his ability to think “outside-the-box” helped to shape the direction of the study especially in terms of data analysis and interpretation. It was his amazing mind for research (and his willingness to share it), his constant motivational pep talks, and his unwavering belief in me that ensured that failure was not going to be an option for me or this study. I am truly grateful.

Dedication

This dissertation is dedicated to my family and especially to my Grandma, Jeannette Yost, who will have the best seat in the house on Graduation Day. Her presence will be felt that day as it is every other day.

Chapter 1 - Statement of the Problem

Multifunction displays (MFD) provide a means through which pilots can access large amounts of information through a single interface, thereby attenuating space constraints inherent in the cockpit. However, because of relatively small MFD display real-estate, information may no longer be directly available through simple visual search, but instead may be hidden within layers of the MFD's information structure (i.e., hierarchical menu structure). Therefore, the MFD must be organized in such a way that the pilot will know where and how to find the desired information.

When interacting with any interface, novel or familiar, pilots bring with them a vast amount of aviation-related knowledge and previous experiences that allow them to form expectations about how information may be organized in that interface. Systems are easier to learn, use, and remember when they are organized in a manner consistent with the knowledge and experiences that the user brings to the situation (Roske-Hofstrand & Paap, 1986). Further, usability bottlenecks can occur when there are disconnects between the design of the system and the user's knowledge and expectations (Norman, 1983). Thus, in order for interface designers to truly design for the user, they must first have a way of deriving a representation of what the user knows about the domain of interest.

The term "knowledge structure" is often used to refer to the organization of one's knowledge about a certain domain. Knowledge structures are comprised of information concepts (e.g., *winds* and *jet stream* are two weather information concepts) and the relationships or associations between these information concepts (e.g., jet stream is a type of wind phenomenon) stored in Long-Term Memory (LTM). They consist of both *declarative knowledge* (e.g., facts, personal experiences, and characteristics associated with aircraft and flight) and *procedural knowledge* (e.g., how to use various systems to complete certain tasks). Declarative knowledge provides users with an in-depth understanding of system components

independent of specific tasks, whereas procedural knowledge represents properties of the system that allow the user to perform specific real-world tasks (Van der Veer & Melguizo, 2002). Some researchers further propose that knowledge structures mediate the translation of declarative knowledge into procedural knowledge and facilitate the application of procedural knowledge (Jonassen, Beissner, & Yacci, 1993).

The notion of a knowledge structure is predicated on the idea that information concepts are stored in memory and differ in relatedness or "psychological proximity." Learning and exposure to new experiences can affect the psychological proximity of information concepts, including the addition of new information concepts to the structure and the reorganization of some of the "old" knowledge, based on the new information. Therefore, individuals who vary in their expertise of a given domain will presumably also vary in how domain knowledge is structured in memory (Schvaneveldt, Durso, Goldsmith, Breen, & Cooke, 1985). Chapter 2 provides a more extensive definition of knowledge structure and review of relevant research.

The process of deriving a representation of how knowledge is structured in memory is generally three-fold. First, information concepts representative of the knowledge domain must be identified. Second, the psychological proximity between the concepts must be quantitatively measured. Third, proximity data must be transformed into a meaningful representation of the psychological proximity between the concepts. For the purposes of the current study, the term "knowledge structure" will refer to the "true" organization of information *in memory*. Chapters 7 and 8 review the processes through which fifteen representative weather-related information concepts were identified for use in this study. The derived representation of how knowledge is structured will be referred to as the "conceptual structure" because 1) it is based on a *representative sample* of information concepts assumed to comprise the knowledge structure and 2) its resemblance to the true knowledge structure is dependent, in large part, on the techniques used to elicit the psychological proximity of those concepts. Techniques that rely on

verbal report or intuition can be subject to bias and flaws in memory or intuition (Canas, Antoli & Quesada, 2001; Cooke, 1999). Techniques that infer knowledge structure from judgments of conceptual relatedness have had their relationship to performance questioned (e.g., Geiwitz et al, 1990, as cited in Cooke, 1999). Indeed, the relationship between knowledge structure and conceptual structure in many ways is analogous to the relationship between the population mean and the sample mean in statistics. As Figure 1.1 illustrates below, the conceptual structure represents the knowledge structure plus some "error" introduced via the elicitation techniques and scaling procedures, just as a sample mean represents the population mean plus measurement "error." Thus the process of knowledge elicitation is more conventionally viewed as a process of constructing a model of knowledge (i.e., conceptual structure) rather than a direct extraction of knowledge structure (e.g., LaFrance, 1992), with the resultant model reflecting reality to a varying degree (Cooke, 1999).

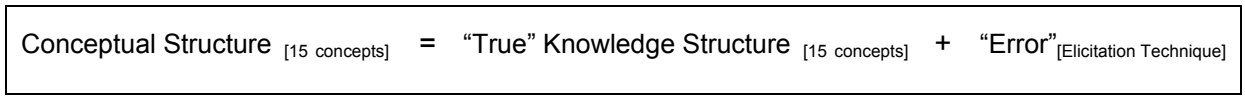


Figure 1.1. Relationship between the *derived* conceptual structure and the *true* knowledge structure.

In Cognitive Psychology literature, there exist several techniques that can be used to elicit or tap into knowledge structures and derive their corresponding conceptual structures. Most knowledge structure elicitation techniques are based on the assumption that knowledge structures are semantic networks in memory comprised of concepts and associations between the concepts that vary in strength. The strength of the association between two concepts is an indication of their similarity (e.g., Anderson & Bower, 1973). Some techniques (e.g., Relationship Judgment and Card Sort) explicitly ask participants to generate judgments of similarity. The explicit techniques are well known and frequently used but it is unclear whether

the high-order cognitive processing needed to generate similarity pairings/groupings affects the validity of the resultant conceptual structure. An alternative technique may be the use of response times and priming to implicitly reflect the strength of relationship between concepts in semantic memory. The Prime Recognition Task, borrowed from the associative memory literature, has been used to test hypotheses about underlying knowledge structures (e.g., Navarro-Prieto & Canas, 2001) but it has not been used to derive knowledge structures themselves. Therefore, the validity of the Prime Recognition Task as a knowledge elicitation technique has not been fully explored. Chapter 3 provides a review of relevant explicit and implicit knowledge elicitation techniques (KETs).

KETs differ in several key ways that may have an impact on the validity of the data collected (i.e., the extent to which the conceptual structure resembles the true knowledge structure). First, some KETs present stimuli for judgment simultaneously (e.g., Card Sort) while others present stimuli one at a time or in pairs (e.g., Relationship Judgment). Second, some KETs allow for judgments to be changed (e.g., Card Sort) or revised while others do not (e.g., Relationship Judgment). Third, some KETs may involve time pressure (explicit or implied) that may affect the extent of cognitive processing the participant engages in when making judgments. Fourth, the total number of similarity judgments may vary considerably across KETs, which in turn, may influence the extent to which fatigue, boredom, or vigilance affect the judgments. And lastly, as previously mentioned, many KETs explicitly ask for participants to make similarity ratings, but other types of objective data (e.g., response times) could be used to represent proximity data without the participant explicitly considering similarity as a factor (see Chapter 3 for more information on the potential advantages of implicit over explicit techniques with regard to eliciting knowledge structure). Understanding the consequences of the choice of KET is crucial to the use of conceptual structures to guide interface design. Conceptual structures can only be used to facilitate better interface design if they are able to adequately

reflect the true nature of the knowledge structure from which they were elicited. Some previous studies have examined the effect of KET on the validity of the resultant conceptual structures (e.g., Bijmolt & Wedel, 1995; Dorsey, Campbell, Foster, & Miles, 1999; Rowe, Cook, Hall, & Halgren, 1996). However, more research is needed as the list of KETs is extensive and their utility is often goal- or domain-specific (Hoffman, Shadbolt, Burton & Klein, 1995).

One means of validating conceptual structures and comparing them across KETs is to assess whether or not the conceptual structures can be used to 1) discriminate among participants of certain groups and to 2) predict group membership (Schvaneveldt et al., 1985). Of particular interest to the current study is whether conceptual structures can be used to discriminate and classify pilots in terms of their level of experience. In other words, given the conceptual structure of a “high-time” pilot (i.e., a pilot who has logged a large amount of flight hours), can the pilot be correctly classified as a High-Time pilot? Again, this validation technique is based on the reasonable assumption that pilots with the same level of experience also share certain characteristics in their conceptual structures.

Although validity is an important factor to consider when comparing KETs, it is also important to compare and contrast KETs on some of the more practical aspects of their employment, including the time and resource requirements of each KET. KETs that elicit highly valid conceptual structures may still have their usefulness questioned in the field of Human-Computer Interaction (HCI) if they are typically unable to be employed and analyzed with minimal time and effort required by both the participant and the researcher. Often, research in the more applied settings of HCI is characterized by tight timelines, minimal resources, and research personnel who may have limited formal background in Human Factors, Psychology and/or Statistics. For a KET to be truly valuable to interface design, it must be robust to the constraints posed by applied environments.

Previous research studies have used conceptual structures to suggest avionics menu structure for Flight Management Systems (FMS) (e.g., Roske-Hofstrand & Paap, 1986) and MFDs (e.g., Williams & Joseph, 1998) and to suggest prioritization of information presentation in cockpit information displays (e.g., Schvaneveldt, Beringer, Lamonica, Tucker, & Nance, 2000). However, with the exception of Williams and Joseph (1998), few other studies have included weather-related elements in their list of information concepts. When weather-related concepts were included, the concepts tended to be vague and used nondescript terminology (e.g., "general weather" and "wind"). A large part of the risk that is inherent in General Aviation (GA) stems from the effects of deteriorating weather conditions on flight dynamics. Weather information can be presented on current MFDs in much greater detail than was previously possible in the cockpit. Therefore, the primary area of interest for the current study is aviation weather and how that information is organized in pilots' knowledge structures.

In summary, the main goal of the current research study is to compare and contrast three different KETs for eliciting and representing pilots' knowledge structures for weather-related information concepts. These KETs will be compared and evaluated in terms of the characteristics of the data collected, the validity of the resultant conceptual structures, and the resources required for their employment. Special focus will be placed on how each KET is able to resolve any differences in conceptual structure for pilots of varying levels of experience.

Chapter 2 - Defining Knowledge Structure

Defining Terms

Knowledge structure is a hypothetical construct in that it has no structural referent in the brain. However, it is a very useful construct for describing the way humans organize and retrieve information from LTM. Knowledge structure as a construct has been applied in many domains, from basic research in cognitive psychology to more applied research in Human Factors, HCI, and even Cognitive Science. Each domain may have slight variations in how knowledge structures are conceptualized but most are consistent in their underlying premise of referring to the pattern or organization of concepts and their relationships in LTM (e.g., Preece, 1976; Shavelson, 1972). In addition to *knowledge structure*, other terms in the literature that have been used to refer to this construct include *cognitive structure*, *conceptual knowledge*, and *structural knowledge* (Jonassen et al., 1993).

Knowledge *structure* differs conceptually from knowledge *content*. Knowledge content refers to the amount and type of knowledge that is encoded in a given knowledge structure. Knowledge structure refers to the interrelationships of concepts that comprise that knowledge content (Ye, 1998). Performance on almost any task that requires cognitive effort will be affected by both the knowledge structure and the knowledge content participants have that is relevant for that task, as the knowledge structure (i.e., the set of connections) leads to an understanding of when and how knowledge content applies in a given situation (Baxter, Elder & Glaser, 1996). For example, anyone can develop the knowledge content for how to play chess (e.g., understanding of the rules, knowledge of the terminology). However, people with more experience playing chess may develop a better structure for their knowledge about chess, which may, in turn, yield better performance. For example, experienced chess players are able to relate current chess configurations to past experiences in developing the strategies necessary to decide which chess pieces to move next and/or within the next two or three steps. In other

words, the organization of knowledge stored in memory is of equal or even greater significance to task performance than the amount or type of knowledge (Kraiger, Ford & Salas, 1993). The primary focus of the current study is on knowledge *structure* although initial phases of the study are necessarily devoted to understanding the knowledge *content* pilots have for weather as well.

Theoretical Foundation of Knowledge Structures

Prevailing memory theories are based on the idea that information is represented in memory in an organized network of associations (e.g., Collins & Quillian, 1969). This network is comprised of concepts (nodes) that are "linked" to represent a variety of relationships. The degree of relatedness between two concepts can be represented by either the strength of a link or the number of shared links. The retrieval of information proceeds through spreading activation (Collins & Loftus, 1975). A given concept activates a corresponding starting node and that activation spreads through all the links connected to that node to other nodes and through all of those links to other nodes. These nodes can represent other concepts and/or their properties. The strength of activation decreases as time, distance, and the number of activated unrelated concepts increases. If the activation reaching a given node achieves a threshold value, that corresponding concept will be activated or retrieved from memory (Anderson, 1974).

Within semantic memory, the more nodes or properties that concepts have in common, the stronger they are linked together through those properties. Further, learning can be thought of as a reorganization of the network in semantic memory, as these semantic networks provide an indication of what a person knows and a framework for learning new ideas through adding, deleting, and restructuring associations. Thus, for the present study, semantic networks are analogous to knowledge structures (Jonassen et al., 1993).

Schemas and Knowledge Structures

The notion of knowledge structure is not unlike the construct of a schema. Schema theory (e.g., Rumelhart & Norman, 1985) asserts that knowledge, experiences, and expectations are organized in structures called schemas. Schemas can be thought of as information packets that represent knowledge about an object or an event in the form of attributes or variables whose values assist in object or event recognition. Schemas vary in their complexity (e.g., ice cream cone, sailing on a boat in the bay) and their abstractness (e.g., happiness, paying at a restaurant). They can be embedded with other schemas. For example, a schema for hitting a homerun may be comprised of both a schema based on the procedural knowledge needed to hit the ball and a schema based on the particular home run you saw Alex Gordon hit at the last Royals game you attended. Schemas are active, dynamic and continually changing – new ones can be developed based on existing ones, and existing ones can be altered or adjusted to meet new task or domain demands. Knowledge structure is built through the use of schemas and their interrelatedness (Jonassen et al., 1993).

Mental Models and Knowledge Structures

Research efforts in several domains have focused on how design can best support the user in developing the “right” mental model, including situation awareness (e.g., Jones & Endsley, 2000), HCI (e.g., Norman, 1988), system maintenance (e.g., Kieras & Bovair, 1984), computer programming (e.g., Pennington, 1987), and weather forecasting (e.g., Trafton, 2004). However, the term “mental model” has been used with much impunity within the literature and its use is often not clearly defined or operationalized (Rickheit & Sichelschmidt, 1999; Canas et al., 2001).

Gentner & Stevens (1983) define a mental model as an internal psychological representation of how a person conceives and understands the system to function. Carroll & Olsen (1988) further define a mental model to be a dynamic representation of the components

of a system, how the system works, the relationship between the components, what the internal processes are, and how those internal processes affect the components. Generally, the mental model construct has as its core components working memory (WM), knowledge structures (i.e., LTM), and perceptual processes (Gentner & Stevens, 1983; Johnson-Laird, 1983; Yates, 1985). To perform a given task, a person forms a mental representation of the world (e.g., the system) by combining the information stored in LTM (i.e., knowledge structures, schemas) with the information about the task characteristics that are extracted by the perceptual processes. Knowledge is extracted from LTM into WM based on a series of triggering events, for instance, pattern recognition (Canas et al., 2001). Therefore, for the purposes of the current study, the terms knowledge structure and mental model are not synonymous terms. Rather, knowledge structures are considered and treated as integral components of LTM upon which mental models are derived and employed in WM during task performance. The primary focus of the current study is on knowledge structure, rather than mental model.

Knowledge Structure and Experience

Efforts to understand the variables that differentiate skilled or expert performance from less-skilled or novice performance can be traced back to the late 1800s (Bryan & Harter, 1899, as cited in Beilock & Carr, 2004). Past research has not only focused on measuring the success of overt behavior but also on understanding the cognitive changes that occur as learning progresses and performance improves. Cognitive mechanisms that facilitate planning and execution are what are truly thought to distinguish novice from skilled performance. Memory and attention are two of these important cognitive mechanisms (Beilock & Carr, 2004). Fitts & Posner (1967), Anderson (1983), and Rasmussen (1983) have all proposed theories or frameworks of skill acquisition. While the nature of these theories range from descriptive frameworks (e.g., Fitts & Posner, 1967; Rasmussen, 1983) to simulated models (e.g., Anderson & Lebiere, 1998), all three are similar in that they propose skilled performance to be acquired in

three general stages. With each successive stage, knowledge is represented in an increasingly structured and organized way (Ye, 1997).

Stage I: Novice performance is based on declarative knowledge that is explicitly retrieved from LTM and held in WM where it is consciously attended to in real-time. Fitts & Posner (1967) refer to this stage as the *cognitive stage*, Anderson (1983) as the *declarative stage*, and Rasmussen (1983) as *knowledge-based behavior*.

Stage II: As learning progresses, declarative knowledge is restructured into procedural knowledge through experience and repetition. This stage is characterized by less conscious control of real-time performance as "rules" or "procedures" are developed or compiled that associate task characteristics or situations with the appropriate actions. This stage is referred to as the *associative stage* by Fitts & Posner (1967), the *knowledge compilation* by Anderson (1983), and *rule-based behavior* by Rasmussen (1983).

Stage III: With extended practice, conscious attentional control of real-time performance is no longer necessary as particular actions are automatically executed when confronted by particular task characteristics or situations. Procedural knowledge is reinforced and refined through experience and fine-tuned for efficiency. This stage is analogous to Fitts & Posner's (1967) *autonomous stage*, Anderson's (1983) *procedural stage*, and to Rasmussen's (1983) *skill-based behaviors*.

Previous empirical research studies have provided evidence to suggest that groups that differ in skill level also differ in knowledge structure. Ye (1998) provided an overview of several relevant studies and results. General findings from the review are summarized below:

- **Experts have more well-structured chunks of knowledge in LTM.** Several studies have found that experts exhibit better recall performance than novices but only when pieces of information are represented to participants in a meaningful organization. Experts and novices show similar recall performance when pieces of information are

presented in a random or scrambled organization (e.g., Adelson, 1981; Chase & Simon, 1973; de Groot, 1965; Reitman, 1976)

- **Experts frame problems in terms of abstract principles while novices represent problems in terms of surface or literal characteristics.** Chi, Feltovich & Glaser (1981) found that when asked to categorize physics problems, those with more physics experience categorized problems according to laws or principles of physics, whereas less experienced physicists categorized the same problems in terms of surface features or more literal aspects of the problem. These results imply that experts represented the problems according to deep and highly learned principles, whereas those with less experience represented the physics problems at a more surface level.
- **Experts form fewer but larger chunks of information.** Ye and Salvendy (1994) asked 10 expert and 10 novice C computer programmers to provide relatedness ratings for C programming concepts presented in pairs. Relatedness ratings were averaged across each of the experience groups and submitted to a hierarchical clustering analysis, with the clusters of concepts representing how each group "chunked" knowledge of the C language. Results showed that experts formed fewer but larger chunks than did novices.
- **Variance in knowledge representation decreases with increases in skill level.** Goldsmith & Johnson (1990) asked college juniors and seniors taking a Psychology class to provide relatedness ratings for psychology concepts presented in pairs at different times during the course (1st week, 8th week, 15th week). Relatedness ratings were analyzed using Multidimensional Scaling and Pathfinder Analysis. The coefficient of correlation between the instructor's knowledge structure and each student's knowledge structure increased over the course of the semester. Also, the agreement between the students' knowledge structures increased over the course of the semester as well.

Research studies in aviation provide evidence implying knowledge structure differences with varying levels of experience. Schvaneveldt et al. (1985) collected similarity ratings for pairs of basic flight-related concepts from 10 instructor pilots (*M*=2583 hrs), nine Air National Guard Pilots (*M*=6064 hrs) and 17 undergraduate pilot trainees (*M*=200 hrs). Similarity ratings were submitted to both MDS and Pathfinder Analysis. Analyses suggested that the knowledge structures of experienced pilots and novice pilots differed in their complexity, with novices showing more complex cognitive structures than the experienced pilots (i.e., more links between concepts). The results implied that experienced pilots are able to identify the important critical information and associations which, in turn, yields a simpler cognitive structure, with only the most meaningful associations represented.

One of the typical explanations suggested for these findings is that those with extensive domain experience are able to perceive a more global picture of the domain, which facilitates chunking relevant concepts into larger units compared to novices. Since research suggests that novices have less organized knowledge structures compared to experienced users, they are not capable of encoding information as quickly or in units as large as the experienced users can. Some have also suggested that those with domain experience have memory structures that are hierarchically organized and, therefore, can recall a larger number of chunks. In this case, a high-level chunk may be comprised of several low-level chunks, each further containing chunks at a more detailed (lower) level. This hierarchical structure allows them to store large amounts of information more effectively than novices (Schvanevaldt et al., 1985).

In sum, a knowledge structure is a hypothetical construct with its theoretical foundation in associative memory and it is used for describing the way humans organize and retrieve information from LTM. They, like schemas, are the integral components upon which mental models are derived and employed in WM. Empirical studies have suggested that the enhanced

performance typically seen by experienced participants relative to novices is in large part a function of the experienced participants having more organized knowledge structures.

Chapter 3 - Review of Knowledge Elicitation Techniques

Categories of Techniques

The process of knowledge elicitation was born out of a desire to build knowledge-based applications to facilitate human performance. These knowledge-based applications include expert systems, adaptive user-interfaces, and knowledge-based selection and training protocols (Cooke, 1999). Researchers and practitioners in a variety of disciplines, including human factors, judgment and decision-making, cognitive science, and expert systems, have developed a number of different knowledge elicitation techniques (KETs) designed to extract and preserve domain-specific knowledge underlying human performance, most commonly within the context of experienced-novice differences (Hoffman et al., 1995).

While the general goal of each method is to elicit knowledge, the purpose of elicitation may vary depending on the discipline. For example, from the perspective of naturalistic decision-making, the KETs must possess ecological validity and representativeness and be easily transported from the laboratory setting to the field setting. From an expert systems perspective, KETs must allow the elicitation of the ”important” knowledge directly since the major goal is to develop a high quality and valid knowledge base to be used in an eventual computer system. From the psychological perspective, KETs must reveal information about reasoning strategies and sequences and facts about how knowledge is organized (Hoffman et al., 1995). To that end, Hoffman et al. (1995) define three categories of knowledge elicitation techniques.

Observational Techniques

Observational KETs are designed to understand the tasks that experts perform (i.e., *What do experts usually do?*). Methods such as documentation analysis, task analysis, think aloud problem-solving, and protocol analysis provide insight into what experts do when they

conduct their usual problem-solving or decision making tasks in their natural environments. These observational techniques are typically how knowledge elicitation begins and are especially helpful in providing an overall conceptualization of the domain and some of the relevant constraints or issues to be mindful of in later phases of knowledge elicitation. However, there are some tasks in which observations in natural settings are impossible (e.g., flying single-seat aircraft) and therefore simulated contexts (e.g., simulator) must be used (Cooke, 1999). Observational techniques can vary in terms of what is observed (e.g., all events or pre-defined events), the observer's role (e.g., passive/nonintrusive or participatory), and documentation method (e.g., note-taking, video, photos, audio) (Cooke, 1999).

Interview-based Techniques

A second category of KETs is comprised of interview-based methods (i.e., *What do experts say they usually do?*). Interviews can either be structured or unstructured, individual or group. During the interviews, experts are asked a series of questions that cover a broad range of issues within a domain. Experts may also be asked to generate different types of lists that are related to the domain of interest (e.g., list of important concepts, definition of terms, list of procedures, event recall). Unstructured interviews are free-flowing and are especially useful in learning about a domain and gathering the knowledge necessary to set up structured interview questions or tasks. Structured interviews take on many different forms (Cooke, 1999). For example, in *forward scenario simulations*, the expert is walked through a problem verbally and asked to respond to a series of system and/or environmental events posed by the facilitator. In the *teachback method*, the expert is asked to teach the facilitator something and the facilitator explains it back to the expert for verification. The process continues until the expert is satisfied with the facilitator's explanation. In the *20 questions method*, the expert tries to guess a domain concept by asking the facilitator yes/no questions about the concept. The types of questions asked by the expert reveal information about what attributes are important for distinction within a

domain. While interviews in general are easy to administer compared to other KETs, untrained facilitators can bias the integrity of the data collection by the type of question asked (or not asked) and how the question is framed (e.g., leading questions). Also, analysis of interview data can be time-consuming and tedious.

Indirect Techniques

Indirect[1] KETs are designed to reveal knowledge and reasoning processes *indirectly*, without actually asking about these processes (i.e., *"What do experts do when they are constrained in some way?"*). Methods like decision analysis (e.g., risk analysis, analyses involving probability and utility modeling), group decision making, rating tasks and sorting tasks force the expert to perform completely unfamiliar tasks or familiar tasks that have been modified or the experts themselves are constrained in some way in performing them. These indirect tasks have been shown to reveal experts' knowledge and reasoning. For example, through asking expert and novice chess players to recall game boards in which pieces had been randomly arranged, Chase & Simon (1973) discovered that experts were no better than novices at recalling chess positions when those positions did not follow the conventions of the game. Indirect tasks like card sorts and rating tasks have been used to derive conceptual maps or structures of a domain (Hoffman et al., 1995).

Two factors contributed to the decision to employ only indirect techniques for the current study. First, since the focus of the current study was to understand how knowledge is organized in memory and not necessarily to understand how experts use that knowledge to perform tasks, only indirect techniques were employed. Second, one of the benefits of the indirect techniques is that the data collection is less susceptible to bias or influence from the researcher/facilitator. Conversely, observational and interview-based techniques require extremely well-trained

[1] Hoffman et al. (1995) refers to these techniques as "contrived" techniques, but for the purposes of the current study, they will be referred to as "Indirect" techniques.

facilitators/researchers to conduct the data collection because the data collected can be very easily biased or influenced by even the most unintentional cues given by the facilitator during the session (e.g., framing/terms used in the questions, facial expressions, unspoken body language). Further, the onus is on the facilitator/researcher to identify and interpret what information is important and unimportant for the study and their interpretation could be biased by any preconceived notions of what they might find (i.e., untrained researchers/facilitators may be more likely to focus on or hear only what they expect to hear during data collection). Thus, indirect techniques for KETs may result in more reliable insights into knowledge structures. For more information on the other two categories of KETS, the reader is directed to Hoffman et al. (1995) and Cooke (1999).

Cooke (1999) refers to Indirect techniques as "conceptual methods of knowledge elicitation" when they are used to elicit and represent knowledge structure as domain-specific concepts and their interrelations. As previously noted, these KETs are based on documenting the perceived similarity between pairs of concepts and using similarity data to infer how concepts are organized in the knowledge structure. Indirect KETs can be further sub-divided based on the extent to which the conscious processing of perceived similarity is required. *Explicit techniques* require the participant to consciously compare or evaluate the similarity of items and/or categorize them based on shared properties. Relationship Judgment (e.g., Cooke, 1994) and Card Sort (e.g., Spencer, 2009) are examples of explicit KETs.

For the *Relationship Judgment Task*, participants are asked to rate (typically on a 9-point scale) pairs of representative information concepts in terms of their perceived similarity. In theory, relationship judgments provide the most direct method for rating the similarity between concepts in a participant's knowledge structure. However, several properties of the Relationship Judgment Task have the potential to limit the reliability of the data that are elicited. First, note that if the term "similarity" is not operationally defined, participants are allowed to vary in their

interpretation of the “relationship” used render the comparisons, possibly compromising the informativeness of the resulting conceptual structure. For instance, if asked to rate the similarity of a *plate* and a *fork*, a participant who interprets similarity in terms of shared functions (e.g., *both are used to support eating behavior*) would rate the pair quite high in similarity, while a participant who interprets similarity in terms of shared characteristics may rate the pair quite low in similarity (Roske-Hofstrand & Paap, 1990).However, operationally defining “similarity” in a way that is ultimately not meaningful to the participant or to the domain of interest also may compromise or limit the informativeness of the resulting conceptual structure. Also, similarity ratings have been shown to be limited in reliability, changing with direction of how the pairs are presented (chair-table vs. table-chair) and/or context in which they are collected (Jonassen et al., 1993). Lastly, without an intervening interference task between ratings, the rating attributed to one pair of information concepts may be influenced by the rating attributed to a recent pair of information concepts (Canas et al., 2001).

Another explicit method for indirectly deriving knowledge structures is to ask users to sort information concepts into different conceptual groups based on their perceived shared properties. The *Card Sort Task* is used to estimate semantic distances between categories and the concepts within those categories (Halgren & Cooke, 1993; Spencer, 2009). According to Miller (1969), asking participants to sort concepts according to some similarity of meaning or meaningful criteria results in the identification of concept groups that are assumed to be organized in a hierarchical manner. The groupings provide insight into the meaning that the participant assigns to the concepts and ultimately gives insight into the organization of their knowledge structure. Further, card sorting allows the identification of meaningful criteria upon which sorting is based, such as identifying concepts that are grouped in terms of their functional (i.e., “how to”) usage and concepts that are grouped in terms of their conceptual (“what is”) similarity. In some cases, the organization imposed by the participant during card sorting will

also provide identification of knowledge deficits that may prevent that participant's knowledge structure from facilitating the interaction with a system (Jonassen et al., 1993). The main disadvantages to the Card Sort task are that it restricts the participants to consider only the concepts presented on the cards and that it can be biased by the surface similarity of the terminology used to describe the concepts on the card or by the recent interaction with a familiar system.

In sum, the criticisms associated with explicit KETs can be attributed to the effects that cognitive processes have on the participants' relationship judgments or groupings, specifically those involved in perceiving and evaluating similarity. Further, according to Canas et al. (2001), explicit knowledge elicitation techniques only provide a partial picture of the knowledge structure – a picture that is dependent on the particular task performed by the participants.

Conversely, *implicit* techniques are those that derive knowledge structure through semantic priming of knowledge stored in LTM, thereby negating the need for explicit, effortful cognitive processes to elicit similarity data.[2] Specifically, Navarro-Prieto and Canas (2001) suggest employing a *Prime Recognition Task* in conjunction with explicit KETs in order to study how knowledge is stored and used in a particular domain. For the Prime Recognition Task, participants are shown a set of information concepts to store in memory for a short period of time (i.e., a memory set). Then, in each trial, the participants are presented with the target information concept and their task is to decide as quickly as possible whether or not that target was part of the memory set. The target is preceded by another information item (the prime). The underlying assumption is that if the prime and target are related in the knowledge structure, the activation of the prime will facilitate the activation of the target. In other words, through spreading activation, the activation of one concept (the prime) will result in the activation of highly similar concepts (e.g., a semantically related target).

[2] Note: the use of the terms "explicit" and "implicit" to describe these techniques is the author's decision. Thus the use of these terms for their meaning in this study may not be reflective of their use in the greater cognitive psychology literature.

In past knowledge elicitation research, the Prime Recognition Task has mainly been used to confirm hypotheses about the general organization of knowledge in LTM (Pennington, 1987; Navarro-Prieto & Canas, 2001). The present study, however, explored the use the Prime Recognition Task as an implicit technique to *derive* knowledge structure. Thus, major contributions of the current study will include: 1) the examination of the Prime Recognition Task as a KET, and 2) the identification of any practical differences that exist between conceptual structures when they are derived through explicit and implicit techniques.

Chapter 4 - Analytical Techniques for Knowledge Structure Representation & Evaluation

To review, the process of deriving a representation of how knowledge is structured in memory is generally three-fold. First, information concepts representative of the knowledge domain must be identified. Second, the psychological proximity between the concepts must be quantifiably measured and Chapter 3 reviewed some of the KETs that have been used to measure psychological proximity between concepts in memory. Third, proximity data must be transformed into a meaningful representation of the psychological proximity between the concepts. The first section of this chapter provides an overview of several different scaling techniques that can be used to provide meaningful representations of how those concepts are related given their psychological proximities. The last section of this chapter describes some approaches for evaluating those derived knowledge structures.

Knowledge Structure Representation Techniques

Multidimensional Scaling (MDS), Pathfinder Network Scaling, and Cluster Analysis are three of the most commonly used analysis techniques to reduce the set of psychological proximity data into a graphical form that is easier to visualize, facilitating both qualitative and quantitative interpretation of the resulting conceptual structures.

Multidimensional Scaling

Multidimensional scaling (MDS) refers to a series of data analysis methods that visually represent the underlying structure and relation between objects or events for which data have been collected (e.g., Young & Hamer, 1987). The data consist of proximities placed into matrix format that quantify the strength or degree of relation (e.g., similarity) between objects or events represented in the matrix (Kruskal & Wish, 1978). MDS methods vary in terms of a number of factors including the space used to represent the data structure (e.g., Euclidean, non-Euclidean)

and whether or not they take into account individual differences. However, the one defining element of *all* MDS methods is the spatial representation of data structure along dimensions thought to represent features or attributes that differentiate the concepts (Young & Hamer, 1987). Figure 4.1 provides an illustration of a spatial layout of a two-dimensional MDS solution for proximity data gathered about a list of concepts occurring in nature. In this example, MDS analysis revealed that the features or attributes that define these concepts are 1) whether they are living or non-living and 2) whether they are plants or animals.

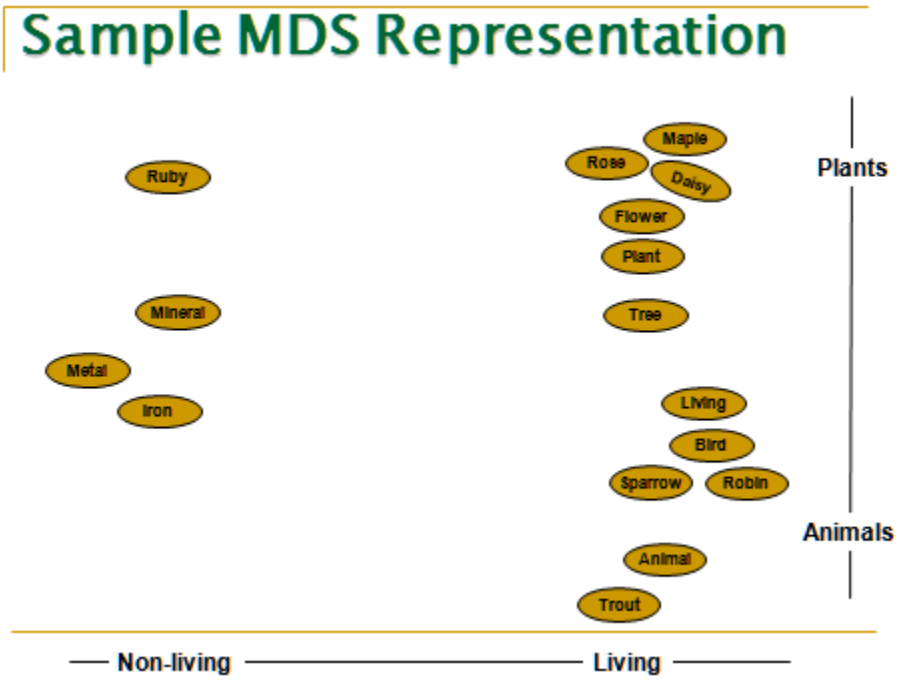


Figure 4.1. Illustration of the spatial layout of an MDS solution in two-dimensions (adapted from Schvaneveldt et al., 1995).

Torgerson (1952) is credited with introducing MDS to the field of Psychometrics, defining the MDS problem space, and providing the first metric solution. His main motivation was to improve upon what he saw as the limits of traditional psychophysical scaling methods,

specifically that traditional methods presuppose knowledge of the area being investigated. In his own words (1952):

> [Traditional psychophysical scaling methods]…require judgments along a particular defined dimension, i.e., A is brighter, twice as loud, more conservative, or heavier than B. The observer, of course, must know what the experimenter means by brightness, loudness, etc. In many stimulus domains, however, the dimensions themselves, or even the number of relevant dimensions, are not known. What might appear intuitively to be a single dimension may not be necessary…it may be that they can be accounted for by linear combinations of others. Other dimensions of importance may be completely overlooked. In such areas the traditional approach is inadequate. (p. 401)

Thus, MDS differs from traditional psychological scaling methods in two important ways (Torgerson, 1952). First, MDS does not require judgments along a given dimension. Rather, MDS utilizes judgments of similarity (or dissimilarity) between the stimuli. Second, the dimensionality and the scale values of the stimuli are determined from the data themselves, rather than having to be predefined.

Pathfinder Network Scaling

Pathfinder Network Scaling is a structural modeling technique developed to derive and represent knowledge structure through the production of link-weighted networks in which concepts are depicted as nodes and relationships are depicted by links connecting the nodes (Schvaneveldt, 1990). Based on any set of proximity data (e.g., similarity ratings, relationship judgments, card sort matrices, etc.), a link is created between each pair of concepts or nodes. Each link is assigned a value or weight that reflects the strength of the relationship between the nodes. Pathfinder removes links on the basis of relative efficiency. That is, a direct link between two nodes is removed if a multi-link pathway exists through the network that is shorter than the

direct link. Thus, a link remains in the network if and only if it represents the minimum-length path between two concepts. The functions by which path length is computed will yield different networks. For instance, the number of links in the resultant network will decrease systematically with decreases in the computed lengths of multi-link paths in the network (Schvaneveldt et al., 1985). Examples of different path length methods include the Minkowski r-metric and the parallel method. Figure 4.2 provides an illustration of a network representation of proximity data gathered about a list of concepts occurring in nature series of nature concepts, the same concepts used in the illustration in Figure 4.1 above.

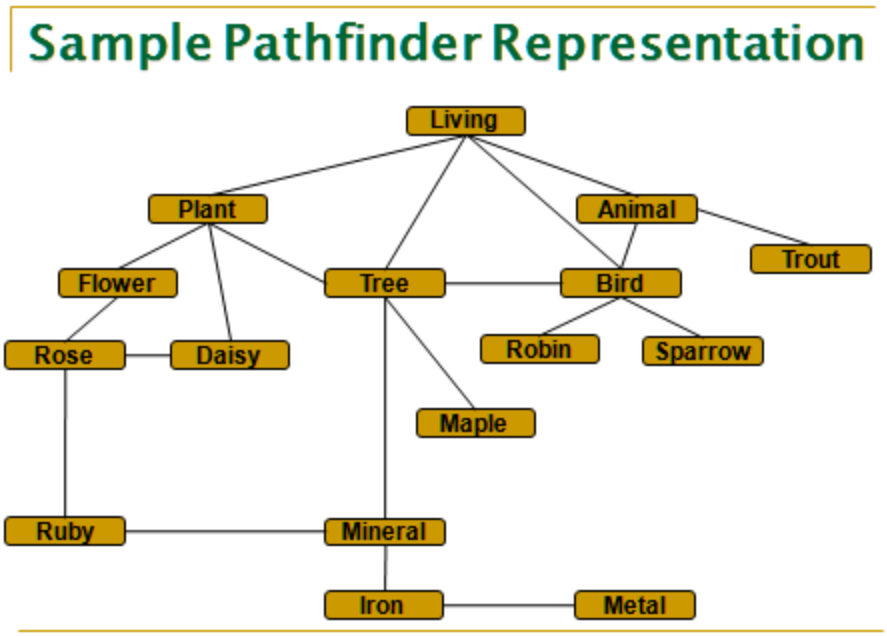


Figure 4.2. Illustration of a Pathfinder network representation (adapted from Schvaneveldt et al., 1995).

Cluster Analysis

Another common technique for mapping participants' knowledge structures is *cluster analysis* (e.g., Richard, Kleiss, & Bittner, 2004). Cluster analysis is an exploratory data analysis technique that can be used to sort different concepts into groups based on the degree of association (i.e., relatedness) between them. Two objects belong to the same group if they have maximum association and belong to different groups if they have minimal association. For instance, Lightning and NEXRAD radar returns are types of weather information and may be perceived to be highly related and would therefore end up being clustered together within the same group. Conversely, traffic awareness may not be perceived to be as related to NEXRAD information and therefore they would probably not end up being clustered together into the same group. There are several cluster analysis methods (tree clustering, block clustering, k-Means clustering) that differ in the procedures for defining clusters.

The Use of MDS as the Representation Technique

Comparing MDS to Cluster Analysis

Like MDS, Cluster Analysis uses proximities or distances between cases or variables. However, there are several deficiencies of Cluster Analysis when compared to MDS. First, in Cluster Analysis, it is typical to have a qualitative grouping of items into clusters. However, Cluster Analysis provides no real explanation about why that structure exists. MDS analysis results in a graphic representation that allows both visual and quantitative examination of the distances between each concept, with the nature of the dimensions providing some explanation as to why the structure exists in that form.

Also, MDS can be performed on multiple matrices, while Cluster Analysis can handle only one matrix. Therefore, in order to perform a Cluster Analysis using data from multiple participants, data must be aggregated into a mean matrix. This mean matrix attenuates the

variance among participants within a group. Consequently, the resultant group hierarchy gives an incomplete picture of the conceptual structures of the group (Ye, 1998). Not only does MDS provide the opportunity for individual differences scaling, it also provides a quantitative indication of how well each individual participant fits the overall model representation. Lastly, there is no way to identify whether or not differences in clustering across groups is statistically significant (George & Mallery, 2009).

Past research has shown that menu structures based on Cluster Analysis derived from end-user conceptual structures support improved performance and facilitation of menu structure learning (Lewis, 1991; McDonald, Stone, Liebelt, & Karat, 1982). However, for more complex systems, the advantages of cluster analysis may be diminished. Cluster Analysis forces an individual function or information element into a single cluster only, even though it may also be closely associated to functions or items in other clusters. This potential masking of other possible relationships between functions and elements may result in an oversimplified menu design, devoid of interconnections and redundancies that are increasingly common and important in complex MFDs (Richard et al., 2004).

In sum, MDS was chosen over Cluster Analysis as the representation technique because MDS 1) provides some insight into the structure (i.e., identifies underlying dimensions), 2) accounts for individual differences in the model and provides a metric for how well each individual's data fits the model, and 3) reduces the probability of oversimplified structures because it does not force items to fit into just one group.

Comparing MDS to Pathfinder

Both MDS and Pathfinder use estimates of psychological proximity as data with the end goal of reducing this large amount of proximity data into a meaningful and interpretable form. However, the two techniques approach achievement of that goal from different perspectives. MDS is concerned with identifying the semantic dimensions that underlie a set of concepts and

representing how those concepts cluster on those dimensions. The dimensions themselves provide an indication of how the concepts are organized in memory, with the Euclidean distance between the points representing the psychological distance between the concepts. Pathfinder judges the importance of the relationship between items in each pair of concepts and builds a network representation based on those individual relationships. Links between two concepts are only included in a network if that link represents the minimum length path between those concepts. Because MDS uses a least-squares technique to determine the location of *all* concepts in some k-dimensional space, each rating between all pairs of concepts is treated the same and has the same amount of influence on the spatial solution. This, in turn, could result in some distortion in the representation of strong associations as the solution has been made to fit all ratings data (Branaghan, 1990).

Several previous research studies have compared Pathfinder and MDS as techniques for representing knowledge structures (e.g., Branaghan, 1990; Cooke, Durso, & Schvaneveldt, 1986; Goldsmith, Johnson, & Acton, 1991; Gonzalvo, Canas, & Bajo, 1994; Schvaneveldt et al., 1985). In general, most research studies have concluded that both MDS and Pathfinder approaches provide valid techniques for assessing and representing knowledge structures (although see Arabie, 1993 for a critique of Pathfinder). However, previous research has suggested that because of their differing perspectives on how to represent knowledge structure (i.e., MDS focuses on underlying dimensions while Pathfinder focuses on network structure), the techniques provide different structural aspects of the knowledge domain. Specifically, MDS seems to capture more global information about knowledge structure and Pathfinder networks seem to represent more local relationships between concepts in knowledge structure (Gonzalvo, Canas & Bajo, 1994). For example, in tasks that require a global analysis of concepts (e.g., categorization, analogy completion), MDS representations are predictive of performance (e.g., Rips, Shoben, & Smith, 1973; Rumelhart & Abrahamson, 1973). For tasks

that rely on relationships between individual concepts such as free recall and paired associates, Pathfinder representations are predictive of performance (e.g., Cooke et al., 1986; Branaghan, 1990).

In order to compare and contrast the KETs, it was necessary to control for as many of the potential confounding variables as possible. One necessary control was to limit the number of weather-related concepts upon which to assess knowledge structure to 15 (Chapter 9 provides more information about the necessity of this constraint). These 15 concepts, while still representative of pilots' knowledge structure for weather, did not constitute a comprehensive representation by any means. Because the concepts were limited to 15, it was decided to focus primarily on assessing the global relationship between these concepts rather than the local relationships because it was assumed that any insight gained about local relationships would be largely dependent upon the concepts that were actually included in the analysis. For example, if two additional concepts were added to the study and the study was replicated, the local relationships between the concepts would be more likely affected by the inclusion of these two extra concepts than would the global relationships between the concepts. Thus it was feared that the information gained from an analysis of the local relationships between this limited number of concepts would be less indicative of pilots' actual knowledge structure for weather information than would the global relationships between these limited number of concepts. Since MDS has been shown to provide more quality information about global relationships between concepts relative to Pathfinder Network Scaling (e.g., Schvaneveldt et al., 1985), MDS was chosen for use in this study.

Approaches to KET Evaluation

Most approaches to evaluating conceptual structures share the common theme of evaluating one conceptual structure within the context of some standard. One approach is to quantify the similarity between the empirically derived conceptual structure and the actual or

known structure of the domain content (e.g., Shavelson, 1972). However, often the actual structure of the domain content is unknown. Therefore, a second approach is to compare different types of conceptual structures (Goldsmith et al., 1991). For example, conceptual structures of novices have often been compared to conceptual structures of experts. Also, conceptual structures of the same individuals have been compared before and after training. Correlations (e.g., Goldsmith et al., 1991; Gonzalvo et al., 1994) and the Mantel Test (e.g., Valero & Sanmartin, 1999) are two techniques that have been used by previous studies to compare conceptual structures. Both techniques will be employed in the current study.

Correlations

A correlation statistic determines a coefficient of correlation between two matrices of proximity data. One can compare intergroup correlations with intragroup correlations, which provides an understanding of the relative closeness of the conceptual structures for individual participants who may share a common grouping characteristic (e.g., experts or novices). However, correlations provide no indication about the nature of the differences in the conceptual structures between the groups of participants (Ye, 1998).

Mantel Test

The Mantel Test (Legandre & Legandre, 1998; Mantel 1967; Sokal & Rohlf, 1995) uses sampled randomization techniques to test whether the association or relationship between two matrices is stronger than what would be expected by chance. Unlike conventional statistical analyses, the Mantel Test does not assume independence of samples. Another advantage of the Mantel test is that it can be applied to different types of data (e.g., categorical, rank, interval-scale) as long as those data can be transposed into a distance measure (e.g., dissimilarity matrix). See Chapter 10 for more explanation of the Mantel Test.

Chapter 5 - Application of Knowledge Structures to HCI

The field of human factors has been at once a *benefactor* and *beneficiary* of knowledge elicitation research. Human factors and basic psychological research have contributed significantly to the development of knowledge elicitation methodology. Methods like structured interviews, card sort, and semantic scaling with their foundation in human memory and perception have made the process of knowledge elicitation more efficient than the unstructured interviews that were relied upon by the computer scientists who showed initial interest in the application of knowledge elicitation (Hoffman, 1998). Previous chapters of this document have provided an overview of the current methods or techniques for knowledge elicitation (Chapter 3) and representation (Chapter 4) of which Human Factors is partially responsible for creating. However, the fields of Human Factors and HCI have also benefitted from the application of these KETs, especially within the context of interface design and training. This chapter provides a select review of some of those most relevant contributions.

Contributions to Interface Design

The application of the knowledge structure construct to HCI most notably provides the theoretical foundation for the “mental model hypothesis” in system display design (e.g., Norman, 1988). The “mental model hypothesis” in HCI is based on the assumption that faulty or incomplete representations (or misconceptions) of knowledge structures in system design lead to user errors. The types of errors that users make can be understood (and may ultimately be eliminated) by deriving a model of their knowledge structures (Kellogg & Breen, 1990). In other words, when learning to use a new system, exposure to new technology requires the user to add new knowledge to their existing knowledge structures and modify the old knowledge structures in order to incorporate the newly learned information. Thus, new technology should be easier to learn, use and remember (and require less training time) if its content and

organization is consistent with the content and organization of the user's existing knowledge structure. Further, if system designs adhere to the *principle of consistency* (Wickens, Gordon, & Liu, 1998), users will be able to correctly infer how to complete a novel task based, in part, on the knowledge acquired through previous experiences (Roske-Hofstrand & Paap, 1986). Thus, not only will an interface that capitalizes on and incorporates users' existing knowledge structures support quicker learning, but it should also facilitate remembering and elicit a more efficient interaction with the system.

Since most of the interaction between a user and a system takes place primarily through menu selections, it is imperative that the menu structure and terminology be consistent with the user's knowledge structure. Previous research on menu layout supports the importance of interface organization. Several studies have found that users are able to locate items more quickly in menus that were categorically organized, compared to alphabetical or random (e.g., McDonald, Stone, & Liebelt, 1983; Salmeron, Canas, & Fajardo, 2005; Halgren & Cooke, 1993). Researchers reasoned that the benefit to performance was because the categorical arrangement of the menus had a high correspondence with the structure of the user's knowledge about the domain. However, Parkinson, Sisson, and Snowberry (1985) caution against blanket generalization of these findings to all forms of menu selection or implementation, as they found that categorization does not always show superior performance to alphabetized lists, especially when other design and situational factors are considered (e.g., spacing between groups, row vs. column arrangement, familiarity of terms, type of menu selection task).

A few attempts have been made to derive users' knowledge structures for the specific purpose of improving the menu organization (e.g., Williams & Joseph, 1998; Schvaneveldt et al., 2000; Jonsson & Ricks, 1995; McDonald, Dearholt, Paap, & Schvaneveldt, 1986). One seminal study that examined the application of knowledge structure as a guide to menu

organization in an aviation-related context was Roske-Hofstrand and Paap (1986). They created 34 panels, each containing a “chunk” of information representative of the types of information a Flight Management System (FMS) would display. Four pilots with varying levels of flight experience rated the similarity of each pair of panels using a 9-point scale with smaller numbers indicating less similarity. The ratings were submitted to a Pathfinder analysis which yielded meaningful information about local relationships between information elements, but the Pathfinder analysis had to be supplemented with Graph Theory in order to identify the items that would appear on a “main menu” page (highest level of organization) as Pathfinder is less useful for providing meaningful information about global organization. Thus, based on the Pathfinder and Graph Theory results, three menu structure prototypes were created based on users’ conceptual structure. Each prototype differed in the level of redundancy of the links between information concepts. A fourth prototype was created based on the recommendations of a design team rather than on derived user knowledge. Results showed that a menu structure based on conceptual structure that offers the maximum number of meaningful pathways from the top level to the bottom level (i.e., high redundancy) was easiest to learn and use.

Williams and Joseph (1998) performed a similar study with respect to the organizational design of multifunction displays (MFDs). They had 148 pilots of varying experience levels perform a card sort task of 48 flight related data elements that could be transmitted to the cockpit via data-link. Pilots also rated their familiarity with each item and the importance of each item with respect to performing three flight-related functions (communication, navigation, or surveillance). Pathfinder analysis was performed on the similarity matrices resulting from the card sort task. Their results suggested a foundation upon which to base MFD menu structure designs, however, no menu structure prototypes were constructed and validated as part of their study.

Contributions to Training and Instruction

As previously discussed, people with different levels of experience with a particular domain and/or tasks within that domain may develop different representations of that domain knowledge which may then lead to different levels of performance on a task (Ye, 1998). The correlation between experience and enhanced performance is most often attributed to experienced users having better organized knowledge representations that facilitate better performance. Therefore, investigating the differences in knowledge structure between groups of different skill levels (e.g., experts and novices) can lead to insights into how training and support tools may be designed to help facilitate enhanced performance by novices. The previous section reviewed some of the literature regarding how knowledge structures have been applied to interface design. This section will review a few examples of how derived knowledge structures have been used to facilitate training or instruction.

Knowledge structure assessment has been used to impact the content and organization of training courses. For example, knowledge structures can be assessed before any training occurs to identify how best to structure the training so that it capitalizes on existing knowledge (Whitener & Brodt, 1994; Dorsey et al., 1999). Also, knowledge structures can be used to identify which concepts are not well understood by novices compared to those who are more experienced in the domain. Note that since experienced users have developed a large database of knowledge, it is important that the knowledge essential to domain expertise be distinguished from knowledge that is not. The scaling-based KETs (e.g., card sort, similarity ratings) may be helpful in identifying which conceptual relationships are present in multiple experts' knowledge structures, suggesting high importance of those relationships to domain expertise (Schvaneveldt et al., 1985).

Knowledge structure assessment has also been used to assess the consequences or impact of specific types of training. For example, Kraiger, Ford & Salas (1993) proposed a

three-dimensional approach to training evaluation, with cognitive, skill-based, and affective components. Achievement tests, the most typical assessment of cognitive learning (i.e., knowledge acquisition), were questioned as to their ability to discriminate high levels of cognitive development. Therefore, the approach of Kraiger and colleagues relied on the assessment of knowledge structure, specifically the comparison of knowledge structures before and after training, to assess the impact of training on knowledge acquisition.

Selecting the Appropriate KET

As reviewed in Chapter 3, there are a multitude of KETs that have been developed with the intent to provide insight into knowledge structure organization. However, no one method has achieved universal acceptance (Rowe et al., 1996). The KETs vary in terms of several key characteristics, including 1) level of ecological validity of the situation, 2) level of involvement of the researcher, and 3) the level of behavior/introspection required by the user. These characteristics may affect the type of knowledge assessed (e.g., procedural vs. declarative) as well as the reliability and validity of that assessment. Consequently, KET acceptability is dependent on the goals of knowledge elicitation, with some KETs more equipped to support some goals than others.

Some research studies have been devoted to the evaluation of some KETs in terms of their reliability and validity. However, more research is needed. For example, Rowe and colleagues (1996) compared four KETs (laddering interview, diagramming task, think aloud task, and similarity ratings) and found that the laddering interview and the similarity ratings were the only techniques that were predictive of troubleshooting performance. Dorsey and colleagues (1999) compared concept mapping and similarity judgments with traditional measures of declarative knowledge. They found a low level of convergence between concept mapping and relatedness judgments, suggesting that the two KETs were tapping into different types of knowledge. Further, they found only small correlations between concept maps and

similarity ratings with traditional measures of declarative knowledge, suggesting that the KETs are also assessing a construct different from declarative knowledge.

Fiore, Fowlkes, Martin-Milham & Oser (2000) were interested in understanding whether card sort and similarity ratings would measure different aspects of “expert” knowledge. They found a significant correlation between the raw data gathered from the two KETs. However, when looking at the intersubject correlations between high and low-time pilots (based on total flight hours), card sort correlations were able to reliably distinguish among the different groups but the similarity rating correlations were not. However, this study used a relatively small number of participants (*N*=14) and the average total flight hours that distinguished “more experienced” pilots (*M*=2142 hrs) from “less experienced” pilots (*M*=1540 hrs) was relatively small. Thus, this study provides some interesting insights into the understanding of KETs and their interaction with experience level but more research is needed.

Chapter 6 - Research Objectives & Overview of Methods

Pilots learn about and use weather information in virtually every phase of flight. Previous research has found that flight experience plays a role in weather-related decision-making, especially in terms of being able to recognize and diagnose in-flight weather problems (e.g., Burian, Orasanu, & Hitt, 2000; O'Hare, Owens & Wiegmann, 2001). However, research has also shown that even experienced pilots lack some confidence in their abilities to diagnose problems related to weather (Goh & Wiegmann, 2002). One key component of the general decision-making process is the ability to diagnose a problem and evaluate pieces of information pertaining to that problem. According to the information processing model of decision making proposed by Wickens and Hollands (2000), LTM and WM influence this diagnostic stage of the process. Therefore, understanding how weather-related information is structured in the memory of both novice and experienced pilots could provide insight into how decision aids (e.g., MFDs) and training materials should be designed to facilitate weather-related decision-making and problem-solving.

As reviewed in Chapter 3, several different techniques have been used to elicit knowledge structures. The choice of technique may influence how well the conceptual structure represents the knowledge structure (e.g., Bijomolt & Wedel, 1995). Further, some techniques may be more adept than others at resolving differences in knowledge structure as a function of domain experience (e.g., Fiore et al., 2000). While previous research provides insights into differences between KETs, there is still more to be learned. Thus, the current study seeks to expand the understanding of how the choice of technique affects the validity of the resultant conceptual structures for pilots of varying levels of experience.

Some techniques (e.g., Relationship Judgment, Card Sort) require the pilot to engage in higher-order cognitive processing (e.g., generate scenarios where both concepts exist) to generate similarity data, which may, in turn, jeopardize the reliability and/or validity of the

judgments (Reitman & Rueter, 1980). The Prime Recognition Task implicitly derives a pilot's knowledge structure from response times generated from priming pairs of concepts. The task, borrowed from the associative memory literature, has been used to confirm hypotheses about underlying knowledge structures (e.g., Navarro-Prieto & Canas, 2001) but its use in the current study as a knowledge elicitation technique is novel and unexplored.

Thus, the *primary goal* of the current study was to compare and contrast three different KETs (Relationship Judgment, Card Sort, Prime Recognition Task) for eliciting and representing the weather-related knowledge structures of pilots of varying experience. More specifically, the current study was designed to address the following research objectives:

1) Explore the similarities and differences between the three KETs in terms of 1) the extent to which the proximity data are correlated and 2) the ability to identify groups of pilots that maintain similar identifiable knowledge structures as a function of experience (Chapter 10).
2) Identify the factors or dimensions underlying pilots' knowledge structure for weather information and how those dimensions are impacted by 1) KET and 2) pilot experience (Chapter 11).
3) Validate and compare the three KETs in terms of their ability to 1) discriminate among pilot experience groups and 2) predict pilots' experience group based on their knowledge structures (Chapter 12).
4) Compare and contrast the KETs on the more practical aspects of their employment, including a) time and resource requirements for the researcher and the participant, b) data formatting and management requirements, and c) participants' subjective perceptions of the experience (Chapter 13).

The research objectives were achieved through the completion of three phases. *Phase I* was an information needs analysis designed to generate a master list of weather-related

information requirements for maintaining a safe flight. *Phase II* identified the most familiar and important weather-related items in that master list. *Phase III* employed each of the three KETs to elicit psychological proximity data for each of the item pairs.

Defining Experience

The design of the current study is predicated on the fundamental tenet that pilots who differ in experience will also differ in their knowledge structure. Ultimately, this tenet is an assumption but has its foundation in empirical research and practical application. Several studies in cognitive psychology have attributed the differences between expert and novice performance in large part to the differences with which information is organized in memory (Chapter 2 provides a review of some of these studies). The issue for the current study is how to operationally define "domain specific experience" such that differences in knowledge structure can be logically assumed.

Several different indicators have been used throughout aviation research to try to capture the notion of "domain-specific experience". For example, several studies have relied upon measures of *flight hours* to indicate domain experience, such as total flight hours (e.g., Goh & Wiegmann, 2002; Fiore et al., 2000), cross-country flight experience (Wiggins & O'Hare, 1995), and hours flown recently (Goh & Wiegmann, 2002). *Pilot certification* (i.e., private or commercial license) and *pilot ratings*[3] (e.g., VFR, IFR) have also been used (Goh & Wiegmann, 2002). However, it is unclear which of these measures of flight experience best captures domain-specific experience. In fact, Goh & Wiegmann (2002) hypothesize that the importance of certain measures of flight experience may be task-dependent, with some measures being more important for certain tasks and not for others.

[3] VFR stands for Visual Flight Rules and means that a pilot is only allowed to fly when they can maintain a constant visual with the ground (typically visibility of 3 mi and ceiling of 1000 ft. IFR stands for Instrument Flight rules and means that a pilot is able to fly using the instruments and, thus, do not have to maintain a constant visual of the ground.

The use of total flight hours as an indication of domain specific experience has been a fairly common practice in aviation research and within studies focused on knowledge structures (e.g., Fiore et al., 2000, Schvaneveldt et al., 1995). In addition, all pilots keep a log of their flight hours and the FAA has specific standards and requirements that involve achieving a certain number of flight hours in order to achieve different ratings and certificates. It is standard practice in the FAA to use total flight hours as an indication of where the pilots are in their career, although it is often paired with other measures as well to provide a more comprehensive assessment (T. McCloy & C. Donovan – FAA Scientific and Technical Advisors for Human Factors, personal communication, April 28, 2011). However, it must be noted that while total hours is a necessary component underlying pilot experience, it is certainly not sufficient for expertise to develop. If flight time does not occur under increasingly challenging conditions (when more domain knowledge and richer experiences can be gained), then increasing total flight hours will not be indicative increasing experience level (i.e., an asymptote in experience level will be quickly achieved).

Given that recruiting difficulties made it impossible to get an equal representation of pilots in terms of other potential indicators of experience level (e.g., ratings, cross-country flight experience, seasons of flying) total number of flight hours was used to capture and quantify domain-specific experience for this study. Thus, while the fundamental tenet that pilots who differ in experience will also differ in their knowledge structure is technically an assumption, it is a logical assumption given 1) previous research in expert-novice studies and memory organization and 2) its use as an indication of experience within the domain of aviation.

Chapter 7 - Weather Information Needs Assessment (Phase I)

The purpose of Phase I was to generate a master list of aviation weather-related information concepts. These concepts represent the type of weather information pilots need to access and/or maintain awareness of in order to maintain safe flight.

The first step in this assessment was to conduct an inventory of the weather-related information currently accessible to pilots through different forms of technology and support systems used both during flight planning and while enroute. System designers of commercially available MFDs (e.g., Honeywell/Bendix King) were informally interviewed regarding how the menu structure and rules of information organization and design were developed for their respective systems. In addition, as part of earlier FAA grant-funded work, menu structures were inventoried for three commercially available MFDs that display text and graphical weather (i.e., the complete menu structures were documented). Also, Williams and Joseph (1998) inventoried the various types of data available from the Operational and Supportability Implementation System (OASIS) to support their interface design research. OASIS is a computer-based system that allows the FAA's Automated Flight Service Stations (AFSSs) to provide weather briefing and flight planning assistance to the GA pilot population through data-link technology. Their inventory was consulted to inform the current study as well. Lastly, an inventory was taken of the types of data-linked weather information available through the Flight Information Service Data Link (FISDL) program.

The second step in this assessment was to consult previous information needs analyses performed for flight-related activities in the GA cockpit. Several previously conducted information needs analyses for flight-related tasks were used as an initial starting point (e.g., Groce, as cited in Jonsson & Ricks, 1995; Schvaneveldt et al., 2000). Also, Latorella, Pliske, Hutton and Chrenka (2001) conducted a cognitive task analysis of business jet pilots' weather

flying behaviors and identified some primary weather information requirements in the cockpit (see also Latorella, Lane & Garland, 2002; Vigeant-Langlois & Hansman, 2002).

The system inventories and research reviews provided a starting point for understanding the type of weather to which pilots needed access in the cockpit. However, each left questions unanswered. For example, system inventories only provided insight into weather information to which pilots had access, but not necessarily the type of information they *needed* for flight. Second, with the possible exception of Latorella et al. (2001), most previous information needs assessments were not specifically focused on understanding weather-related needs. Therefore, the third step to Phase I of this study involved applying design research techniques in order to more thoroughly understand pilots needs, goals, and context of using weather-related information.

Method

Participants

Five discussion groups were held with a total of 16 participants, 15 of whom were GA pilots. The additional participant was an aviation meteorologist who had flown with GA pilots many times but did not, himself, hold a private pilot's license. Table 7.1 summarizes the demographics and other relevant characteristics of the pilots participating in each discussion group. As Table 7.1 shows, the pilots represented various levels of aviation experience.

Table 7.1. *Pilot Demographic Data for Discussion Groups.*

Date/Time of Discussion Groups	Location	Number of pilots	Age *(avg)*	Total Hrs *(avg)*	Rating
8/18/05 (AM)	Garmin (Olathe KS)	2	26.5	297.5	All VFR-rated
8/18/05 (PM)	Garmin (Olathe KS)	3	52*	7750*	All IFR-rated
8/19/05	Aviation Weather Center (Kansas City , MO)	3	45.5**	2500**	2 IFR-rated; 1 aviation meteorologist (non-pilot)
9/9/05	Douglas Aviation (Memphis, TN)	6	24	1603	2 VFR-rated 4 IFR-rated
9/22/05	Kansas State Univ. (Manhattan, KS)	2	21.5	67.5	2 VFR-rated

** Based on data from two pilots; one pilot failed to turn in a demographics questionnaire.*
*** Based on data from two participants; the non-pilot participant (the aviation meteorologist) did not complete the demographics questionnaire.*

Procedure

Each discussion group began with a brief 15 minute presentation about the study. This presentation was designed to inform the pilots about the problem (i.e., organization of information in MFDs) and their role in the effort to solve the problem (i.e., the tasks in which they will be asked to participate). After the presentation, pilots were asked to sign the informed consent form. Discussion groups were video-recorded.

The discussion group itself was comprised of three different exercises (Word Association Task, Forward Scenario Simulation, structured group discussion), each designed to understand more about what weather information pilots use for flight planning and enroute. Each exercise approached knowledge gathering from different perspectives, with the goal of gaining a more thorough assessment of aviation weather-related information needs.

Exercise #1: Word Association Task

The Word Association Task was used as a method to brainstorm an initial list of important weather-related concepts (see Jonassen et al., 1993 for more information). During the task, pilots were shown a series of 28 weather- and/or aviation-related terms presented via

Power Point. For each weather information concept (the cue), they were asked to write down as many related terms as they could think of in one minute. They were told to write down terms immediately as they came to mind. After that one minute, they were asked to go back over the terms they generated and rank order the terms from *most* to *least* related to the original concept. Table 7.2 presents a listing of the 28 terms used as cues in the Word Association Task.

Table 7.2. *Weather / Aviation-Related Terms Comprising the Word Association Task.*

Alternate Airport	METAR	Forecasts	Autopilot
PIREPs	Altimeter Setting	Relative Humidity	Turbulence
SIGMET	Heading	Traffic	Fuel On-Board
Airspeed	Ambient Temperature	Precipitation	Winds Aloft
Low Pressure Center	Visibility	Ceiling	Taxiway
Altitude	Convective Activity	Waypoints	Icing
Estimated Time of Arrival	Radio Communications	Horizontal Situation Indicator	Minimum Safe Altitude

The lists produced by pilots could contain any number of the following: 1) alternative terminology for the cued concept, 2) other concepts that are related to the cued concept in some way (e.g., used to perform the same function, used in combination to perform a function, etc.), 3) concepts that have no real meaningful relationship to the cued concept (e.g., only recalled because the terms were learned in conjunction with each other, or they remember hearing them mentioned at the same time in the past). Therefore, the lists produced by the pilots were mainly used to create a more exhaustive list of weather information concepts and to identify more appropriate or frequently used terminology to describe or refer to the weather information.

Exercise #2: Forward Scenario Simulation

After the Word Association Task, pilots participated in a structured interview employing the Forward Scenario Simulation method (Burton & Shadbolt, 1987). Pilots were presented with

a scenario and then were asked to verbally identify the steps they would take to resolve the situation. This technique was chosen because it provides situational context for pilots to describe and elaborate on their flight-related problem-solving and decision making strategies without having to actually perform tasks (i.e., flight) during the session. It also does not require the researcher/facilitator to have prior knowledge about how the task is performed.

For each scenario, pilots were given basic information about the flight (aircraft type, how many pilots aboard, destination and departure airport, flight plan, weather conditions and flight route) as well as a narrative about the type of problem they encountered (see Appendix A) for a description of the scenarios used and the information provided). Pilots in each of the groups were free to discuss amongst themselves the possible actions appropriate for the given scenario.

Exercise #3: Structured Discussion

Lastly, the pilots were asked to discuss their general perspectives on and information usage pertaining to weather and aviation, including what types of weather information and weather sources they rely on to support their weather-related decisions, define their personal minimums for flight, etc. This time was also used to clarify any questions about information gathered during the Word Association Task and/or the Forward Scenario Simulation.

Results

Data from the Word Association Task were analyzed in the following way. Each concept generated for each cue word by each pilot was entered into an MS Excel file, along with the order in which it was generated (i.e., was it the first, second, third, etc.) and the rank order it was given by the pilot. Then, two Aviation Human Factors researchers independently went through each of the generated terms and grouped them into similar concepts (e.g., the concepts of “visibility” and “distance at which you can see” were grouped as the same concept of “visibility”).

The two researchers then met to review their individual groupings and identified and resolved discrepancies. The result was a master list of all concepts generated through the word association task. Terms for information concepts were changed to more appropriately match how pilots refer to them if necessary. Lastly, a frequency measure was calculated by summing the number of times each of these concepts was generated.

Tapes were viewed from the discussion sessions and notes were taken regarding pilots' problem-solving and decision making strategies voiced during the Forward Scenario Simulations and their answers to the structured discussion questions posed to them at the end of the session. The notes were reviewed and information extracted about the weather concepts and sources that pilots said they rely on for weather-related decisions. No efforts were made to gather frequency data on the number of times each concept was mentioned during either exercise. Rather, the end result was a listing of important weather/aviation-related information concepts mentioned at least once by at least one pilot during the exercises.

The resulting lists from the Word Association Task and the discussion exercises were cross-referenced with a list of important weather/aviation-related concepts identified in previous information needs and knowledge elicitation studies (e.g., Roske-Hofstrand & Paap, 1986; Williams & Joseph, 1998; Groce, as cited in Jonsson & Ricks, 1995; Schvaneveldt et al., 2000) to ensure consistency. The end result of Phase I was a comprehensive list of 68 unique information requirements pilots rely on for weather-related decision-making (see Table 8.2 in Chapter 8). These concepts provided the basis for the stimulus lists used in Phase II and ultimately in Phase III as well.

Chapter 8 - Identification of the Primary Weather Information Concepts (Phase II)

The purpose of Phase II was to identify the information concepts to be used in each of the three KETs to elicit pilots' knowledge structures for weather information in Phase III. The final list was identified through two steps. First, an online survey was conducted to identify the importance of the 68 concepts identified through Phase I. Second, because the goal of this study was to make comparisons across three different KETs, it was necessary to use the same list of concepts in each of the three KETs. Given the nature of two of the KETs, the list of concepts had to be reduced to 15 in order to accommodate pilots' time availability and to guard against fatigue.

Method

Participants

GA pilots were recruited to participate in this study. Emails were sent to nearby pilot organizations (e.g., KSU Salina, FAA employees, nearby flying clubs, etc.) and those who received the email were encouraged to forward it to other pilots. Advertisements were also placed in lobby areas of several local GA airports and posted to several GA pilot online forums. The only qualification for participation was that they had to hold a current flight license.

Sixteen pilots ultimately volunteered to participate (14 males, 2 females). Of these 16, all but one had their instrument rating. The IFR-rated pilots averaged 1096 total hours (SD=764 hrs), while the one VFR pilot reported having 225 total hours. Three of the 15 IFR-rated pilots reported not having flown at all under IFR during the previous six months. Table 8.1 provides a summary of the pilots' demographic information.

Table 8.1. *Demographic information for pilots participating in Phase II.*

	Number of pilots	**Gender**	**Age *(average)***	**Total Hours *(average)***
IFR-Rated pilots	15	13 male 2 female	32 yrs (stdev=13) Range: 21-56	1096 hrs (stdev=764)
VFR-Rated pilots	1	1 male	30	225 hrs

Procedure

An on-line survey was used to identify the most important weather-related concepts from the 68 concepts identified from the information needs assessment conducted in Phase I. The on-line survey consisted of a brief demographics questionnaire, followed by a listing of the 68 terms. Pilots were asked to rate the importance of each term for weather avoidance on 9 point-Likert scale (1 = not important at all; 9 = extremely important). They were given the following instructions: "For the following section, imagine that you are approximately 45 minutes into a scheduled 2 hour flight and it is your desire to avoid hazardous weather that may or may not affect your flight. For each of the following information concepts, please rate how important each is to the task of avoiding hazardous weather and maintaining safe flight."

The survey was on-line for approximately 3 weeks. Due to time constraints, data had to be extracted and analyzed at the end of 3 weeks' time. By this time, 16 pilots had completed the survey.

Results

Average importance ratings were calculated for each of the 68 terms that were rated. The terms were then divided into two categories – weather concepts and non-weather concepts. Weather concepts were defined as actual weather phenomena or reports of weather phenomena. Non-weather concepts were aviation-related terms that were not specific weather

events but could be affected by weather or could affect weather avoidance. Table 8.2 provides the mean importance ratings for each of the weather-related and non-weather-related concepts.

Table 8.2. *Pilots' Average Importance Ratings for Information Concepts (1-9 Likert Scale).*

Weather Concepts			***Non-Weather Concepts***	
Term	**Rating**		**Term**	**Rating**
Icing	8.58		Fuel On-board	8.67
Precipitation Type (e.g., rain, snow, hail, etc.)	8.42		Type of Flight (IVR, VFR)	8.00
Thunderstorms	8.42		Restricted Airspace	7.33
Freezing Level	8.33		Altitude AGL	7.17
Ceiling	8.08		AWOS	7.17
Ambient Temperature	8.00		NAVAID Operational Status	7.08
Lightning	7.75		ASOS	7.08
Convective SIGMET	7.58		ATIS	7.08
Cloud Type (e.g., scattered, overcast, etc.)	7.50		Minimum Safe Altitude (MSA)	7.00
Visibility	7.50		Approach Types	6.92
Weather Forecasts	7.42		Alternate Airport	6.83
Cloud Proximity	7.25		Margin Above Stall	6.58
Wind Velocity	7.17		NOTAMs	6.50
Terminal Aerodrome Forecast (TAF)	7.08		Runway Information (e.g., length, orientation, surface, active, etc.)	6.50
Surface Observation	7.00		Destination	6.25
Temperature/Dewpoint Spread	7.00		Radio Frequencies (e.g., FSS, Flight Watch, etc.)	6.17
Weather Trend (e.g., developing or dissipating)	7.00		Traffic Conflicts	6.17
SIGMET	6.92		Ground Track	6.08
PIREPs	6.83		Indicated Airspeed	6.08
Dewpoint	6.67		Nearest	6.00
METAR Report	6.67		Estimated Time of Arrival (ETA)	5.92
Weather Fronts	6.67		Altitude MSL	5.83
Turbulence	6.58		Estimated Time Enroute (ETE)	5.83
Wind Shear	6.58		Flight Plan	5.83
Wind Direction	6.50		Heading	5.75
AIRMET	6.33		Rerouting	5.75
Cloud Tops	6.33		Terminal Information	5.75
Radar	5.75		GPS Operational Status	5.67
Area Forecast (FA)	5.67		Engine RPM	5.58
NEXRAD	5.58		Traffic Information	5.50
Runway Visual Range (RVR)	5.42		Direct-to	5.17
			Airspace Class	5.08
			Barometric Pressure	4.83
			Datalink Status	4.75
			Departure Time	4.67
			Traffic Patterns	4.67
			Departure Point	4.08

Finalizing the List of Weather-Related Concepts

Because the goal of this study was to make comparisons across three different KETs, it was imperative that the same list of concepts was used in each of the three KETs. However, one of the ways the KETs differ is with respect to how the items are presented to participants. Specifically, the Relationship Judgment Task and the Prime Recognition Task both require the pair-wise assessment of each of the items within the concept list, whereas the Card Sort Task does not. The number of pair-wise combinations generated for a given list length (N) is defined by the following formula:

$$\text{Total \# of pair-wise comparisons} = \frac{N*(N-1)}{2} \quad (1)$$

Thus, the length of the study was directly proportional to the number of concepts. For example, a list of 30 concepts results in 435 pair-wise comparisons and a list of 40 concepts results in 1225 pair-wise comparisons. Given concerns about fatigue and the limited availability of GA pilots, the goal was to ensure that the Relationship Judgment Task and the Prime Recognition Task were each able to be completed in less than one hour. Therefore, the final list of weather-related concepts was limited to 15, resulting in 105 different pair-wise comparisons.

The final list of 15 most important weather-related concepts was derived from the information in the left column of Table 8.2 above. Concepts included in the final list had to be highly rated as important and they also had to satisfy the following constraints: 1) only concepts consisting of one or two words were used, and 2) if two concepts were too conceptually similar only one was used in the final list (e.g., thunderstorms and Convective SIGMET). These constraints were put into place so that the items in the final list would be appropriate for use in the Prime Recognition Task (the implicit elicitation task). The logic and importance behind these constraints are further discussed in Chapter 9 under the description of the Prime Recognition

Task. Table 8.3 lists the final 15 items used as stimuli for the knowledge elicitation tasks in Phase III.[4] A Professor from the Kansas State University (KSU)-Salina Aviation Program helped in the applying the constraints to the final concept list.

Table 8.3. *Fifteen Weather-Related Concepts to be used as Stimuli in the Phase III Knowledge Elicitation Tasks.*

Ambient Temperature	Dewpoint	Cloud Proximity
Precipitation Type	Icing	Turbulence
Freezing Level	Thunderstorms	Lightning
Visibility	Ceiling	Wind Direction
Wind Velocity	Sky Conditions	TAF

[4] Many of the concepts in Table 8.2 were used as distractor concepts in the Prime Recognition Task in Phase III.

Chapter 9 - Knowledge Elicitation (Phase III)

The information needs assessment conducted in Phase I and the importance ratings study in Phase II culminated in a list of 15 important weather-related concepts to be used as stimuli for the knowledge elicitation tasks employed in Phase III. Again, the purpose of Phase III was to derive pilots' knowledge structures for weather-related information.

To review, three techniques were used to elicit pilots' knowledge structures for weather-related information. Two of the techniques, the *Relationship Judgment Task* and the *Card Sort Task*, are well established within the psychological and HCI domains as techniques for eliciting knowledge structures. However, the major drawback of these techniques is that each explicitly asks pilots about how they organize information or how they perceive information to be related, and therefore, may be influenced by transitory contextual factors or by the cognitive processes necessary for completing these tasks. Consequently, the way pilots *say* the information is organized in their memory may not be the way they actually *think* about that information when posed with a real-world situation, especially under time stress and high workload. The *Prime Recognition Task* implicitly derives knowledge structures by assessing relationships or associations in memory through examining priming effects on response times to a basic memory task. Pilots are unaware of the intent to examine relationships between information concepts; therefore, the results should be less susceptible to influences of transitory or experimental factors affecting cognitive processing. Thus, comparisons will be made between conceptual structures derived using the Relationship Judgment Task, Card Sort Task and the Prime Recognition Task.

Hypotheses

The current study was largely exploratory in nature. Therefore, very few hypotheses were generated going into the data collection. However, based on previous research regarding

knowledge structures and experience, a few hypotheses were set forth across the four objectives of the study:

Objective #1: Explore the similarities and differences between the three KETs in terms of 1) the extent to which the proximity data are correlated and 2) the ability to identify groups of pilots that maintain similar identifiable knowledge structures as a function of experience.

- Hypothesis #1: Relationship Judgment and Card Sort should show more similarity in how information concepts are related (i.e., the underlying dimensions) compared to the Prime Recognition Task. The similarity will be due to the fact that they both require pilots to explicitly consider how the concepts are similar to each other.
- Hypothesis #2: Given previous research on the effect of experience on knowledge structures, the more experienced group of pilots will show less variability in their knowledge structures than will the less experienced groups of pilots.

Objective #2: Identify the factors or dimensions underlying pilots' knowledge structure for weather information and how those dimensions are impacted by 1) KET and 2) pilot experience

- No specific hypotheses were generated for what dimensions or factors may underlie pilots' knowledge structures for weather information. However, given that the 15 weather concepts were identified at least in part because of their importance for avoiding hazardous weather, it is expected that the concept of severity or hazard would be represented in some manner in conceptual structures from valid KETs. Further, pilots with more experience should have a different interpretation of severity than less experienced pilots.

Objective #3: Validate and compare the three KETs in terms of their ability to 1) discriminate among pilot experience groups and 2) predict pilots' experience group based on their knowledge structures.

- Hypothesis #3: Conceptual structures derived from knowledge structures using the Prime Recognition Task should be more consistent among pilots of similar experience levels because they should be uninfluenced by variability introduced by the cognitive processing required of the explicit techniques. Therefore, given that previous research suggests that experts have more well organized knowledge structures than novices, the Prime Recognition Task should be able to better discriminate among pilots of various experience groups based on the differences in their knowledge structure organizations than the other KETs.

Objective #4: Compare and contrast the KETs on the more practical aspects of their employment, including 1) time and resource requirements for the researcher and the participant, 2) data formatting and management requirements, and 3) participants' subjective perceptions of the experience.

- No specific hypotheses were generated for this objective. A qualitative assessment and comparison between the KETs will be made based on the effectiveness and efficiency with which each supports elicitation of knowledge.

Method

Participants

Recruiting posed a significant challenge to this study, both in the ability to access GA pilots and in the need to convince them to volunteer significant amounts of their time for no monetary compensation. Therefore, to ensure adequate numbers of participants, the only qualification for participation was that they had to at least hold a current GA private pilots'

license. GA pilots were recruited to participate through multiple different avenues. Some were students and instructors from the KSU-Salina Aviation Program, some were affiliated with the FAA and/or other government agencies involved in aviation, and some were recruited based on their affiliation with local flying clubs. Data collection locations are listed in Table 9.1 below. In total, 53 pilots participated in data collection for Phase III. After data collection was complete, however, seven participants' data were excluded from further analysis because it was discovered that these pilots did not hold a current flight license.[5]

Table 9.1. *Data Collection Locations.*

MITRE	*Tyson's Corner, VA*
FAA Headquarters	*Washington, DC,*
General Aviation Conference	*Alexandria, VA*
Garmin	*Olathe, KS*
KSU Salina Campus	*Salina KS*
KSU Department of Psychology	*Manhattan, KS*

Pilots were classified into one of three Experience Groups based on their total number of hours flown (see Chapter 6 for more discussion on using total number of hours flown as an indication of pilot experience). When specifying cutoffs between groups, care was taken to minimize the standard deviation in total hours among pilots in each group while also ensuring relatively similar numbers of pilots within each of the groups. Table 9.2 provides a summary of the demographics information for the three Pilot Experience Groups.

Six of the 46 pilots were VFR-rated only, meaning they were only rated to fly under Visual Flight Rules (VFR). Five of these pilots were categorized as Low-Time pilots and one

[5] During data collection, all pilots who volunteered were allowed to participate, regardless of whether they had maintained currency with their license. Once data collection was complete it was decided that enough current pilots had participated so those without current licenses were dropped from the study. The majority of those excluded from the analysis were pilots who had retired from the commercial airline industry with over 10,000 hrs of total flight time.

was categorized as a High-Time pilot, based on their total flight hours. The rest of the 46 pilots had their Instrument Rating (IFR).

Table 9.2. *Demographic information for the Pilot Experience Groups based, in large part, on total number of hours flown.*

Group	# in Group	Ave Age *(yrs)*	Ave Total Hrs Flown	# of hrs flown in last 90 days	# of hrs flown in last 6 mo	Range of total hrs
Low-Time Pilots	15	25.5 (SD=9.9)	208 (SD=76.7)	3.5 (SD=3.8)	37.1 (SD=33.4)	65-310
Mid-Time Pilots	14	40.9 (SD=15.9)	700 (SD=284.1)	16.7 (SD=23.8)	95.3 (SD=109.7)	336-1185
High-Time Pilots	17	49.4 (SD=10.4)	5722.2 (SD=3577)	17.3 (SD=30.7)	98.3 (SD=100.8)	1660 – 15,500

Note that the High-Time group consisted of pilots with the widest range of total hours flown (1660 – 15,500 hrs). While this large range constituted a potential concern, the decision was made to conduct the initial analyses with pilots divided into these three groups in order to maximize the number of participants in each group. It was decided that additional analyses could be conducted with the High-Time group subdivided into two smaller groups if the initial results imply it is necessary. For example, it may be that knowledge structures of pilots with 1660 hrs vary widely from knowledge structures of pilots with over 10,000 hrs, resulting in too much variability to identify dimensions that underlie knowledge structures of experienced pilots. In that case, additional analyses could be conducted with this group further divided into two subgroups to see if consistency in knowledge structure improves, but the analysis will require a sacrifice in sample size. Alternatively, it may be the case that once a certain level of experience is achieved, knowledge structures become more consistent and therefore a 1000 hr pilot may be more similar to a 10,000 hr pilot than a 100 hr pilot in terms of how information is structured in memory. Therefore, initial data analysis was conducted with the three Pilot Experience

Groups as defined above with the possibility of conducting additional analysis with the High-Time pilot group sub-divided if the results from the initial data analysis indicate a need.

Procedure

A general experimental procedure was followed for all three KETs. Pilots were run one or two at a time.[6] The Relationship Judgment Task and the Prime Recognition Task were both performed on a laptop computer and keyboard with E-Prime Psychology Software (Schneider, 2000) used to present the stimuli for the task and collect the data (Schneider, 2000). Card Sorts were performed by hand using concepts printed individually on index cards. Depending on time availability, one, two, or three Card Sorts could be completed. Each Card Sort differed in the number of cards needed to be sorted: 15, 36, and 80 cards.[7] Only the data from the 15-card Card Sort were included in the final analysis. The stimuli for the Relationship Judgment, Card Sort, and Prime Recognition Tasks consisted of the 15 concepts listed in Table 8.3.

Each data collection session began with a brief 15 minute presentation about the study. This presentation was designed to inform the pilots about the problem (i.e., organization of information in MFDs) and their role in the effort to solve the problem (i.e., the tasks in which they will be asked to participate). Pilots were told that this was a study designed to understand how pilots think about weather information. They were also told that the study was funded by an FAA grant. After the presentation, pilots were asked to sign the informed consent form and then asked to fill out a brief demographics questionnaire.

---

[6] Even when pilots were run two at a time, they were isolated from each other and each completed the KETs separately, without input or interference from the other pilot.

[7] The 36-card and 80-card card sorts consisted of additional weather-related concepts and aviation related concepts identified from the information needs analysis in Phase I. These data were collected as part of a separate study.

Pilots were pseudo-randomly[8] assigned to either the Relationship Judgment Task or the Prime Recognition Task (Table 9.3). Because the Relationship Judgment Task only required approximately 40 min, all but one pilot in the Relationship Judgment Task also completed at least one of the three Card Sorts. Because the Prime Recognition Task lasted over one hour, pilots assigned to that group were only asked to perform Card Sort(s) if their schedules allowed for it. Data collection sessions lasted one to three hours, depending on the number of KETs completed. Overall, 75% of the 46 pilots completed two KETs – the Card Sort and either the Relationship Judgment or Prime Recognition Task. No pilot participated in ***both*** Relationship Judgment and the Prime Recognition Task. Table 9.4 provides a summary of the number of participants in each Pilot Experience Group to complete two KETs as well as the number of pilots completing just one of the KETs.

Table 9.3. *Number of pilots from each Pilot Experience Group who participated in each of the KETs. Note: Pilots often participated in more than one KET.*

Pilot Experience Group	**Relationship Judgment (RJ)**	**Card Sort (CS)**	**Prime Recognition Task (PRT)**
Low-Time Pilots	6	12	9
Mid-Time Pilots	5	12	8
High-Time Pilots	8	14	7
Total	**19 pilots**	**38 pilots**	**24 pilots**

Table 9.4. *Number of pilots who completed each KET or combination of KETs.*

Pilot Experience Group	**Pilots completing 2 KETs**		**Pilots completing only 1 KET**			
	RJ & CS	*PRT & CS*	*RJ only*	*CS only*	*PRT only*	*Total*
Low-Time Pilots	6	6	0	0	3	*15*
Mid-Time Pilots	5	6	0	1	2	*14*
High-Time Pilots	6	6	2	2	1	*17*
Total number of pilots	17	18	2	3	6	46
	35 pilots		11 pilots			

[8] During the last several data collection sessions, care was taken to assign pilots to KETS so that there would be relative equality in pilot experience (total numbers of hours flown) across each KET.

Relationship Judgment Task (RJ)

The Relationship Judgment Task began with brief but explicit instructions on how to complete the task on the computer. Pilots were familiarized with the 1-9 Likert scale on which to base their judgments (1=not related at all; 9 = highly related). The pilots were not given an explicit definition of what "relatedness" meant. Instead, they were just told to make their judgment for a pair of concepts based on whatever the term "relatedness" meant to them. After they read the instructions, they were given three practice trials using concepts not appearing in Table 8.3 and were told to ask for clarification about the task if needed. They then completed the 105 relatedness ratings (pairwise comparisons with each of the 15 concepts). The relatedness of each pair of concepts was judged only once.

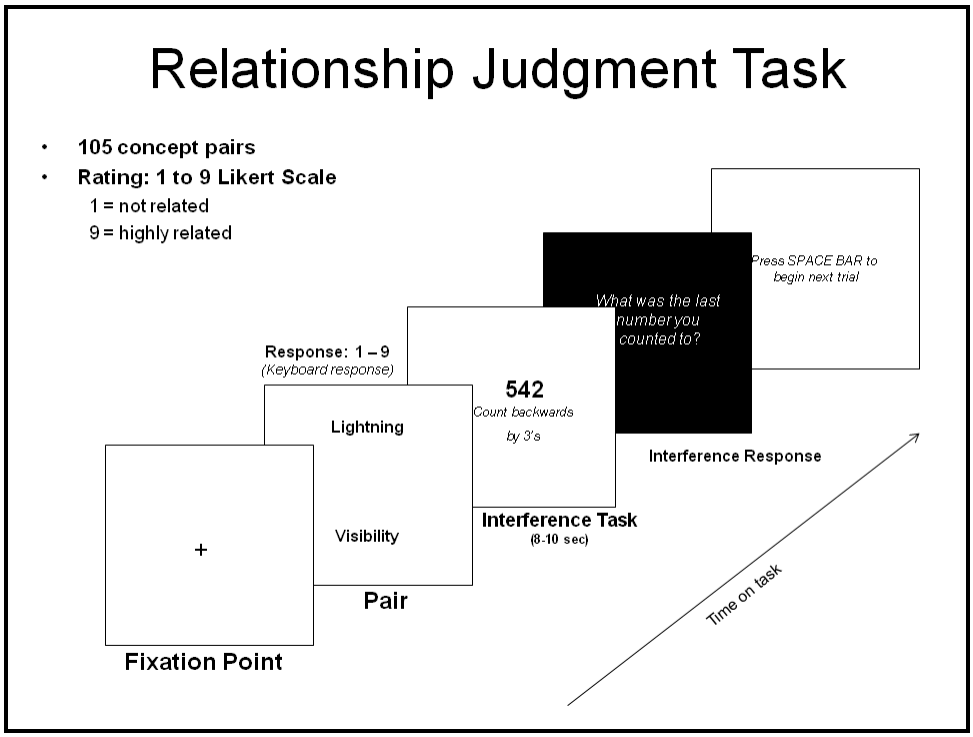


Figure 9.1. Illustration of the procedure for a Relationship Judgment procedure.

Figure 9.1 provides an illustration of the procedure for a given trial. Each trial began with a presentation of a fixation point (1000ms), followed by a pair of information concepts. Pilots

were instructed to provide their rating of the relationship between the two concepts as quickly but as truthfully as possible. After they entered their 1-9 rating via computer keyboard, they were presented with a three digit number and asked to count backwards by 3's until prompted to stop (8-10 sec). After 8-10 seconds, a screen appeared asking them to enter via the computer keyboard the last number to which they counted. After they entered the number, they were asked to press the space bar to begin the next trial. This "interference task" of counting backwards by 3's was designed to make it difficult for pilots to remember their previous judgment, as previous research has shown that a relationship judgment can be unintentionally influenced by the immediately preceding judgment (Canas et al., 2001). Thus, the interference task was used in an attempt to eliminate any traces left in WM from a previous pair-wise comparison rating before making the next pair-wise comparison. The Relationship Judgment Task took approximately 40 minutes to complete.

Prime Recognition Task (PRT)

As Figure 9.2 illustrates, a trial in the Prime Recognition Task started with the presentation of a memory set of four information concepts. The pilot was then presented with a fifth information concept not in the memory set (the prime) and was told to memorize all five concepts. After the prime, the pilot was presented with another information item (the target) in red font. On 60% of the trials, the target was in the memory set (i.e., target trials) while on the other 40% the target was not in the memory set (i.e., foil trials). The pilot's task was to decide as quickly as possible whether or not the target was part of the memory set by responding "Yes" or "No" using the appropriately labeled key on the computer keypad. The underlying assumption was that if the prime and target are related in the knowledge structure, the activation of the prime will facilitate the activation of the target, resulting in a shorter time to respond that the target was in the memory set.

To eliminate some of the variability associated with response time measures, participants were instructed to complete the task with each index finger constantly positioned on the "Yes" and "No" buttons. For half of the participants, the "Yes" button was mapped to their right index finger and for the other half of participants it was mapped to their left index finger. This arrangement was to guard against any possible confounds associated with response button placement.

To reduce the possible build-up of proactive interference, the pilots performed one of two interference tasks after they made their decision about whether or not the target was in the memory set. On 60% of the trials, they were asked to complete the same interference task as in the Relationship Judgment Task (i.e., they were prompted to count backwards by three's from a randomly generated three-digit number for 8-10 seconds). On 40% of the trials, they were asked to complete a search task in which they were presented with a display of different-colored asterisks for approximately 2 seconds and then had to respond with how many red asterisks they counted. Both tasks were used to ensure memory traces were cleared in both the auditory and visual components of WM.

Each of the 15 weather-related concepts was used as either a prime or a target, resulting in 105 prime-target pairs (i.e., target trials) whose data were used to construct the conceptual structures. Fifteen additional target trials were added but were not used to construct the conceptual structures, bringing the total number of target trials to 120. In addition, several other information concepts from Phase I were used as distracters in the memory set[9], primes, and as probes, to comprise a total of 79 foil trials. Thus, the total number of trials seen by each pilot was 199. The Prime Recognition Task took approximately 60-70 minutes to complete.

Measures were taken in constructing the memory sets to protect against participants responding whether or not the target was in the memory set based on physical or visual

[9] Memory sets for the 105 prime-target pairs ("target trials") were comprised of combinations of the 15 weather-related information characteristics.

characteristics of the concepts rather than their meaning. For example, if the target was a two-word concept, then all of the items in the memory set were two-word concepts. If the target was one word, then all of the items in the memory set were one word. If the target was an acronym, then all of the items in the memory set were acronyms. Also, the distractor items in the memory set were chosen to have relatively low conceptual similarity to the target (e.g., if "precipitation" was the target, then the memory set would not include distractor items like "rain" or "snow" or "fog").

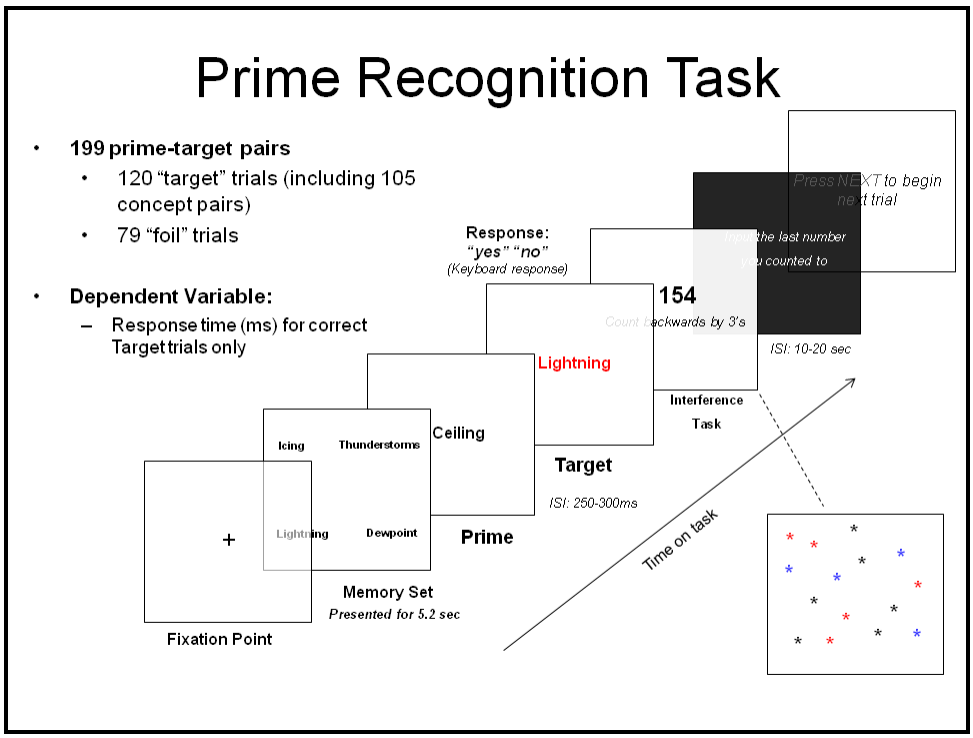


Figure 9.2. Illustration of the procedure for a Prime Recognition Task trial.

Card Sort Task (CS)

Pilots completed up to three card sorts, depending on their time availability. Card Sort #1 consisted of the same 15 concepts from Phase II that were used in the other two KETs. Card Sort #2 consisted of 36 weather-related concepts (the 15 original concepts and 11 additional weather-related concepts from Phase II). Card Sort #3 consisted of 80 concepts (the 36

weather-related concepts in Card Sort #2 and 44 additional non-weather-related concepts from Phase I). [Only data from Card Sort #1 were analyzed as a part of the current study.]

To begin the Card Sort task, each pilot was presented with the stack of cards along with several blank cards (Figure 9.3). Each card had an information concept printed on one side. The card sort was an "open" card sort, meaning that participants create their own groups (as many or as few as they like) and apply labels to each of the group. This is different than a closed card sort in which participants are given pre-determined categories into which they are to sort a stack of cards.

Pilots were instructed to read every card and sort the cards into groups that seemed appropriate. They were given no further instructions on how the cards were to be sorted or the number of groups that could be created. Pilots were instructed to not be concerned with trying to organize the information as they have seen it organized in MFDs or other displays. Rather, the pilots were told to organize the cards as they perceived the relatedness of the concepts.

Since the goal of the card sort was to understand how participants think about and relate the concepts to each other (i.e., elicit their knowledge structure of the concepts), very few restrictions were placed on their ability to organize the cards. They were given several allowances with their Card Sort for fear that some procedural restrictions may artificially influence their groupings:

- They were allowed to create a duplicate card if they felt the concept belonged in more than one category.
- If they encountered a concept that absolutely did not belong in a cockpit display, they were allowed to create an "outlier" pile
- They were allowed to create hierarchies of card groups (i.e., subgroups of cards within a larger group).

- If a pilot found a card whose label did not seem appropriate or descriptive of the concept, he/she was instructed to write a “better” label on the card.

Once the groups were established, pilots were asked to provide labels for each group and/or subgroup. The number of times each information concept was grouped with another was calculated by hand and input into a matrix as proximity data.

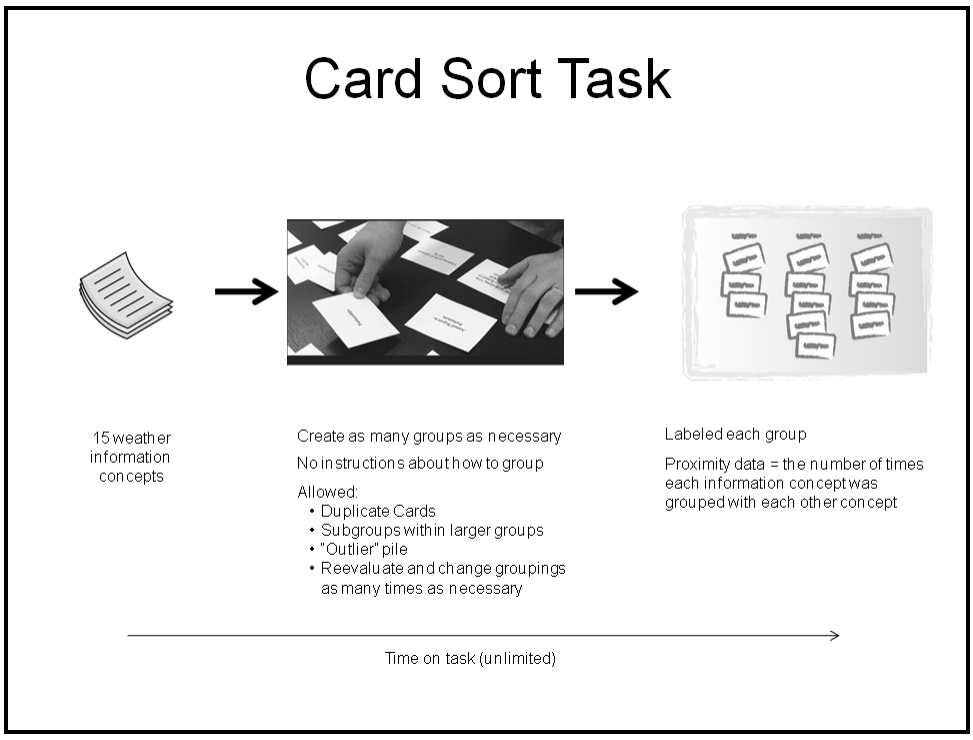


Figure 9.3. Illustration of the procedure for the Card Sort task.[10]

Preparation of Data for Analysis

Raw data collected from each of the three KETs were submitted to initial analyses that served two major purposes. First the analyses helped determine the best approach to coding and/or trimming the data. Second, the analyses were conducted to explore whether the Pilot Experience Groups differed in their general approach or ability to perform the KETs. The

[10] (Images courtesy of the following websites: http://nform.ca/tradingcards/2007_01.jpg, http://www.officeclipart.com/office_clipart_images/a_stack_of_papers_and_documanets_in_an_office_0515-1007-3003-0814_SMU.jpg, http://www.foviance.com/wp-content/uploads/2009/03/card-sorting-3.jpg)

following section provides a brief description of data preparation procedures and a summary of the results from the initial explorations into differences between pilot experience groups on raw data from each of the three KETs. Appendix B provides more detailed discussion of the results.

Data Coding & Formatting

While all data from each of the KETs had to be prepared for analysis, some KETs required more data preparation than others. The goal of data preparation was to transform the raw data into dissimilarity data (i.e., higher numbers correspond with greater dissimilarity) because it is recommended that MDS be applied to dissimilarity data rather than similarity data. The Relationship Judgment Task required the least amount of data preparation while the Card Sort task required the most data preparation.

Relationship Judgment. When the data were extracted from the E-Prime Psychology Software program, it was in the form of similarity judgments between pairs of concepts because participants were asked to rate the relatedness with higher numbers indicating greater relatedness or similarity on a 9-point scale. Thus, each relationship judgment data point was recoded to represent a dissimilarity judgment by subtracting it from 10 (dissimilarity = 10 – similarity judgment).

Prime Recognition Task. Data extracted from the E-Prime Psychology Software program were in the form of response times between prime-target pairs of concepts. Response times were fairly uniform without the presence of many extreme outliers. To help decide whether a trim should be applied to the data, some initial exploratory analyses of variance (ANOVAs) were conducted. Response times were compared between the Pilot Experience Groups on both target and foil trials. One analysis used the untrimmed data while the other analysis was used data in which a two standard deviation trim was applied to each individual participant's data (i.e., response times that were greater than or less than two standard deviations away from the participant's mean were replaced by that outer fence value). A total of 4.6% of the data points

were replaced with the trimmed 2 standard deviation value. All of the replaced data points were on the upper tail of the distribution. Trimming the data had no effect on the results, relative to the untrimmed data (Appendix B provides more information on the results of this analysis). Therefore, results based on untrimmed data are reported to maintain consistency with the idea that response time is a measure of semantic distance between pairs of items in memory. Trimming the response times could artificially affect the validity of that measure of proximity in ways that are not clearly understood. Prime Recognition Task data were already in the form of dissimilarity data when they were collected, as larger response times between prime-target pairs represented greater dissimilarity.

Card Sort. Because participants physically organized cards into groups, rather than using an online or computerized tool, the first step of data coding required that the relationship (or lack thereof) between each of the concepts be recorded by hand into an Excel spreadsheet. However, allowing duplicate cards, outlier piles, and hierarchies significantly complicated the data coding procedure. Therefore, the data were coded using two different procedures. The first procedure was consistent with much of the usability literature (e.g., Spencer, 2009) and involved collapsing hierarchical groups into a single level. Items that occur within the same parent group were treated the same regardless of if they were in the same or different sub-groups. Pairs of items that occurred within the same group were assigned 1s and pairs of items that did not occur within the same group were given 0s.

The second procedure used Jaccard scoring to code the groupings (Capra, 2005). Jaccard scoring accounts for hierarchal groups on a continuous scale in that it places different weights on items depending on whether they occur within the same group or within different subgroups under the same parent group. Items occurring in different subgroups under the same parent group are given scores greater than 0 (meaning there is some relationship between the items) but less than 1 (meaning they did not occur in the same immediate group). Similarity

scores calculated using Jaccard scoring were highly correlated with similarity scores calculated using traditional card sort coding procedures (r=.97, $p \leq .05$). However, because the Jaccard scoring takes into account the hierarchical groups created by pilots, it was decided to use the Jaccard scoring procedure as it provided a more accurate representation of how the pilots truly sorted the concepts. (See Appendix C for more information about how Jaccard similarity scores are calculated and Appendix B for more information on the results of the exploratory analyses). Jaccard similarity scores were then recoded into dissimilarity scores by subtracting each score from 1 (dissimilarity = 1-similarity). Thus, the Jaccard Card Sort data were on a continuous scale with 0 indicating not dissimilar (i.e., two items placed in the same group) and 1 indicating highly dissimilar (i.e., two items were not placed in the same group).

Explorations of Pilot Experience Group Differences

One-factor (Pilot Experience Group) ANOVAs were conducted on each of the data sets from the three KETS – the dissimilarity judgments from the Relationship Judgment Task, the Jaccard similarity scores from the Card Sort, and the accuracy and response time data from the Prime Recognition Task. There was no effect of Pilot Experience Group on the dissimilarity judgments or on the Jaccard similarity scores (p=*n.s.*). There was also no effect of Pilot Experience Group on accuracy for the 105 target trials in Prime Recognition Task or in the average response times for those trials. However, upon further analysis of the response time data, it was discovered that one Low-Time pilot had an average response time that was more than two standard deviations longer than the mean for Low-Time pilots. When the analysis was redone with the removal of this pilot (analysis based on 8 pilots rather than 9), there was still no effect of Pilot Experience Group on accuracy, but there was a main effect of Pilot Experience Group on response time ($p \leq .05$). Low-Time pilots had the quickest average response times. Mid-Time pilots and High-Time pilots did not significantly differ in average response time. See Appendix B for a more detailed discussion of the results for each of the three KETs.

Chapter 10 - KET Data Exploration (Phase III)

Initial data explorations and comparisons between the data elicited from each KET were designed to address two specific research questions:

1) To what extent are the data elicited by each technique correlated for the 15 weather-related items?
2) Do pilots with similar levels of experience maintain similar knowledge structures and, if so, are there differences in how the techniques are able to identify these similarities in knowledge structure?

Correlations between KETs

Overall Correlations

Within each KET, raw data[11] for each of the 105 weather-related items were averaged across all pilots, regardless of Pilot Experience Group. Thus, each of the 105 concept pairs had an average Relationship Judgment score, an average response time score (from the Prime Recognition Task), and an average Card Sort similarity score. Significance was assessed at an alpha level of .05.

The overall correlation between *Relationship Judgments and Card Sort* was significant (r=.46) and positive, meaning that higher dissimilarity relationship judgments corresponded with higher card sort dissimilarity scores. The correlation between *Card Sort* and the *Prime Recognition Task* was marginally significant (r = -.17, p<.10). Also note that the correlation is negative, meaning that longer response times were associated with smaller Card Sort dissimilarity scores. Longer response times were hypothesized to indicate weaker relationships (i.e., less similarity) between the prime-target pairs of weather-related items. Therefore, if Card

[11] Data from each KET were on different scales (Relationship Judgment: 1-9 Likert scale; Prime Recognition Task: response time (ms) data bounded only on the lower tail (0) and positively skewed; Card Sort: 0-1). Reported analyses were performed using raw data. However, separate analyses revealed that standardizing the KET data to z-scores had no effect on the strength or the outcome of the correlations.

Sort and Prime Recognition Task were eliciting similar structures, longer response times would be expected to be associated with larger Card Sort dissimilarity scores (i.e., a positive correlation). This negative correlation, although only marginally significant, may suggest that Card Sort and Prime Recognition Task are tapping into different types of knowledge structures. The correlation between *Prime Recognition Task* and *Relationship Judgment* was not significant (r = -.02, p=*n.s.*).

Correlations between KETs within each Pilot Experience Group

For this analysis, data from each of the 105 weather-related items were averaged across pilots within each of the three Pilot Experience Groups for each KET. In other words, for each Pilot Experience Group, each of the 105 concept pairs had an average Relationship Judgment score, an average response time (from the Prime Recognition Task), and an average Card Sort similarity score. Again, significance was assessed at an alpha level of .05.

Relationship Judgment and Card Sort. The correlation between Relationship Judgment and Card Sort was significant for all three levels of Pilot Experience Group: Low-Time (r = .43), Mid-Time (r =.26), and High-Time (r = .53). The correlation was strongest for High-Time pilots, which is not surprising considering that previous research has shown that pilots with more experience exhibit more consistency in their knowledge structures. Thus, if both techniques are tapping into the same type of knowledge structure, one would expect the correlation to be stronger when there is less variability in the knowledge structures themselves. However, it is also interesting to note that Relationship Judgments and Card Sort scores were more strongly correlated for Low-Time pilots than Mid-Time pilots. Skill acquisition literature provides one possible explanation for the lower correlation for Mid-Time pilots. Perhaps these pilots are in the process of transitioning between declarative and procedural knowledge which may lead to higher variability within this group relative to the Low-Time pilots who may be heavily relying on

declarative knowledge (e.g., Rasmussen,1983). The higher within-group variability for Mid-Time pilots may have led to the lower correlation between the two KETs.

Prime Recognition Task and Card Sort. The correlation between the Prime Recognition Task and Card Sort failed to reach significance for any of three pilot experience groups (*p*=*n.s.*). The correlation was weak regardless of pilot experience.

Relationship Judgment and Prime Recognition Task. There was no correlation between Relationship Judgments and response times from the Prime Recognition Task for any of the Pilot Experience Groups (*p*=*n.s.*). Again, the correlation was weak regardless of Pilot Experience Group.

Correlations between KETs for Within-Subjects Data

Recall that 75% of the 46 pilots who participated in the study actually participated in two of the three KETs – Card Sort and either Relationship Judgment or Prime Recognition Task (no pilot participated in ***both*** Relationship Judgment and the Prime Recognition Task). Therefore, correlations were examined between Card Sort and Prime Recognition Task and between Card Sort and Relationship Judgment using data from pilots who completed both techniques. Again, significance was assessed at an alpha level of .05. Table 10.1 shows correlations coefficients for all data and within-subjects data.

Relationship Judgment and Card Sort. There was a significant positive correlation between Relationship Judgment and Card Sort (r = .60). As dissimilarity judgments increased, Jaccard dissimilarity scores increased. This significant positive correlation was consistent across all three levels of Pilot Experience. Similar to the results that were based on all participants, the correlation was again strongest for the High-Time pilots (r = .54) followed by the Low-Time pilots (r = .48) and was weakest, but still significant, for Mid-Time pilots (r = .41). Not surprisingly, the correlation between Relationship Judgment and Card Sort when collapsed

across Pilot Experience Group was stronger when based on within-subjects data (r = .60) compared to when it was based on data from all pilots for those KETs (r = .46).

Prime Recognition Task and Card Sort. There was no correlation between Prime Recognition Task data and Card Sort data when collapsed across Pilot Experience or as a function of Pilot Experience Group.

Summary. Analyses revealed a relatively strong correlation between the data elicited by the Relationship Judgment Task and the Card Sort task. Correlations were highest for High-time pilots. These results suggest that Relationship Judgment and Card Sort may be tapping into similar types of knowledge and the structure of that knowledge gets more consistent as pilot experience increases. The Prime Recognition Task, however, may be tapping into a different type of knowledge compared to the other two techniques. However, because the data used in this analysis were averaged across participants on each of the 105 individual item pairs, it provides only a cursory glance into pilots' knowledge structures for weather information. The analysis discussed in the next section provides more insight into comparisons of knowledge structures between individual pilots.

Table 10.1. *Correlations between each Knowledge Elicitation Technique on all data and on within-subjects comparisons (where each pilot completed both techniques).*

KET Comparison	All Data				Within Subjects Data			
	Collapsed Across all Pilots	*Low-Time Pilots*	*Mid-Time Pilots*	*High-Time Pilots*	*Collapsed Across all Pilots*	*Low-Time Pilots*	*Mid-Time Pilots*	*High-Time Pilots*
Relationship Judgment & Card Sort	r =.46*	r = .43*	r =.26*	r = .53*	r = .60*	r = .48*	r = .41*	r = .54*
Prime Recognition Task & Card Sort	r = -.17^	r = -.14	r =- .11	r =- .06	r = -.15	r = -.10	r = -.09	r = -.12
Relationship Judgment & Prime Recognition Task	r = -.02	r = .02	r = -.05	r = .003	N/A	N/A	N/A	N/A

* p<.05
^ p<.10

Identifying Similarities in Conceptual Structures

The objective of this analysis was to assess whether conceptual structures were more similar between pilots of the same experience group than between pilots of different experience groups and, if so, whether this greater similarity within pilot group occurred consistently throughout all three KET datasets. In order to assess similarities between pilots' conceptual structures, an index of similarity between pilots was first obtained. For each KET dataset, correlations were performed between all pairs of pilots who completed that KET. Pilots were paired with every other pilot whether they were in their same experience group or not. The correlations were based on each pilots' data for each of the 105 concept pairs. This analysis produced Pearson *r* correlation coefficients for each pair of pilots. High *r* values indicated pairs of pilots who were similar in their judgments/responses when completing that particular KET.

Because the data were symmetrical (i.e., the correlation between P1 and P2 was the same as the correlation between P2 and P1), the correlation coefficients were organized and placed into half-matrices for each KET. Each matrix was constructed with pilots ordered based on the Pilot Experience Group to which they were assigned (see Figure 10.1). Values along the diagonal of each matrix (i.e., the white cells in the bottom half of the matrix) corresponded to members of the same Pilot Experience Group (i.e., intragroup pairs). Values toward the lower left corner of each matrix (i.e., the gray cells in the bottom half of the matrix) corresponded to pairs of participants who were from different Pilot Experience Groups (i.e., intergroup pairs).

	Low-Time Pilots						Mid-Time Pilots						High-Time Pilots						
	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19
P1	1																		
P2	r_{12}	1																	
P3	r_{13}	r_{23}	1																
P4	r_{14}	r_{24}	r_{34}	1															
P5	r_{15}	r_{25}	r_{35}	r_{45}	1														
P6	r_{16}	r_{26}	r_{36}	r_{46}	r_{56}	1													
P7	r_{17}	r_{27}	r_{37}	r_{47}	r_{57}	r_{67}	1												
P8	r_{18}	r_{28}	r_{38}	r_{48}	r_{58}	r_{68}	r_{78}	1											
P9	r_{19}	r_{29}	r_{39}	r_{49}	r_{59}	r_{69}	r_{79}	r_{89}	1										
P10	r_{110}	r_{210}	r_{310}	r_{410}	r_{510}	r_{610}	r_{710}	r_{810}	r_{910}	1									
P11	r_{111}	r_{211}	r_{311}	r_{411}	r_{511}	r_{611}	r_{711}	r_{811}	r_{911}	r_{1011}	1								
P12	r_{112}	r_{212}	r_{312}	r_{412}	r_{512}	r_{612}	r_{712}	r_{812}	r_{912}	r_{1012}	r_{1112}	1							
P13	r_{113}	r_{213}	r_{313}	r_{413}	r_{513}	r_{613}	r_{713}	r_{813}	r_{913}	r_{1013}	r_{1113}	r_{1213}	1						
P14	r_{114}	r_{214}	r_{314}	r_{414}	r_{514}	r_{614}	r_{714}	r_{814}	r_{914}	r_{1014}	r_{1114}	r_{1214}	r_{1314}	1					
P15	r_{115}	r_{215}	r_{315}	r_{415}	r_{515}	r_{615}	r_{715}	r_{815}	r_{915}	r_{1015}	r_{1115}	r_{1215}	r_{1315}	r_{1415}	1				
P16	r_{116}	r_{216}	r_{316}	r_{416}	r_{516}	r_{616}	r_{716}	r_{816}	r_{916}	r_{1016}	r_{1116}	r_{1216}	r_{1316}	r_{1416}	r_{1516}	1			
P17	r_{117}	r_{217}	r_{317}	r_{417}	r_{517}	r_{617}	r_{717}	r_{817}	r_{917}	r_{1017}	r_{1117}	r_{1217}	r_{1317}	r_{1417}	r_{1517}	r_{1617}	1		
P18	r_{118}	r_{218}	r_{318}	r_{418}	r_{518}	r_{618}	r_{718}	r_{818}	r_{918}	r_{1018}	r_{1118}	r_{1218}	r_{1318}	r_{1418}	r_{1518}	r_{1618}	r_{1718}	1	
P19	r_{119}	r_{219}	r_{319}	r_{419}	r_{519}	r_{619}	r_{719}	r_{819}	r_{919}	r_{1019}	r_{1119}	r_{1219}	r_{1319}	r_{1419}	r_{1519}	r_{1619}	r_{1719}	r_{1819}	1

Figure 10.1. Illustration of how the matrix of correlations between participants was formatted. Cells in white around the diagonal indicate pairs of participants from the same group (i.e., intragroup pairs). Cells in gray, further away from the diagonal, indicate pairs of participants from different groups (i.e., intergroup pairs). (Illustration is representative of the Relationship Judgment Task that 19 pilots completed).

Assessing Intragroup Homogeneity

If pilots with similar total hours flown have similar knowledge structures, then it would be expected that correlations between pilots within the same Experience Group would be generally higher than correlations between pilots who are not in the same Experience Group. The term "intragroup homogeneity" refers to the situation when correlations are higher between members of the same group than between members of different groups (Valero & Sanmartin, 1999).

One drawback to using correlation coefficients as an index of similarity to examine intragroup homogeneity is that the correlations between each pair of pilots are non-independent. To meet the assumption of statistical independence, the occurrence of one event within a dataset cannot make the occurrence of another event more or less probable. In other words, changing the value of one event should not affect the values of the other events in the dataset.

Since the current analysis uses correlation coefficients between all pairs of pilots, if the responses of one pilot are altered, the correlation coefficients of all pairs that involve that pilot would also be altered. Thus, matrix data for this analysis are not independent and therefore, traditional parametric tests are inappropriate for use, although the issue of non-independence has been frequently overlooked in previous research (e.g., Cooke & Schvaneveldt, 1988; Fiore et al, 2000). Valero & Sanmartin (1999) suggest the use of the Mantel Test to examine intragroup homogeneity of conceptual structures when correlation coefficients are used as indices of similarity between participants.

The Mantel Test (Mantel, 1967; Legandre & Legandre, 1998; Sokal & Rohlf, 1995) is a statistical test designed to assess the similarity between two matrices by using sampled randomization techniques to test whether the association between the matrices is stronger than what would be expected by chance. It is frequently used in the domain of Ecology because, unlike conventional statistical analyses, the Mantel Test does not assume independence of samples and frequently the ecology samples that are being compared are non-independent. Another advantage of the Mantel Test is that it can be applied to different types of data (e.g., categorical, rank, interval-scale), as long as that data can be transposed into a distance measure (e.g., dissimilarity matrix).

There are several different forms of the Mantel Test but the form most relevant for the current research study is its use as a nonparametric equivalent of an analysis of variance (Sokol & Rohlf, 1995; Hubert, Golledge, & Costanzo, 1982). The Mantel Test can be used as a formal hypothesis test by assessing the strength of the relationship between an observed similarity matrix and a matrix posed by a model or hypothesis. Because the objective was to assess whether conceptual structures of pilots within the same Pilot Experience Group are more similar than are conceptual structures of pilots in different groups, the model matrix consists of 1's and 0's, with 1's placed in cells corresponding to the correlations between pairs of pilots within the

same Experience Group (i.e., intragroup pairs) and 0's placed in cells corresponding to correlations between pairs of pilots in different Experience Groups (i.e., intergroup pairs) (see Figure 10.2). The null hypothesis for any Mantel Test is that there is no association between the elements in the observed and model matrices (Sokal & Rohlf, 1995). A rejection of the null would indicate that the correlations between pilots within the same group are higher and more frequently associated with the 1's in the model matrix than the correlations between pilots within different groups.

	Low-Time Pilots						Mid-Time Pilots						High-Time Pilots						
	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15	P16	P17	P18	P19
P1	1																		
P2	1	1																	
P3	1	1	1																
P4	1	1	1	1															
P5	1	1	1	1	1														
P6	1	1	1	1	1	1													
P7	0	0	0	0	0	0	1												
P8	0	0	0	0	0	0	1	1											
P9	0	0	0	0	0	0	1	1	1										
P10	0	0	0	0	0	0	1	1	1	1									
P11	0	0	0	0	0	0	1	1	1	1	1								
P12	0	0	0	0	0	0	1	1	1	1	1	1							
P13	0	0	0	0	0	0	0	0	0	0	0	0	1						
P14	0	0	0	0	0	0	0	0	0	0	0	0	1	1					
P15	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1				
P16	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1			
P17	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1		
P18	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	
P19	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1

Figure 10.2. Illustration of the model matrix to test for intragroup homogeneity.

Mantel Test Calculations

The Mantel Test involves the calculation of a z test value (z_M) that is based on the cross-product of values within the two matrices that are being compared:

$$z_M = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} x_{ij}\, y_{ij} \qquad (2)$$

where x_{ij} is the correlation between a given pair of pilots and y_{ij} is the corresponding value in the model matrix ("1" if the pilots are in the same experience group or "0" if the pilots are in different groups). High values of z_M indicate a high correspondence between 1s in the model matrix and high correlation (*r*) coefficients between pairs of pilots in the observed matrix.

The z_M value is typically then normalized using the following formula:

$$r_M = \frac{1}{(d\text{-}1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{(x_{ij} - \bar{x})}{S_x} - \frac{(y_{ij} - \bar{y})}{S_y} \qquad (3)$$

where x and y are the average of the observed (X) and model (Y) matrices, respectively, S_x and S_y are standard deviations of the X and Y matrices, and $d = [n(n\text{-}1)/2]$ is the number of distances in the half matrix (lower triangular part of each matrix) (Legendre & Legendre, 1998).

Once r_M is calculated, the next step is to conduct the equivalent of an analysis of variance over the matrices. However, because the distribution of r_M is unknown, the hypothesis test is conducted through permutation procedures. The rows and columns of one of the matrices are randomly rearranged and the r_M values are recalculated thousands of times to generate a distribution of r_M values. It does not matter which matrix gets rearranged. Figure 10.3 provides a schematic illustration of a distance matrix (a) and a random permutation of its rows and columns (b).

a. Distances

	p1	p2	p3	p4	p5
p1	1				
p2	0.20	1			
p3	0.41	0.39	1		
p4	0.12	0.25	0.53	1	
p5	0.13	0.14	0.45	0.17	1

b. Random permutation of rows and columns

	p2	p1	p5	p4	p3
p2	1				
p1	0.20	1			
p5	0.14	0.13	1		
p4	0.25	0.12	0.17	1	
p3	0.39	0.41	0.45	0.53	1

Figure 10.3. A schematic distance matrix (a) and an example of a random permutation of its rows and columns (b). Schematic representation adapted from Sokol & Rohlf (1995, p.817).

The significance is specified by the proportion of times the permuted r_M is above or below the original r_M value. The probability (p) of the observed value is calculated using the formula:

$$p = \frac{(n_T + 1)}{(N + 1)} \tag{4}$$

where n_T is the number of randomized r_M values equal to or above (or equal to or below) the observed value of r_M . If $p < .05$, then the observed and model matrix are correlated, meaning that the probability of the observed r_M is statistically larger than what we would normally observe through chance. XLSTAT was used to perform the Mantel Test analyses.[12]

All p values were calculated using a distribution of r_M values estimated from 10,000 permutations and are evaluated at a significance level of .05. All analyses were also conducted with 1000 permutations (recommended to be evaluated at an alpha level of .05) with no change in the results. Thus, all reported analyses are from the 10,000 permutation analysis.

As Figure 10.4 and Figure 10.5 show, average correlations between pilots were highest in the *Relationship Judgment Task***,** regardless of intergroup or intragroup pairings. Intergroup

[12] XLSTAT demonstrations and tutorials for the Mantel Test suggest the use of a full matrix to conduct the analysis, even when the data is symmetric as is the case in the current research study. Thus, all analyses were actually conducted on full matrices.

and intragroup correlations were all around *r* = .5. The Mantel Test confirmed that intragroup correspondence was not greater than intergroup correspondence, meaning that there was no homogeneity within the Pilot Experience Groups (r_M =.06, p=.45). In general, pilots from the same Experience Group showed no greater correlation with each other than did pilots from different Experience Groups with respect to their Relationship Judgments. This finding suggests that the Relationship Judgment may tap into more general aviation knowledge that is unaffected or uninfluenced by the level of pilot experience. Note that there seems to be more variability in the correlations between pairs of Mid-Time Pilots (MT/MT) than between pairs of Low-Time Pilots (LT/LT) or High-Time (HT/HT) pilots (i.e., the 95% confidence interval is larger for MT/MT than the LT/LT or HT/HT in Figure 10.4). Again, this finding is consistent with skill acquisition research which suggests that the transition between declarative and procedural knowledge that Mid-time pilots are in the process of making results in higher variability in their knowledge structures.

Average correlations between pilots were very low for the *Prime Recognition Task* response times regardless of intergroup or intragroup pairings (all were very near *r* = .1). The Mantel Test revealed no homogeneity within the Pilot Experience Groups (r_M =-.001, p=.973), meaning that pilots from the same Experience Group showed no greater correlation with each other than did pilots from different Experience Groups on their Prime Recognition Task data. Thus, pilots showed a general lack of similarity in their response times for prime-target pairs, regardless of whether they shared similar levels of experience.

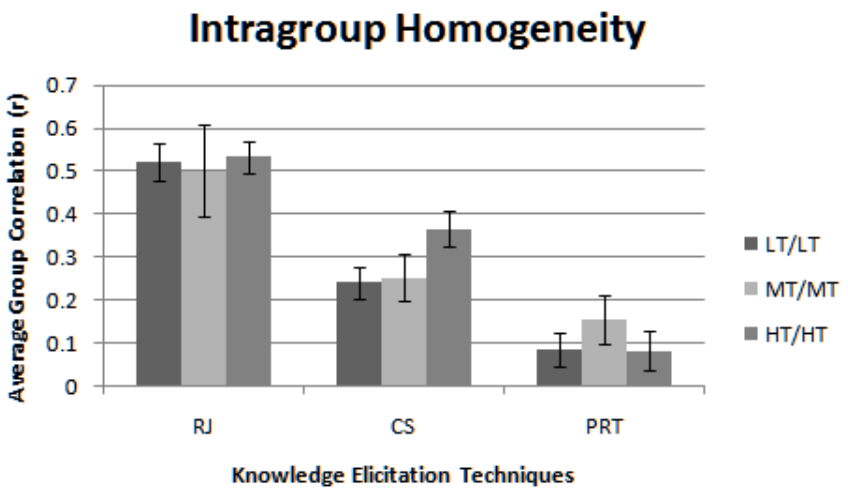


Figure 10.4. Average Group Correlations for intragroup pilot pairings across the three KETs (LT/LT=Low-Time pilots paired with other Low-Time pilots; MT/MT = Mid-Time pilots paired with other Mid-Time pilots; HT/HT =High-Time pilots paired with other High-Time pilots). Means are depicted with 95% confidence intervals.

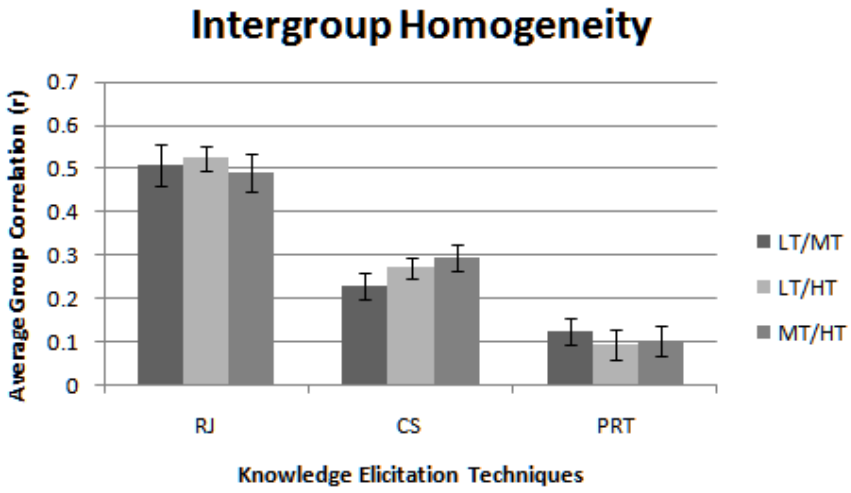


Figure 10.5. Average Group Correlations for intergroup pilot pairings across the three KETs (LT/MT=Low-Time pilots paired with Mid-Time pilots; LT/MT = Low-Time pilots paired with Mid-Time pilots; MT/HT =Mid-Time pilots paired with High-Time pilots). Means are depicted with 95% confidence intervals.

Overall, correlations between pilots based on their Card Sort scores were higher than correlations based on their Prime Recognition Task data but not as high as the correlations based on the Relationship Judgment task. The Mantel Test revealed that intragroup correspondence was slightly greater than intergroup correspondence in the Card Sort data, meaning that there was a marginally significant homogeneity within the Pilot Experience Groups (r_M =.07, *p*=.06). This result implies that, in general, Card Sorts by pilots from the same Pilot Experience Group show slightly greater correlation with each other than do Card Sorts by pilots from different Pilot Experience Groups.

To identify which of the three Pilot Experience Groups account for the intragroup homogeneity in the Card Sort, subsequent Mantel Tests were conducted on pairwise comparisons between each of the three Pilot Experience Groups. Thus, three different observed matrices were created – one comparing Low-Time to Mid-Time pilots, one comparing Low-Time to High-Time pilots, and one comparing Mid-Time to High-Time pilots. Figure 10.6 shows an example of an observed matrix comparing Low-Time pilots to High-Time pilots. The model matrix (Figure 10.7) was again constructed such that cells corresponding to pairs of pilots originating from different Experience Groups were given 0's and cells corresponding to pairs of pilots from the same Experience Groups were given 1's (Legendre & Legendre, 1998).

	Low-Time Pilots						High-Time Pilots						
	P1	P2	P3	P4	P5	P6	P13	P14	P15	P16	P17	P18	P19
P1	1												
P2	r_{12}	1											
P3	r_{13}	r_{23}	1										
P4	r_{14}	r_{24}	r_{34}	1									
P5	r_{15}	r_{25}	r_{35}	r_{45}	1								
P6	r_{16}	r_{26}	r_{36}	r_{46}	r_{56}	1							
P13	r_{113}	r_{213}	r_{313}	r_{413}	r_{513}	r_{613}	1						
P14	r_{114}	r_{214}	r_{314}	r_{414}	r_{514}	r_{614}	r_{1314}	1					
P15	r_{115}	r_{215}	r_{315}	r_{415}	r_{515}	r_{615}	r_{1315}	r_{1415}	1				
P16	r_{116}	r_{216}	r_{316}	r_{416}	r_{516}	r_{616}	r_{1316}	r_{1416}	r_{1516}	1			
P17	r_{117}	r_{217}	r_{317}	r_{417}	r_{517}	r_{617}	r_{1317}	r_{1417}	r_{1517}	r_{1617}	1		
P18	r_{118}	r_{218}	r_{318}	r_{418}	r_{518}	r_{618}	r_{1318}	r_{1418}	r_{1518}	r_{1618}	r_{1718}	1	
P19	r_{119}	r_{219}	r_{319}	r_{419}	r_{519}	r_{619}	r_{1319}	r_{1419}	r_{1519}	r_{1619}	r_{1719}	r_{1819}	1

Figure 10.6. Illustration of how the matrix of correlations between pilots was formatted when conducting the Mantel Tests to compare between groups (a comparison between Low-Time and High-Time pilots is depicted).

	Low-Time Pilots						High-Time Pilots						
	P1	P2	P3	P4	P5	P6	P13	P14	P15	P16	P17	P18	P19
P1	1												
P2	1	1											
P3	1	1	1										
P4	1	1	1	1									
P5	1	1	1	1	1								
P6	1	1	1	1	1	1							
P13	0	0	0	0	0	0	1						
P14	0	0	0	0	0	0	1	1					
P15	0	0	0	0	0	0	1	1	1				
P16	0	0	0	0	0	0	1	1	1	1			
P17	0	0	0	0	0	0	1	1	1	1	1		
P18	0	0	0	0	0	0	1	1	1	1	1	1	
P19	0	0	0	0	0	0	1	1	1	1	1	1	1

Figure 10.7. Illustration of model matrix to test for comparisons between Pilot Experience Groups.

Table 10.2 provides a summary of the individual Mantel Tests for each pairwise comparison between the three Pilot Experience Groups on the Card Sort scores as well as the average correlations between and within each Pilot Experience Group. Note that the average

correlations between pilots within the same Experience Group (i.e., r_{11}, r_{22}, r_{33}) and between pilots in different Experience Groups (i.e., r_{12}, r_{13}, r_{23}) shown in Table 10.2 are also represented graphically in Figure 10.4 and Figure 10.5 above. High-Time pilots showed the highest average correlations (r_{33} = .37).

As Table 10.2 indicates, the Mantel Tests indicated a significant result between High-Time and Low-Time pilots (r_m = .125, p=.03), thus implying that there are differences in conceptual structures between High-Time pilots and Low-Time pilots[13]. In other words, Low-time pilots and High-time pilots had higher correlations when paired with pilots of similar experience level (LT/LT and HT/HT) than when paired with pilots of different experience level (LT/HT). This result implies that Low-time pilots and High-time pilots have least similarity in their conceptual structures.

When taken together, the Mantel Test results and the average correlations between and within the Pilot Experience Groups suggest that the Card Sort technique may be tapping into a part of the knowledge structure that is affected or shaped by pilot experience. High-Time pilots have a greater intragroup homogeneity with respect to their Card Sort data (r_{33}=.37) than Low-Time pilots (r_{11}=.24) and Mid-Time pilots (r_{22}=.25), even though the range of total hours flown was much larger for High-Time pilot group than it was for the Low-Time and Mid-Time pilot groups. Further, the Mantel Test results suggest that the conceptual structures of High-Time pilots are incompatible with the conceptual structures of Low-Time pilots when elicited by Card Sort. Mid-Time pilots seem to share some similarity in conceptual structure with both Low-Time Pilots and High-Time pilots, as Mantel Test results were not significant for any group comparison involving Mid-Time pilots.

[13] Note that r_m=.13 is not a Pearson correlation coefficient and therefore does not imply that the relationship between High-Time and Low-time pilots explains 1.7% of the variance. As the explanation on Mantel Test calculation indicates, r_m is just the normalization of the cross product between the observed and model matrices.

It is somewhat surprising to find that Mid-Time pilots show higher average correlations when paired with High-time pilots (r_{23} = .30) compared to when they are paired with other Mid-Time pilots (r_{22} = .25). Also, Low-time pilots show higher average correlations when paired with High-time pilots (r_{13}=.27) compared to when they are paired with other Low-Time pilots (r_{11} = .24). However, according to the Mantel Test, these average correlations between groups were not significantly different from each other.

Table 10.2. *Mean correlations for the Pilot Experience Groups (intragroups) and between members of different groups for each of the group comparisons conducted (intergroups). Also included are results from the three Mantel Tests performing pairwise comparisons between the three Pilot Experience Groups (** indicates a significant result, p≤.05).*

	Low-Time Pilots	Mid-Time Pilots	High-Time Pilots
Low-Time Pilots	r_{11}=.24		
Mid-Time Pilots	r_M 0.048 p-value (Two-tailed) 0.429 alpha 0.05 r_{12}=.23	r_{22}=.25	
High-time Pilots	r_M 0.125 p-value (Two-tailed) **0.027** ** alpha 0.05 r_{13}=.27	r_M 0.054 p-value (Two-tailed) 0.332 alpha 0.05 r_{23}=.30	r_{33} = .37

Assessing the Similarity between KETs using Within Subjects Comparisons

Several pilots participated in two KETs, Card Sort and either Prime Recognition Task or Relationship Judgment. The Mantel Test was used to see whether pairs of pilots who were highly correlated in one KET were also highly correlated in another KET. In other words, instead of using the Mantel Test to evaluate the relationship between an observed matrix of correlations between pilots and a model matrix, the Mantel Test was used to assess the similarity between a

matrix of correlations based on pilots' Card Sort data and a matrix of correlations for the same pilots on either their Relationship Judgment or Prime Recognition Task data.

Seventeen pilots participated in both the Card Sort and the Prime Recognition Task. Correlations between each pair of those 17 pilots on their Card Sort data were arranged into a matrix. Pilots' pairwise correlations on their Prime Recognition Task data were arranged into another matrix. The Mantel Test was used to test the strength of the association between these two data matrices. Mantel Test found no homogeneity between pilots correlated in the Prime Recognition Task and Card Sort (r_M =-.08, p=.37). In other words, the Mantel Test found no relationship between the Prime Recognition Task and Card Sort matrices, implying that pilots who were highly correlated based on their Prime Recognition Task data were not necessarily highly correlated based on their Card Sort data.

An additional 17 pilots participated in both the Card Sort and the Relationship Judgment Task. Correlations between each pair of those 17 pilots on their Card Sort data were arranged in one matrix and their correlations based on their Relationship Judgment data were arranged in a second matrix. The Mantel Test was again used to test the strength of the association between these two data matrices. Mantel Test did find homogeneity between pilots correlated in the Card Sort and pilots correlated in the Relationship Judgment matrices (r_M =.20, p=.02). In other words, the Mantel Test found a relationship between the Card Sort and Relationship Judgment matrices, meaning that pilots who were highly correlated based on their Card Sort data were also highly correlated based on their Relationship Judgment data.

Summary

Relationship Judgment and Card Sort techniques elicit proximity data that are highly related, both in terms of the raw proximity data and with respect to individual pilot performance. This suggests that the Relationship Judgment and Card Sort Techniques tap into a similar type of knowledge structure relative to the Prime Recognition Task, as hypothesized. Card Sort was

the only technique where conceptual structures were differentiated by pilot experience (as indicated by the Mantel Test). Conceptual structures elicited by Relationship Judgment and by Prime Recognition Task seemed relatively uninfluenced by different levels of pilot experience, as defined by total number of hours flown. Taken together, these results suggest that each of the three techniques may be tapping into a different type of knowledge, knowledge that for at least two of the KETs may not be as influenced or affected by the total number of hours flown as Card Sort was.

The Mantel Test was based on calculating Pearson correlation coefficients between each pair of pilots to see whether the pattern of responses across the 105 items was the same. However, correlation coefficients ignore information about the actual magnitude of the scores. They do not provide any information about the actual structure of the weather items in memory. They also do not provide any information on the factors that pilots used to assess the relationship between items. Correlations only specify the extent to which the response patterns themselves are similar. Thus, these initial analyses provide some insight into the similarities in knowledge structure across Pilot Experience Groups but there is still more to be learned. The next chapter describes analyses designed to provide insight into the features or characteristics that influence how knowledge is structured in memory. Also, the chapter describes attempts to assess the validity of each of the KETs.

Chapter 11 - Dimensions Underlying Pilots' Knowledge Structures for Weather Information (Phase III)

The second major objective of this research was to understand more about pilots' knowledge structures for weather information and how the elicited conceptual structure may be affected by pilot experience and type of KET. Multidimensional scaling (MDS) is one of the most commonly used methods for extracting the latent structure from the proximity data and representing that structure in a spatial form. The application of MDS to the current research study had two purposes. First, MDS was used to create spatial representations of conceptual structures that provide insight into how weather information is structured in pilots' memory. Spatial representations were created based on data from all pilots and also for each pilot experience group to see if there were any differences in spatial layout as a function of experience. Those results are reviewed in the current chapter. Second, the MDS results were used to help provide validation of each of the data collection methods as a knowledge elicitation technique. Validation occurred by examining how well each KET conceptual structure could be used to 1) discriminate among pilot experience groups and 2) predict pilot experience group membership. The validation results are reviewed in Chapter 12.

Procedure for Analysis

Proximity data collected from each KET were formatted and placed into 15 x 15 half matrices where each of the 15 weather concepts was crossed with all other concepts. One data matrix was created for each participant. These individual matrices were submitted to MDS ALSCAL individual differences scaling procedure (i.e., Weighted MDS or WMDS). Individual ALSCAL WMDS analyses were conducted on data from each KET separately. Nonmetric WMDS scaling solutions were developed from two to six dimensions for matrix conditional data. The convergence criterion was set to .001, the maximum number of iterations was set to 30 and

the minimum s-stress value was set to .005, all default settings for SPSS 13. Table 11.1 provides a summary of the characteristics of the data collected in each KET and the parameters used for the ALSCAL WMDS analysis (see Appendix D for more explanation about the data characteristics and analysis decisions).

Table 11.1. *Characteristics of the Data Collected in each KET.*

	Knowledge Elicitation Technique (KET)		
	Relationship Judgment (RJ)	*Card Sort (CS)*	*Prime Recognition Task (PRT)*
Data Characteristics:			
Type of proximity data	Dissimilarity (Judgments on 1-9 scale)	Dissimilarity (Jaccard scores, range 0-1)	Response Times (longer = less similar)
Data matrix shapes	square	square	square
Number of ways	three-way	three-way	three-way
Presence of missing data	Yes	no	yes
# of judgments per stimulus pair (recommended > 9)	Total: 19 Low-time pilots: 6 Mid-time pilots: 5 High-time pilots: 8	Total: 38 Low-time pilots: 12 Mid-time pilots: 12 High-time pilots: 14	Total: 24 Low-time pilots: 9 Mid-time pilots: 8 High-time pilots: 7
Analysis Decisions:			
MDS Procedure	non-metric (ordinal)	non-metric (ordinal)	non-metric (ordinal)
Conditionality	matrix-conditional	matrix-conditional	matrix conditional
MDS model	Weighted MDS	Weighted MDS	Weighted MDS
Approach to ties in data	untie	untie	untie

ALSCAL WMDS provides a spatial representation of the relationships between each of the weather-related concepts. That spatial representation is called a *stimulus space*. Fenker (1975) identified three properties of stimulus spaces:

1) Characteristics or features of the domain of interest are represented by a set of dimensions, and information about the domain of interest (represented as concepts or items) is organized and interpreted on the basis of those dimensions.
2) The dimensions can be represented in n-dimensional space.

3) There are many relationships between concepts and the stimulus space identifies or depicts a relationship. However, the nature behind how the concepts are related (i.e., the meaning of the underlying dimensions) has to be interpreted.

Thus, the dimensions in the stimulus space reveal the underlying features or characteristics used by pilots to make judgments of similarity between the pairs of weather-related concepts. The placement of each concept in the stimulus space is based on data from all pilots included in the analysis. ALSCAL WMDS also provides a *participant space* that consists of points along the same dimensions as the stimulus space, but this time the coordinates of each point represent the relative salience of each dimension to the pilot each point represents. In sum, the stimulus space provides insight into how knowledge is structured in memory (i.e., through identification of an overall conceptual structure based on data aggregated across pilots) and the participant space provides insight into how individual pilots value the underlying dimensions of that aggregated conceptual structure.

Background on Procedures for MDS Interpretation

The following sections provide a brief description of the procedures for interpreting MDS analyses and representations. See Appendix D for more explanation on each of these procedures.

Identifying Optimal Dimensionality

Several criteria can be used to identify optimal dimensionality but ultimately the decision is left up to the researcher. The term *optimal dimensionality* refers to the number of dimensions that provides the best model fit to the proximity data while also lending itself to meaningful interpretation. Ultimately, the determination of optimal dimensionality is subjective. Measures of fit such as stress and variance accounted for (R^2) can be calculated to evaluate the fit of the model solution to the original proximity data. *Stress* is defined as the square root of a

normalized “residual sum of squares” and it is the measure that the computer programs attempt to minimize through the iterative procedure in which the model configuration is modified step-by-step to increase its correspondence with the original proximity data. Stress values greater than .20 are widely considered to indicate poor model fit (Kruskal & Wish, 1978). R^2, also referred to as the *squared correlations*, is an indicator of the proportion of variance of the disparities accounted for by the MDS model. Thus higher numbers of R^2 indicate a better fit of the model to the data (George & Mallery, 2009; Schiffman, Reynolds, & Young, 1981).

A higher dimensional solution generally increases the fit of the model. However, higher dimensional solutions are not necessarily desirable because of the difficulty in interpreting such configurations. Thus, the reliance on measures of fit is tempered by considerations for ease of use and ease of interpretability when deciding which dimensional solution is optimal. Also, the MDS literature suggests the following guideline for appropriate dimensionality (D) given the number of stimuli used (I): $I - 1 > 4D$. Generally, the number of stimuli (I) minus one should be at least four times as great as the dimensionality, D (Kruskal & Wish, 1978). The current study used 15 stimuli and thus this guideline would suggest that the optimal solution should not exceed 3 dimensions.

Interpreting Dimensionality

While MDS provides a systematic procedure for creating a graphical representation of the underlying relationships in the proximity data, the process of interpreting the meaning of the dimensions that specify the relationships is less systematic. One way of identifying meaningful features or characteristics that define dimensions is to look at the orderings or groupings of the stimuli along the dimensions separately, based on their coordinates. Figuring out what distinguishes the items on the extremes of the dimensions should help in interpreting the dimension meaning. Often, subject matter experts (SMEs) are relied upon to help provide this understanding of the dimensions.

Of course, in some cases, it may be impossible to identify the meaning of a dimension. In other words, items may be configured along a dimension but there is not enough information to be able to identify exactly what feature or characteristic that dimension actually represents. Failure to identify or interpret a dimension does not necessarily invalidate the MDS analysis (e.g., Schiffman et al., 1981; Schvaneveldt et al., 1985).

Representing Conceptual Structure

Procedure

MDS was used at two different levels to uncover latent structure in the proximity data elicited by each KET. First, MDS was used to uncover how the 15 weather-related concepts were organized in memory based on data from all pilots without regard for level of experience. This analysis provided insight into the *general* conceptual structure of weather information and how that general structure may differ as a function of the KET. Second, MDS was used to examine how the organization of those weather concepts differed among the Pilot Experience Groups.

ALSCAL WMDS analyses were conducted using SPSS 13 on data from each of the KETs. For each KET, ALSCAL WMDS analysis was applied to the entire dataset (collapsed across Pilot Experience Group). Conceptual structures and participant spaces were constructed for WMDS solutions ranging from two to six dimensions. Then, individual ALSCAL WMDS analyses were conducted on data sets specific to each Pilot Experience Group. Conceptual structures and participant spaces were constructed for WMDS solutions ranging from two to six dimensions for each Pilot Experience Group. Thus, a total of 12 WMDS analyses were conducted – four analyses based on data (Overall, Low-Time only, Mid-Time only, High-Time only) from each of the three KETs.

Assessing Complexity of Conceptual Structure

Table 11.2 provides the Stress-1 and R^2 values as a function of increasing dimensionality of the WMDS solutions for each KET collapsed across Pilot Experience Group. For each of the three KETs, Stress-1 and R^2 improved with increasing dimensionality. Scree plots were constructed but as is often the case, there were no clear "elbows" apparent to help in identifying optimal dimensionality for any of the KETs. Although the six-dimensional configuration resulted in the best Stress-1 value for all three KETs, a six-dimensional solution is not interpretable and therefore is of no practical use for this study. Stress-1 levels were less than .20 (indicating at least a fair model fit) for the 3D solution in Relationship Judgment, the 2D solution in Card Sort, and the 4D solution in the Prime Recognition Task. Note, too, that the R^2 values for the threshold at which Stress-1 indicates fair model fit are higher for Card Sort (R^2 = .85) compared to Relationship Judgment (R^2 = .74) and much higher compared to Prime Recognition Task (R^2 = .21). Thus, the Card Sort required the least complex solution to achieve at least a fair model fit to the overall dataset.

Table 11.2. *Stress-1 and variance accounted for (R^2) for each WMDS solution based on two- to six-dimensions for Relationship Judgment, Card Sort, and Prime Recognition Task data. Dimension at which MDS identified a "Fair" model fit (Stress-1 ≤ .20) for each KET dataset is shaded in gray and bolded.*

Dimensions (D) in solution	**Relationship Judgment (19 pilots)**		**Card Sort (38 pilots)**		**Prime Recognition Task (24 pilots)**	
	Stress-1	R^2	*Stress-1*	R^2	*Stress-1*	R^2
2	.22	.70	**.18**	.85	.36	.16
3	**.17**	.74	.14	.88	.26	.21
4	.13	.76	.11	.87	**.20**	.21
5	.11	.79	.09	.89	.16	.23
6	.10	.80	.08	.91	.13	.27

Table 11.3 provides the Stress-1 and R^2 values as a function of increasing dimensionality of the WMDS solutions for each KET x Pilot Experience Group. Again, Stress-1 and R2 improved with increasing dimensionality for each KET x Pilot Experience Group. Scree

plots were constructed but, as was the case for the analysis on data collapsed across experience group, there were no clear "elbows" apparent to help in identifying optimal dimensionality for any of the KET x Pilot Experience Group. Not surprisingly, the six-dimensional solution resulted in the highest Stress-1 and R^2 values for nine analyses but it is uninterpretable. Fair model fits (*Stress-1* $\leq$.20) were found in the 3D solution for Relationship Judgment, 2D solution for Card Sort and 4D solution for Prime Recognition Task. The dimensions at which the models indicated fair fit were consistent across the three Pilot Experience Groups for each KET. Again, Card Sort required the least complex (fewest dimension) solution to achieve at least a fair model fit, and the High-Time pilots required the least complex solution with the highest R^2 value (R^2 = .91) of all KET x Pilot Experience Groups.

Table 11.3. *Stress-1 and R^2 values as a function of increasing dimensionality of the solution space for each KET and each pilot experience group: a) Low-Time pilots, b) Mid-Time pilots, c) High-Time pilots. Dimension at which MDS identified a "Fair" model fit (Stress-1 $\leq$ .20) for each KET dataset is shaded in gray and bolded.*

a) Low-Time Pilots

Dimensions (D) in solution	Relationship Judgment (6 pilots)		Card Sort (12 pilots)		Prime Recognition Task (9 pilots)	
	Stress-1	R^2	*Stress-1*	R^2	*Stress-1*	R^2
2	.21	.76	**.15**	.89	.36	.14
3	**.15**	.79	.12	.91	.26	.20
4	.12	.82	.08	.94	**.20**	.25
5	.10	.84	.07	.94	.16	.27
6	.08	.86	.06	.96	.13	.31

b) Mid-Time Pilots

Dimensions (D) in solution	Relationship Judgment (5 pilots)		Card Sort (12 pilots)		Prime Recognition Task (8 pilots)	
	Stress-1	R^2	*Stress-1*	R^2	*Stress-1*	R^2
2	.22	.73	**.18**	.84	.34	.22
3	**.15**	.80	.14	.87	.25	.25
4	.11	.85	.10	.93	**.18**	.34
5	.08	.89	.07	.95	.15	.36
6	.07	.89	.06	.95	.12	.42

c) High-Time Pilots

Dimensions (D) in solution	Relationship Judgment (8 pilots)		Card Sort (14 pilots)		Prime Recognition Task (7 pilots)	
	Stress-1	R^2	*Stress-1*	R^2	*Stress-1*	R^2
2	.22	.70	**.14**	.91	.37	.17
3	**.17**	.75	.10	.94	.26	.23
4	.14	.76	.08	.96	**.20**	.29
5	.11	.81	.06	.97	.15	.31
6	.09	.83	.05	.97	.14	.31

Identifying Optimal Dimensionality for Interpretation

Based on the following criteria, the two-dimensional solutions for each KET and each KET x Pilot Experience Group were identified as optimal dimensionality for interpretation:

- *Measures of fit:* As discussed in the previous section, Fair model fits occurred in the 3D solutions for Relationship Judgment, 2D solutions for Card Sort, and 4D solutions for Prime Recognition Task. It was determined that the R^2 values did not show a large enough increase between the 2D and 3D solutions to justify the extra complexity of interpreting Relationship Judgment with three dimensions. Also, the R^2 value for Prime Recognition Task was going to be much lower than the other two KETs regardless of what dimensionality was chosen as optimal.
- *Interpretability / Ease of Use:* Convention states that any solutions that involve more than three dimensions are difficult if not impossible for humans to process and evaluate (see Appendix D for more information). Since no other supporting data was collected to help understand dimension meaning (e.g., importance ratings, phase of flight usage, frequency of use, etc.), it was decided that two-dimensions would result in easier interpretability and ease of use for both interpretation and application.

- *Number of stimuli:* Given the number of stimulus items used (15 weather-related information concepts), MDS guidelines suggest that only the two and three-dimensional solutions are appropriate for interpretation.

Interpreting Conceptual Structure

Figures 11.1 – 11.3 show the spatial representations of the conceptual structures based on the 2D WMDS solutions for proximity data elicited from Relationship Judgment, Card Sort, and Prime Recognition Tasks, respectively. The larger graph in the upper left hand corner depicts the conceptual structure based on all pilots' data. Graphs depicting the conceptual structures based on data from the individual Pilot Experience Groups are also included in Figures 11.1 – 11.3. The X and Y axes represent dimensions (i.e., features, characteristics, etc.) that underlie the relationships between the weather information concepts.

Item placement and clustering along the dimensions are the most informative pieces of graphical representation of an MDS analysis. Recall that MDS can distort the local relationships (i.e., the distance between any particular pair), so MDS is much more useful for understanding the global structure among concepts rather than the local structure (Schvaneveldt et al, 1985). For example, in the Relationship Judgment and Card Sort conceptual structures, items like *wind direction* and *wind velocity* are often clustered together and placed on the opposite end of a dimension from items like *freezing level* and *icing*. This graphical representation provides some insight into how knowledge is structured on a global level (*freezing level* and *wind direction* are opposites on some dimension meaning). The fact that *wind direction* and *wind velocity* are placed slightly further apart in the Card Sort conceptual structure compared to the Relationship Judgment conceptual structure is less informative and potentially unreliable in an MDS analysis.

The conceptual structures of Card Sort and Relationship Judgment (Figure 11.1 and Figure 11.2) appear to be more qualitatively similar in layout compared to the Prime Recognition Task (Figure 11.3). For example, while *wind direction* and *wind velocity* are generally clustered

in the Card Sort and Relationship Judgment conceptual structures, they occur on opposite ends from each other in the Prime Recognition Task conceptual structure. *Icing* and *freezing level* also appear to be on opposite ends of a dimension in the Prime Recognition Task conceptual structure, whereas they are generally clustered in Card Sort and Relationship Judgment conceptual structures. The numerous visual differences between the Prime Recognition Task conceptual structure and the Relationship Judgment and Card Sort conceptual structures may be at least partially due to the relative lack of fit of the 2D solution for Prime Recognition Task data (R^2 = .16) compared to the fits of the 2D solutions for Card Sort (R^2 = .85) and Relationship Judgment (R^2 = .70) data. However, note that even the 6D solution for Prime Recognition Task data (R^2 = .27) still provided much less fit than even the 2D solutions for Relationship Judgment and Card Sort.

Three pilots and one FAA Engineering Manager were asked to assist as subject matter experts (SMEs) in interpreting the meaning of the dimensions for the 2D solutions. Demographic and background information for each of the SMEs is provided in Table 11.4 below. These individuals were chosen to be SMEs because of their extensive experience as pilots and instructor pilots and/or because of their work with the FAA in developing guidance for pilot training and policy for new aircraft technology.

SMEs were given a short description of the goal of the study and an overview of the characteristics and demographic information for the pilots who participated in the research. They were shown example output of an MDS analysis not based on any of the concepts used or data collected for this study to use as practice for interpretation of the dimensions. They were told to focus on assessing two factors: 1) dimension meaning, and 2) effect of pilot experience. To help identify dimension meaning, SMEs were told to think about what characteristics distinguish weather-related items on the opposite sides of the dimensions and what features the clusters of items have in common. Regarding the effect of pilot experience, SMEs were asked to

identify any differences they saw in how the items are structures in data from High-Time pilots compared to the less experienced pilots.

Table 11.4. *Demographic information for the Subject Matter Experts (SME) asked to interpret the meaning of the dimensions underlying the 2D conceptual structures resulting from each KET.*

SME	**Total hrs flown**	**Certificates / Ratings Held** ***Job Description***
1	1650	Private Pilot with single and multi-engine land ratings with an instrument rating; Also hold Certified Flight Instructor, Flight Instructor Instrument, and Multi-engine Instructor Ratings *Flight Safety International Simulator and Ground Instructor for the Cessna C208 Caravan. Responsible for taking initial and recurrent pilots through both classroom curricula and the simulator training course*
2	17,700	Airline Transport Pilot with type ratings for: LR-Jet, IA-Jet, HS-125, G-IV, CL-600, BBD-700, DHC-6, B-757/767, single engine land and sea commercial privileges; also hold a Flight Instructor Certificate for single engine, multi-engine and instrument airplanes *FAA Operation's Specialist responsible for developing training, checking and currency requirements for new and retrofitted legacy airplanes.*
3	600	Private Pilot with Instrument Rating and Glider Rating *FAA flight test engineer with responsibilities for rules, guidance, and policy for the flight test issues in Part 23 aircraft. Also flight tests new technology.*
4	n/a	Student Pilot[14] *FAA Engineering Manager (Programs and Procedures) responsible for certification of new technology and the FAA's NextGen air traffic systems as it is implemented for small planes. Also spent 15 years as an FAA flight test engineer.*

Conceptual Structures Elicited by the Relationship Judgment Task

Figure 11.1 depicts the 2D conceptual structures elicited by the Relationship Judgment Task for data collapsed across pilot experience groups and for data from each Pilot Experience Group. For the conceptual structure based on all pilots, the SME pilots agreed that Dimension 1

[14] Note that this person's inclusion as an SME was due to his knowledge and experience as a result of his role as an FAA Engineering Manager responsible for certification of new technology, rather than for his minimal flight experience as a student pilot.

was representative of a “Severity” construct. SME pilots were less confident about the meaning of Dimension 2 but did agree that Dimension 2 seemed to represent some type of “Seasonal” construct or a “Go/No-go Flight Decision” construct which for some pilots, especially Low-Time pilots, may be affected by seasonal conditions. For example, Low-Time pilots may decide not to even fly in the winter because of the possibility of icing and the decreased amount of daylight. Further, SME pilots also perceived that the “Severity” and “Seasonal”/”Flight Decision” dimensions also seemed to interact to form two diagonal “hybrid” dimensions (superimposed on Figure 11.1 with a dotted line). SME pilots named the left-most diagonal “Medium to High Risk” and the right-most diagonal “Low to Medium Risk” indicating the two general types of consequences that could occur for weather item’s Severity x Flight Decision coordinates.

When looking at the conceptual structures at the level of the Pilot Experience Group, the SME pilots found no perceivable or interpretable differences between the experience groups in terms of dimension meaning or in how concepts were clustering. For Low-Time pilots, Dimension 2 more clearly signified a “Go / No-go Flight Decision” characteristic in that items that would most likely affect a Low-Time pilot’s decision to fly were grouped toward the bottom end of the dimension (e.g., *lightning*, *turbulence*, *thunderstorms*) and items that they would need to consult while flying were grouped toward the top end of the dimension (e.g., *wind speed*, *wind direction*, *visibility*). Dimension 1 again signified “Severity.” The general meanings of Dimensions 1 and 2 held fairly consistently throughout the conceptual structures of all three Pilot Experience Groups. This consistency was a surprise to some of the SME pilots who expected that the weather items would cluster differently on the “Go/No-go Flight Decision” dimension for High-Time pilots since their decision to fly would be much less affected by the convective items on the “No-go” end of the dimension (e.g., *lightning, thunderstorms*).

Conceptual Structures Elicited by the Card Sort Task

Figure 11.2 depicts the 2D conceptual structures elicited by the Card Sort Task for data collapsed across Pilot Experience Groups and for data from each Pilot Experience Group. Consistent with Relationship Judgment, for the conceptual structure based on data from all pilots the SME pilots agreed that Dimension 1 was again defined by some type of "Severity" construct. However, a few pilots went on to further describe the dimension as having a "Strategic-Tactical" component in addition to Severity. Weather items like *TAF*, *ceiling*, *sky conditions* and *visibility* are pieces of information that pilots use strategically when planning their flights, whereas information like *turbulence*, *thunderstorms* and *icing* are used on a more tactical basis, specifically in trying to avoid them because they may pose some level of severity. The meaning of Dimension 2 was less clear, but all SME pilots agreed that the dimension represented some type of "Seasonal" construct.

When Card Sort conceptual structures were examined at the level of the individual Pilot Experience Group, Dimension 2 was still loosely defined as "Seasonal" across all experience groups, but the "Severity / Strategic-Tactical" dimension was more clearly defined and seemed to vary at least slightly across experience groups. For the Low-Time pilots, this dimension appeared to function at a level one step higher than "Severity/Strategic-Tactical." One pilot described the dimension more descriptively for Low-Time pilots as a "Level of Importance" dimension. Items on the left of the dimension (*thunderstorms, lightning, icing, freezing level*) would actually be relatively unimportant to Low-Time pilots when flying because they will not or cannot attempt to fly during those conditions. In other words, rather than being strategic, tactical, or even severe, the weather items at that one end of the dimension are just irrelevant because Low-Time pilots will never even encounter them when flying. The items on the right side of the dimension, however, are important for Low-Time pilots to attend to during flight. SME pilots noted that the overall conceptual structures appeared to be quite similar between the Low-

Time and Mid-Time pilots along this dimension, except that the placements of the items were reversed (i.e., “unimportant” items were now on the left side of the dimension for Low-Time pilots and the right side of the dimension for Mid-time pilots).

For the High-Time pilots, this Dimension 1 seemed to maintain its general meaning of “Severity / Strategic – Tactical” but it did vary a bit from the other two groups. In the conceptual structure for High-Time pilots, *turbulence*, *lightning*, and *thunderstorms* are very isolated from the rest of the items. The SME pilots interpreted this dimension to adopt a “Severity for Passengers” emphasis that the Low-time and Mid-Time pilots did not or would not show based on their more limited experience. Most High-Time pilots likely have passengers with them and passengers tend to become concerned or upset by convective weather. Therefore, it is in the High-Time pilot’s best interest to avoid convective weather, not necessarily because they are concerned about their ability to handle it but because they do not want to cause problems or concerns for their passengers.

Conceptual Structures Elicited by the Prime Recognition Task

Not surprisingly, it was impossible for the SME pilots to interpret the meaning of either dimension in the Prime Recognition Task conceptual structure (Figure 11.3). The lack of interpretation was most likely due to the lack of fit of the model to the data. MDS stimulus spaces tend to appear as if the items are placed in circular form when the fit of the model to the data is poor.

Summary

This section described the conceptual structures elicited from the three KETs as represented by the WMDS analysis. The number of dimensions that are needed to fit a solution can be considered an indication of the complexity of the conceptual structure that is needed to represent the inherent knowledge structure. The Card Sort task appeared to elicit the least

complex conceptual structures. Further, the High-Time pilots seemed to be best fit by the least complex conceptual structure, which is consistent with past research that has shown knowledge structures to decrease in complexity as experience increases (e.g. Schvaneveldt et al, 1985). Schvaneveldt and colleagues interpreted their results to imply that expertise is not necessarily synonymous with complex structure. Rather, experts tend to identify the information that is critically associated with their task, which may lead to a simpler representation of that information in knowledge structure. The fact that when dimension (complexity) was held constant, the model fit (i.e., Stress-1) did not increase with experience in Relationship Judgment but did slightly in Card Sort also suggests that Card Sort may be better able to tap into knowledge that is shaped by experience. The next section provides a deeper look into validating each of these techniques.

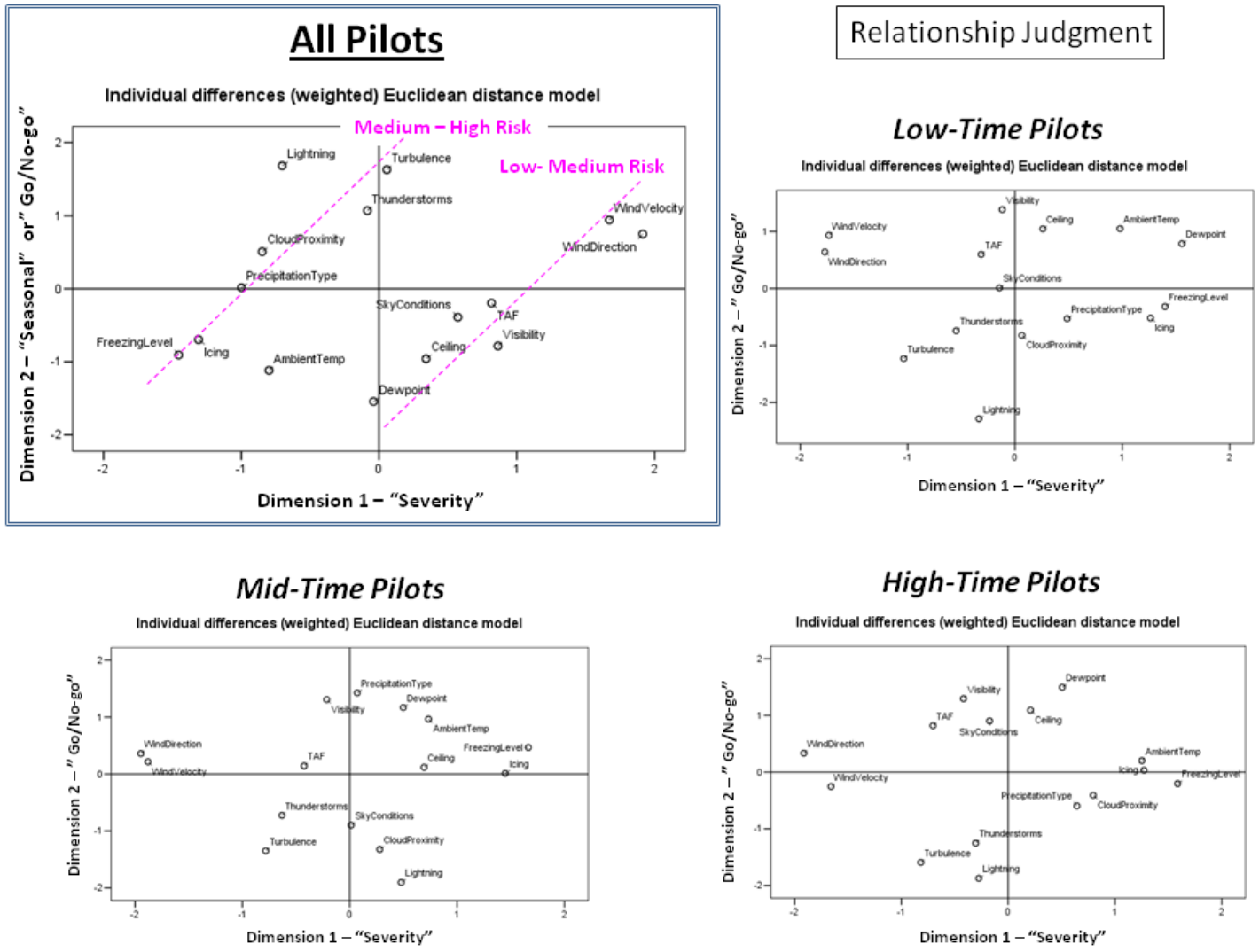


Figure 11.1. Conceptual structures based on 2D WMDS solutions for the Relationship Judgment Task when based on all pilots and when based on pilots within each individual Pilot Experience Group.

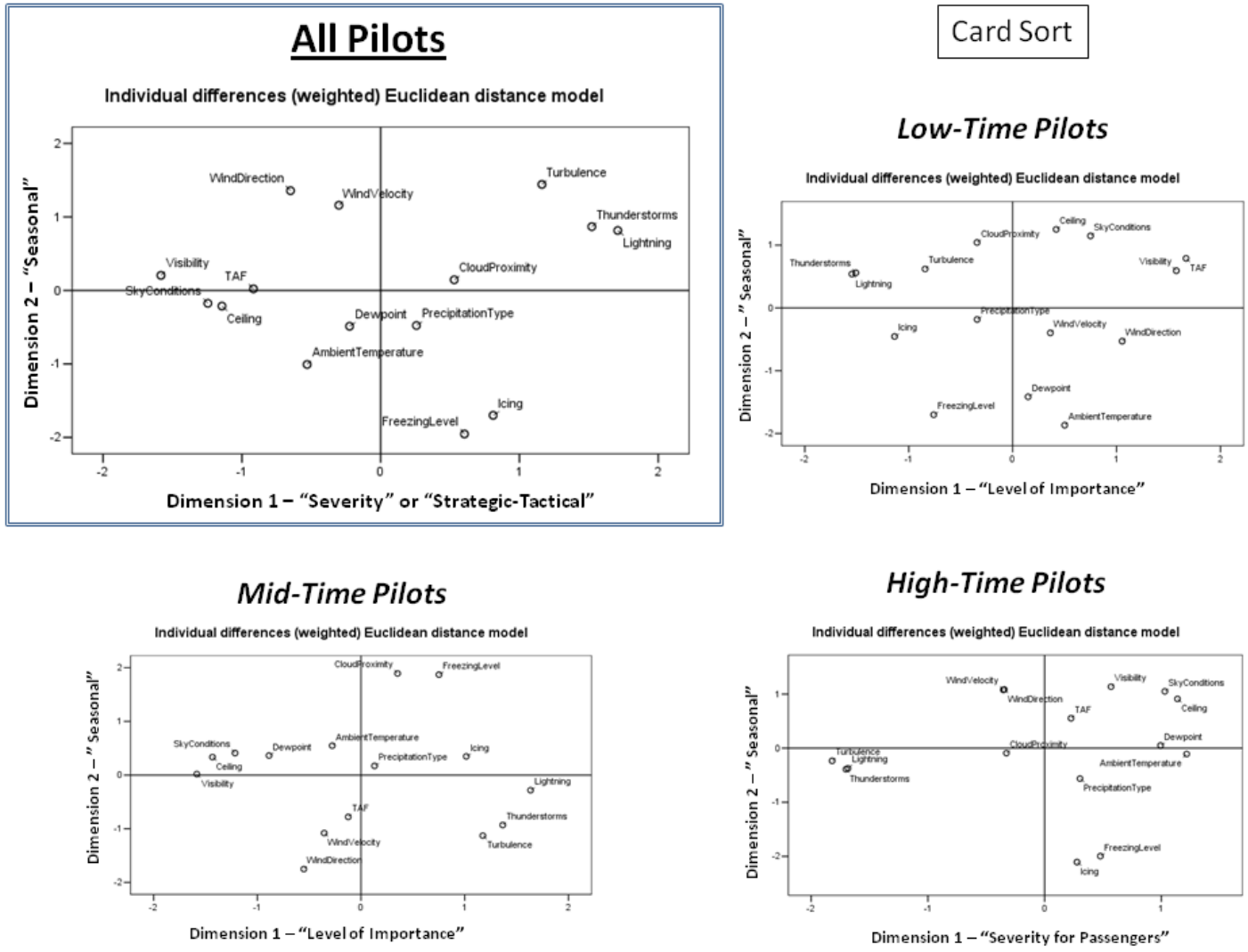


Figure 11.2. Conceptual structures based on 2D WMDS solutions for the Card Sort Task when based on all pilots and when based on pilots within each individual Pilot Experience Group.

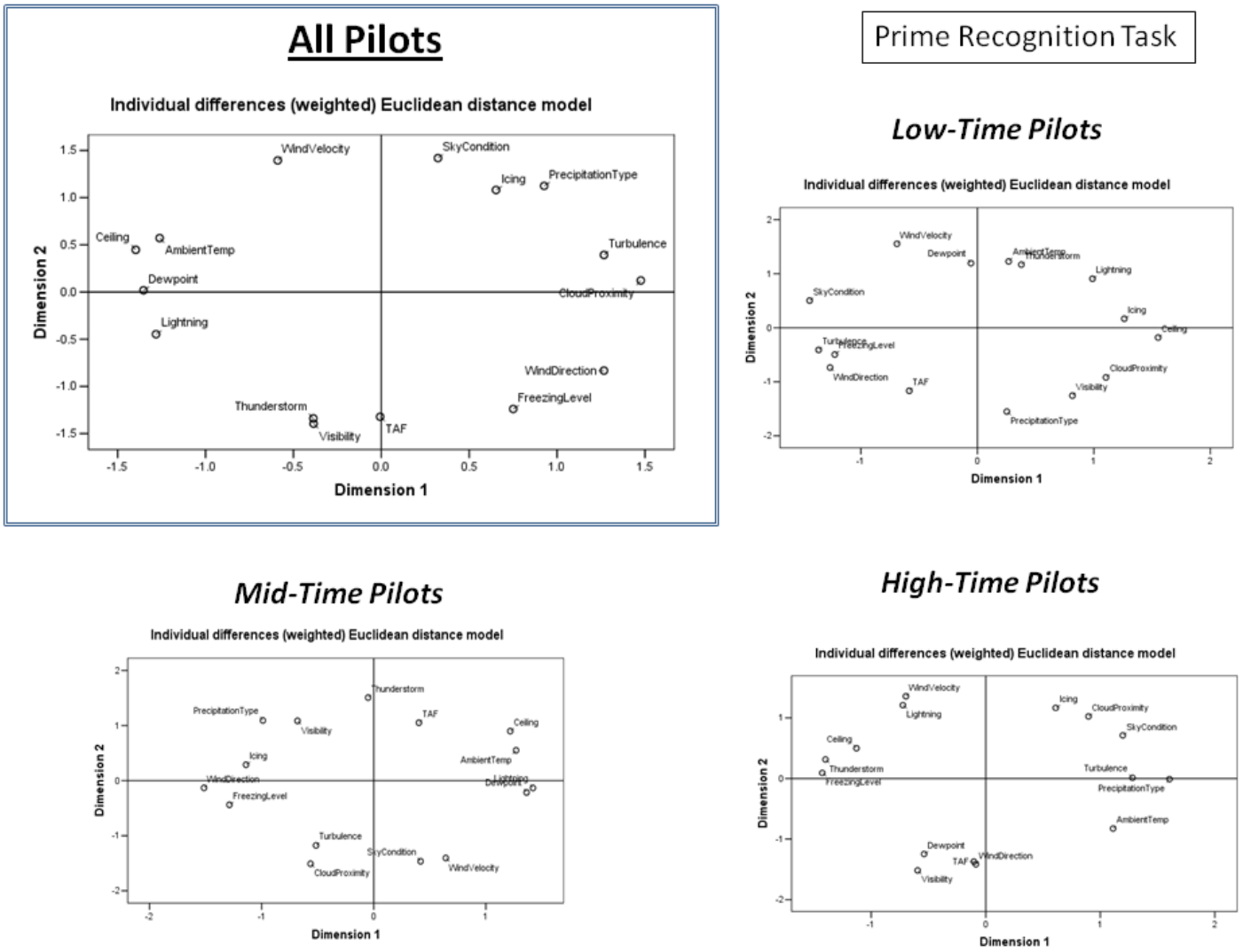


Figure 11.3. Conceptual structures based on 2D WMDS solutions for the Prime Recognition Task when based on all pilots and when based on each individual Pilot Experience Group.

Chapter 12 - Validation of the KETs (Phase III)

Schvaneveldt and colleagues (1985) described and employed a method of validating and comparing conceptual structures that result from different techniques (in their case, they were interested in comparing the validity of analysis techniques – MDS and Pathfinder). This validation method is predicated on the assumption that differences in experience level among pilots should translate into differences in their elicited conceptual structures if the elicitation technique was valid. Further, pilots of similar experience level should share certain characteristics in their conceptual structures that were shaped by common experiences. Thus, if the conceptual structures elicited by a given KET are valid, it should be possible to use those conceptual structures to 1) *discriminate* pilots with less experience from pilots with a lot of experience, and 2) *classify* the experience level of a pilot.

Assessing KET Ability to Discriminate Based on Pilot Experience

Procedure

The first step to assessing validity was to examine which of the KETs (if any) allowed the conceptual structures of pilots with less experience to be *discriminated* from conceptual structures of pilots with a lot of experience. First, proximity matrices of High-Time pilots from each KET were submitted to an ALSCAL WMDS analysis to define the “expert” conceptual structure elicited by each KET. Consistent with the overall conceptual structures discussed in the previous section, when goodness of fit, ease of use, and ease of interpretability were all considered, the 2D solution was determined as optimal for representing the High-Time pilots’ conceptual structure as elicited from each KET. However, because complexity of conceptual structure was also a factor of interest, the ability of a KET to discriminate levels of experience was assessed for both the 2D and the 3D solutions.

In addition to the conceptual structure, recall that WMDS also calculates dimension weights for each pilot included in the analysis. These dimension weights (comprising the participant space) represent the importance each pilot placed on each of the dimensions of the resulting conceptual structure when making their responses during completion of each KET (i.e., how much each dimension influenced their overall judgments, categories, or response times). Dimension weights were on a continuous scale and could range from 0 (meaning the dimension was not at all a factor in the pilots' judgments) to 1 (meaning that the dimension was extremely important in the pilots' judgments). The dimension weights were recorded for each High-Time pilot for the "expert" conceptual structures in both the 2D and 3D solutions.

Next, after defining the "expert" conceptual structure based on data from the High-Time pilots, the proximity matrix from each Low-Time and Mid-Time pilot was then added one at a time to the proximity matrices of the High-Time pilots and the WMDS procedure was rerun repeatedly. The 2D and 3D dimension weights for each Low- and Mid-time pilot were also recorded. Table 12.1 shows the average 2D and 3D weights for each of the Pilot Experience Groups for each KET. Thus, values in Table 12.1 provide an indication of how important each of the dimensions of the "expert" conceptual structure were to the judgments of pilots in all three Pilot Experience Groups. If the dimensions of the expert stimulus space are meaningful, one would expect those dimensions to be more critical to (i.e., weighted higher by) the High-Time pilots compared to the less experienced pilots who may lack the domain experience and knowledge organization characteristics that the more experienced pilots have. Dimensions listed in Table 12.1 include the names they were given (if any) by the SME pilots as discussed in Chapter 11.

Table 12.1. *The average dimension weights for each Pilot Experience Group on the dimensions that defined the "expert" conceptual structure from the a) Relationship Judgment, b) Card Sort, and c) Prime Recognition Tasks. Higher weights indicate greater importance of that dimension in pilots' responses while completing each KET.*

a) Relationship Judgment

2 Dimension Solution	*High-time Pilots*	*Mid-time Pilots*	*Low-time Pilots*	*Mean*
Dim 1 – "Severity"	.69	.73	.67	.70
Dim 2 – "Seasonal" or "Go/No-Go"	.46	.38	.50	.45

3 Dimension Solution	*High-time Pilots*	*Mid-time Pilots*	*Low-time Pilots*	*Mean*
Dim 1 – "Severity"	.64	.67	.59	.63
Dimension 2	.44	.35	.40	.39
Dimension 3	.36	.37	.47	.40

b) Card Sort

2 Dimension Solution	*High-time Pilots*	*Mid-time Pilots*	*Low-time Pilots*	*Mean*
Dim 1 – "Severity"	.68	.80	.74	.74
Dim – 2 "Seasonal"	.54	.31	.35	.40

3 Dimension Solution	*High-time Pilots*	*Mid-time Pilots*	*Low-time Pilots*	*Mean*
Dim 1 – "Severity"	.61	.82	.72	.72
Dimension 2	.48	.17	.26	.30
Dimension 3	.38	.27	.28	.31

c) Prime Recognition Task

2 Dimension Solution	*High-time Pilots*	*Mid-time Pilots*	*Low-time Pilots*	*Mean*
Dimension 1	.29	.28	.27	.28
Dimension 2	.28	.30	.28	.29

3 Dimension Solution	*High-time Pilots*	*Mid-time Pilots*	*Low-time Pilots*	*Mean*
Dimension 1	.29	.28	.27	.28
Dimension 2	.27	.29	.29	.28
Dimension 3	.26	.28	.28	.27

Individual two-way repeated measures ANOVAs were conducted on pilots' dimension weights from each 2D and 3D solution for each KET. Within each ANOVA, Dimension was the within-subjects variable, Pilot Experience Group was the between-subjects variable and Dimension Weight was the dependent variable. Results of each ANOVA are described below.

When necessary, Tukey's HSD was used for post-hoc comparisons between Pilot Experience Groups and Paired t-tests were used for post-hoc comparisons between Dimensions for each level of Pilot Experience Group. A significance level of .05 was adopted for all analyses and comparisons.

Ability to Discriminate Pilot Experience: Relationship Judgment Task

Three-dimensional (3D) solution. There was a significant main effect of Dimension ($F(2,32) = 21.51$, $p \leq .05$, partial $\eta^2 = .57$). There was no main effect of Pilot Experience Group (p=n.s.) nor did it interact with Dimension (p=n.s.). Paired t-tests revealed that pilots with different levels of experience relied upon the dimensions differently (Table 12.2). Pilots with more experience tended to rely on Dim 1 ("Severity") significantly more than the other two dimensions when making their judgments. Mid-Time and High-Time pilots showed a greater difference between the weights for Dim 1 ("Severity") and the other two dimensions than did the Low-Time pilots who weighted Dim 1 ("Severity") significantly higher than Dimension 2 but not Dimension 3.

Table 12.2. *Results for paired comparisons (t-tests) between dimension weights for each Pilot Experience Group on each Dimension for the Relationship Judgment Task 3D WMDS solution.*

Pilot Experience Groups	**Paired Comparisons** *(paired t-tests, alpha = .05)*		
	Dimension 1 vs. Dimension 2	**Dimension 1 vs. Dimension 3**	**Dimension 2 vs. Dimension 3**
Low-Time Pilots	$t(5) = 3.28$, $p \leq .05$ Dimension 1 was more important (weighted higher)	*p=n.s.*	*p=n.s.*
Mid-Time Pilots	$t(4) = 4.54$, $p \leq .05$ Dimension 1 was more important (weighted higher)	$t(4) = 2.83$, $p \leq .05$ Dimension 1 was more important (weighted higher)	*p=n.s.*
High-Time Pilots	$t(7) = 4.76$, $p \leq .05$ Dimension 1 was more important (weighted higher)	$t(7) = 4.18$, $p \leq .05$ Dimension 1 was more important (weighted higher)	*p=n.s.*

Two-dimensional (2D) solution. There was a significant main effect of Dimension ($F(1,16) = 35.26$, $p \leq .05$, *partial* $\eta^2 = .69$). Paired t-tests showed that pilots in all three Pilot Experience Groups relied significantly more on Dim 1 ("Severity") than Dimension 2 ("Seasonal" or "Go/No-go") when making their judgments (Table 12.3). There was no main effect of Pilot Experience Group (*p=n.s.*) nor did it interact significantly with Dimension (*p=n.s.*).

Table 12.3. *Results for paired comparisons (t-tests) between dimension weights for Pilot Experience Group on each Dimension for the Relationship Judgment Task 2D WMDS solution.*

Pilot Experience Groups	**Paired Comparisons** *(paired t-tests, alpha = .05)*
	Dimension 1 vs. Dimension 2
Low-Time Pilots	$t(5) = 2.62$, $p \leq .05$ Dimension 1 was more important (weighted higher)
Mid-Time Pilots	$t(4) = 3.31$, $p \leq .05$ Dimension 1 was more important (weighted higher)
High-Time Pilots	$t(7) = 4.18$, $p \leq .05$ Dimension 1 was more important (weighted higher)

In summary, Dimension 1 ("Severity") was relied upon most heavily by all three Pilot Experience Groups for both the 2D and the 3D solutions. However, weights did not differ between Pilot Experience Groups for any of the dimensions in either the 2D or 3D solutions. In other words, none of the Dimensions defining the "expert" conceptual structure were any more important to the High-Time pilots than they were to the Mid-Time or Low-Time pilots. This finding calls into question the validity of the Relationship Judgment task in being able to distinguish between pilots with different levels of experience.

Ability to Discriminate Pilot Experience: Card Sort Task

Three-dimensional solution. Analysis revealed a marginally significant interaction between Dimension and Pilot Experience Group ($F(4, 70) = 2.43$, $p=.06$, *partial* $\eta^2 = .12$) and significant main effects for Dimension ($F(2,70) = 21.92$, $p \leq .05$, *partial* $\eta^2 = .39$) and for Pilot

Experience Group ($F(2, 35) = 4.33$, $p \leq .05$, *partial* $\eta^2 = .20$). Dim 1 ("Severity") was weighted highest of all three Dimensions by all three Pilot Experience Groups. Low-Time and Mid-Time pilots had higher weights on Dim 1 than High-Time pilots but the difference between the means was not statistically significant. Overall, High-Time pilots had significantly higher average weights on the dimensions (M=.49) than either Mid-Time pilots (M=.42) or Low-Time pilots (M=.42). Dimension weight comparisons within each Pilot Experience Group revealed that Low-Time and Mid-Time pilots relied on Dim 1 ("Severity") significantly more than Dimensions 2 and 3 (Table 12.4). There were no differences among the weights for the dimensions by High-Time pilots (p=*n.s.*), meaning that High-Time pilots relied on each of the dimensions fairly equally.

Table 12.4. *Results for paired comparisons (t-tests) between dimension weights for each Pilot Experience Group on each Dimension for the Card Sort Task 3D WMDS solution.*

Pilot Experience Groups	**Paired Comparisons** *(paired t-tests, alpha = .05)*		
	Dimension 1 vs. Dimension 2	**Dimension 1 vs. Dimension 3**	**Dimension 2 vs. Dimension 3**
Low-Time Pilots	$t(11) = 2.77$, $p \leq .05$ Dimension 1 was more important (weighted higher)	$t(11) = 2.60$, $p \leq .05$ Dimension 1 was more important (weighted higher)	*p=n.s.*
Mid-Time Pilots	$t(11) = 7.02$, $p \leq .05$ Dimension 1 was more important (weighted higher)	$t(11) = 4.11$, $p \leq .05$ Dimension 1 was more important (weighted higher)	*p=n.s.*
High-Time Pilots	*p=n.s.*	*p=n.s.*	*p=n.s.*

Two-dimensional solution. There were significant main effects of Dimension ($F(1,35) = 12.98$, $p \leq .05$, *partial* $\eta^2 = .27$) and Pilot Experience Group ($F(1,2) = 3.45$, $p \leq .05$, *partial* $\eta^2 = .17$) on Dimension Weights but no significant interaction between Dimension and Pilot Experience Group (p=*n.s.*). Dim 1 ("Severity") was weighted higher than Dim 2 ("Seasonal") by all three Pilot Experience Groups. Post-hoc comparisons indicated that High-Time pilots showed significantly higher overall averaged Dimension Weights (M=.61) for the dimensions in the "expert" conceptual structure than did Low-Time pilots (M=.54) or Mid-Time Pilots (M=.54), although only

the difference between High-time and Low-time was significant. Paired t-tests confirmed that Mid-time and Low-time pilots relied more heavily on Dim 1 ("Severity") compared to Dim 2 ("Seasonal"), as evidence by the higher dimension weights for Dim 1 (Table 12.5). However, High-time pilots relied on Dim 1 ("Severity") and Dim 2 ("Seasonal") more equally, with the difference between Dim 1 and 2 weights failing to reach significance (p=n.s.).

Table 12.5. *Results for paired comparisons (t-tests) between dimension weights for Pilot Experience Group on each Dimension for the Card Sort Task 2D WMDS solution.*

Pilot Experience Groups	**Paired Comparisons** *(paired t-tests, alpha = .05)* **Dimension 1 vs. Dimension 2**
Low-Time Pilots	$t(11) = 2.06$, $p \leq .05$ Dimension 1 was more important (weighted higher)
Mid-Time Pilots	$t(11) = 3.55$, $p \leq .05$ Dimension 1 was more important (weighted higher)
High-Time Pilots	*p=n.s.*

In sum, High-Time pilots had the highest averaged dimension weights when averaged across all Dimensions in both the 2D and 3D WMDS solutions. Mid-Time and Low-Time pilots tended to mainly rely on Dim 1 ("Severity") to make their judgments. This was true for both the 2D and 3D solutions. High-Time pilots generally seemed to make their judgments while placing more equal importance on all of the dimensions comprising the "expert" conceptual structure (Figure 12.1 provides an illustrated summary of these findings). The fact that the High-Time pilots seemed to weight the dimensions of the "expert" conceptual structure differently than the other experience groups provides evidence that the Card Sort may be a more valid KET than Relationship Judgment.

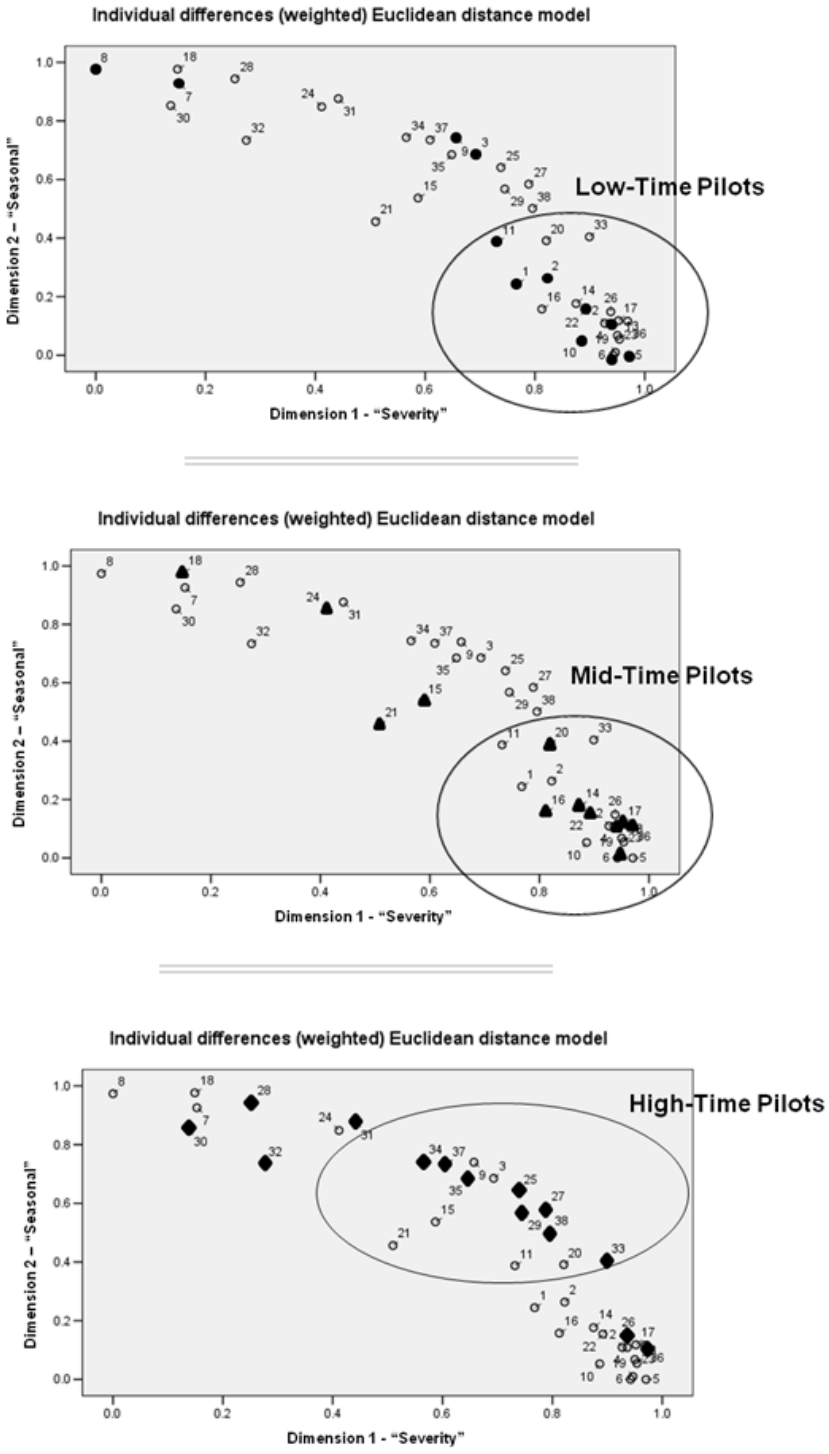


Figure 12.1. Visual illustration of the participant space from the Card Sort 2D WMDS solution. Dimension Weights for pilots within each Pilot Experience Group are highlighted and represent the salience of each dimension for pilots when making their Card Sort groupings.

Ability to Discriminate Pilot Experience: Prime Recognition Task

Three-dimensional solution. There were no significant differences in the Dimension weights as a function of Pilot Experience Group or as a function of Dimension (*p*=*n.s.*).

Two-dimensional solution. Again, there were no significant differences in the Dimension weights as a function of Pilot Experience Group or as a function of Dimension (*p*=*n.s.*).

In sum, Dimension weights did not differ by Pilot Experience Group or by Dimension. Also note that the dimension weights themselves were overall much lower in the Prime Recognition Task compared to the Card Sort and Relationship Judgment Tasks. These lower weights were mostly likely a byproduct of the overall lower fit of the 2D and 3D solutions to the proximity data (as indicated by their respective *Stress* and R^2 values).

Summary of Discrimination Ability

Individual WMDS analyses indicated that Pilot Experience Groups were only distinguishable based on their conceptual structures derived from Card Sort. There were no differences between the Pilot Experience Groups in the importance they placed on any of the dimensions in the 2D or 3D solutions for either Relationship Judgment or Prime Recognition Task. In both Card Sort and Relationship Judgment, Dimension 1 ("Severity") was weighted higher than the other Dimensions. In Relationship Judgment, the reliance on Dimension 1 over the other dimensions was fairly consistent across Pilot Experience Groups. However, in the Card Sort, High-Time pilots tended to weight the dimensions more equally, while Low-Time and Mid-Time pilots really seemed to rely on Dimension 1 much more than the other dimensions. Prime Recognition Task does not appear to tap into any knowledge that is shaped by experience, as the Pilot Experience Groups did not differ at all in their usage of the dimensions. Further, the fact that the weights for all dimensions were rather low and the model fits (i.e., *Stress* and R^2) were much poorer compared to the other two KETs suggests that the nature of

the response times as proximity data precludes MDS from resolving any meaningful relationships in the data.

Thus, assuming that knowledge structures of highly experienced pilot should be different from knowledge structures of inexperienced pilots, Card Sort was the only technique that showed evidence of being a valid KET for this particular purpose in that 1) High-Time pilots had higher overall dimension weights compared to Mid-time and Low-Time pilots and 2) High-Time pilots seemed to rely on more than just Dimension 1 ("Severity") to make their judgments whereas the judgments of the Low-Time and Mid-Time pilots seemed to be more heavily based on Dimension 1 over the other dimensions.

Classification

The use of classification to assess validity is also predicated on the assumption that knowledge about a topic area is organized in memory differently for an expert compared to a novice. Therefore, a KET is valid if and only if it elicits proximity data that reflect these differences in experience. The previous section compared the KETs in terms of how well the Pilot Experience Groups were discriminable based on their conceptual structures. This section describes the assessment of KET validity in terms of in terms of how well a pilot can be classified into a particular experience group based on the individual's conceptual structure derived from the proximity data.

Classification procedures are generally concerned with assigning an object to one of two or more groups based on a combination of attributes. These procedures generally involve comparing each object's combination of attribute values to each group's prototype to find the closest group for membership. Discriminant analysis is one classification technique that has been used to distinguish groups based on MDS dimension weights (Jones & Young, 1972; Kruskal & Wish, 1978) and can be used to test the validity of conceptual structure representations (Schvaneveldt et al., 1985).

Procedure

Discriminant analyses were conducted for each KET to determine whether a pilot's dimension weights from their conceptual structures could predict Pilot Experience Group membership. Although 2D WMDS solutions were identified as optimal for interpretation, it was decided to conduct discriminant analyses on dimension weights from both a 2D WMDS solution and a 3D WMDS solution to explore how classification accuracy for a given KET may differ with complexity of the WMDS solution. Thus, dimensions weights from both the 2D and 3D WMDS solutions for each KET were used in separate discriminant analysis as predictors to Pilot Experience Group membership (Low-Time, Mid-Time, High-Time), resulting in six individual discriminant analyses.

It should be noted that one serious limitation of discriminant analysis is its potential sensitivity to sample size (Mertler & Vanetta, 2001). The suggested ratio of total sample size to the number of variables is 20:1 (Stevens, 1992, as cited in Mertler & Vanetta, 2001). For example, the recommended sample size for a discriminant analysis with 2 predictors (e.g., dimension weights from 2D WMDS solution) would require a sample size of at least 40 pilots and the recommended sample size for 3 predictors (e.g., dimension weights from the 3D WMDS) is 60. Recall the sample sizes for each KET were Relationship Judgment (*N*=19), Card Sort (*N*=38) and Prime Recognition Task (*N*=24), and therefore constitute very small sample sizes compared to what is recommended for discriminant analysis to be stable. Therefore, results of the discriminant analyses are designed to be exploratory in nature and are to be interpreted with caution.

Classification Accuracy: Relationship Judgment Task

Two discriminant analyses were conducted on data from the Relationship Judgment Task to determine whether the dimension weights of the 2D and/or 3D WMDS solutions, respectively, could predict the experience level of a pilot. Bivariate scatterplots indicated that the

assumption of normality of the linear combinations of predictors was not violated for dimension weights for either the 2D or 3D solution. Box's M test failed to reach significance (p>.001) in either solution, indicating that the assumption of homoscedasticity was not violated.

Analysis on 2D WMDS solution, using Dim 1 ("Severity") and Dim 2 ("Seasonal" or "Go/No-Go") Weights as predictors. Two functions were generated but neither function was significant (p=.62, p=.92) indicating that the function of predictors did not significantly differentiate between pilots with different levels of experience. Pilot Experience Group was found to account for 15.6% of Function 1 variance and .06% of Function 2 variance. Standardized function coefficients and correlation coefficients (see Table 12.6) revealed that Dim 2 ("Seasonal" or "Go/No-go") was most associated with Function 1 and Dim 1 ("Severity") was most associated with Function 2. Cross-validated classification results revealed that 66.7% of Low-time pilots, 40% of Mid-time pilots, and 12.5% of High-time pilots were correctly classified as Low, Mid, or High-time (Table 12.7). For the overall sample, 52.6% were correctly classified. Cross-validation derived 36.8% accuracy for the total sample. For Function 1, Low-Time pilots had the highest function mean, indicating that those with high dimension weights on Dim 2 ("Seasonal" or "Go/No-go") were likely to be classified as Low-Time (Table 12.6). For Function 2, the function means of all three Pilot Experience Groups were fairly low, indicating that none of the group differences lend much support to Function 2.

Analysis on 3D WMDS solution, using Dim 1 ("Severity"), Dimension 2, and Dimension 3 weights as predictors). Two functions were generated and both were marginally significant, Function 1: Wilks' Lambda = .45, χ^2 (6, N=19) = 11.96, p=.06; Function 2: Wilks' Lambda = .70, χ^2 (6, N=19) = 5.41, p=.07, indicating that both functions of predictors were marginally significant in differentiating pilots of different experience levels. Pilot Experience Group was found to account for 35.4% of Function 1 variance and 30.3% of Function 2 variance. Standardized function coefficients and correlation coefficients (Table 12.6) revealed that Dimension 2 was

most associated with Function 1 and Dim 1 ("Severity") and Dimension 3 were most associated with Function 2. Cross-validated classification results revealed that 66.7% of Low-time pilots, 40% of Mid-time pilots, and 37.5% of High-time pilots were correctly classified as Low, Mid, or High-time (Table 12.7). For the overall sample, 57.6% were correctly classified. Cross-validation derived 47.4% accuracy for the total sample. For Function 1, Mid-Time pilots had the highest function mean, indicating that those with high dimension weights on Dimension 2 were likely to be classified as Mid-Time (Table 12.6). For Function 2, Low-Time and High-Time pilots had the highest function means, indicating that those with high dimension weights on Dim 1 ("Severity") and Dimension 3 were likely to be classified as either Low-Time or High-Time.

Table 12.6. *Correlation Coefficients, Standardized Function Coefficients, and Discriminant Function Means for Relationship Judgment using dimension weights from a) the 2 Dimensions in the 2D WMDS solution as predictors and b) the 3 Dimensions in the 3D WMDS solution as predictors.*

a) **2D WMDS Solution**

	Correlation Coefficients with Discriminant Function		**Standardized Function Coefficients**	
Dimension Weights	*Function 1*	*Function 2*	*Function 1*	*Function 2*
Dim 1 ("Severity")	-.435	.892	-.021	1.111
Dim2 ("Seasonal" or "Go/No-Go")	1.00	.018	.991	.503

Discriminant Function Means

Pilot Experience Groups	**Function 1** ***("Seasonal" or "Go/No-go")***	**Function 2** ***("Severity")***
Low-Time Pilots	.527	.014
Mid-Time Pilots	-.496	.025
High-Time Pilots	-.085	-.026

***Note: Both functions failed to reach statistical significance (p=n.s.)*

b) 3D WMDS Solution

Dimension Weights	Correlation Coefficients with Discriminant Function		Standardized Function Coefficients	
	Function 1	*Function 2*	*Function 1*	*Function 2*
Dim 1 ("Severity")	-.294	.048	.604	1.803
Dimension 2	.940	-.109	1.236	.413
Dimension 3	.076	.515	.205	1.859

Discriminant Function Means

Pilot Experience Groups	Function 1 *(Dimension 2)*	Function 2 *(Dim 1 "Severity" and Dimension 3)*
Low-Time Pilots	.527	.014
Mid-Time Pilots	-.496	.025
High-Time Pilots	-.085	-.026

***Note: both functions were marginally significant (p=.06, p=.07)*

Table 12.7. *Classification results for the discriminant analyses conducted on Relationship Judgment dimensional weights from a) the two-dimensional WMDS solution, and b) the three-dimensional WMDS solution.*

a) Based on dimension weights from 2D WMDS Solution

		Predicted Group Membership			Total
		Low-time Pilots	*Mid-Time Pilots*	*High-Time Pilots*	
Original Classification	*Low-time Pilots*	83.3%	16.7%	0%	100%
	Mid-Time Pilots	20%	80%	0%	100%
	High-Time Pilots	50%	37.5%	12.5%	100%
Cross-Validation	*Low-time Pilots*	**66.7%**	16.7%	16.7%	100%
	Mid-Time Pilots	20%	**40%**	40%	100%
	High-Time Pilots	50%	37.5%	**12.5%**	100%

52.6% of original grouped cases correctly classified
36.8% of cross-validated grouped cases correctly classified

b) Based on dimension weights from 3D WMDS Solution

		Predicted Group Membership			Total
		Low-time Pilots	*Mid-Time Pilots*	*High-Time Pilots*	
Original Classification	*Low-time Pilots*	66.7%	0%	33.3%	100%
	Mid-Time Pilots	20%	60%	20%	100%
	High-Time Pilots	25%	25%	60%	100%
Cross-Validation	*Low-time Pilots*	**66.7%**	0%	33.33%	100%
	Mid-Time Pilots	20%	**40%**	40%	100%
	High-Time Pilots	25%	37.5%	**37.5%**	100%

57.9% of original grouped cases correctly classified
47.4% of cross-validated grouped cases correctly classified

Summary. Classification accuracy was the same for Low-Time and Mid-Time Pilots whether using dimension weights the 2D or the 3D WMDS solutions. The slight increase in overall classification accuracy for the 3D solution (47.4%) compared to the 2D solution (36.8%) was primarily because of the slightly increased success in classifying High-time pilots when 3 Dimensions were used as predictors (12.5% accuracy with 2D solution, 37.5% accuracy with the 3D solution). There is still quite a bit of confusion in classifying between Mid-Time Pilots and High-Time pilots in both the 2D and 3D solutions. Using both 2D and 3D solutions, just as many Mid-Time pilots were correctly classified as "Mid-Time" (40%) as were incorrectly classified as "High-Time" (40%) and just as many High-Time pilots were correctly classified as "High-Time" (37.5%) as were incorrectly classified as "Mid-Time" (37.5%). The two discriminant functions based on the 2D WMDS solution were not significant, whereas the two discriminant functions based on the 3D WMDS solution were marginally significant, indicating that classification accuracy increases with increasing complexity. This is not surprising, given that only the 3D WMDS solution provided a "Fair" fit to the data, but the 2D WMDS solution did not.

Classification Accuracy: Card Sort Task

Two discriminant analyses were conducted on data from the Card Sort Task to determine whether the dimension weights of the 2D and 3D WMDS solutions, respectively, could predict the experience level of a pilot. Bivariate scatterplots indicated that the assumption of normality of the linear combinations of predictors was not violated for dimension weights for either the 2D or 3D solution. However, scatterplots showed that data appeared slightly less normal in the 3D solution compared to the 2D solution. Box's M test failed to reach significance ($p>.001$) in either solution, indicating that the assumption of homoscedasticity was not violated.

Analysis on 2D WMDS solution, using Dim 1("Severity" or "Strategic-Tactical" and Dim 2 "Seasonal" Weights as predictors. Two functions were generated but only one function (Function 1) was significant, Wilks' Lambda = .704, χ^2 (4, N=38) = 12.12, $p \leq .05$ indicating that

the function of predictors significantly differentiated between pilots of different experience groups. Pilot Experience Group was found to account for 29.5% of Function 1 variance and only 0.11% of Function 2 variance. Standardized function coefficients and correlation coefficients (Table 12.8) revealed that Dim 2 ("Seasonal") was most associated with Function 1 and Dim 1 ("Severity" or "Strategic-Tactical") was most associated with Function 2. Cross-validated classification results revealed that 8.3% of Low-time pilots, 58.3% of Mid-time pilots, and 71.4% of High-time pilots were correctly classified as Low, Mid, or High-time, respectively (Table 12.9). For the overall sample, 55.3% were correctly classified. Cross-validation derived 47.4% accuracy for the total sample. For Function 1, High-Time pilots had the highest function mean, indicating that those with high dimension weights on Dim 2 ("Seasonal") were likely to be classified as High-Time (Table 12.8). For Function 2, the function means of all three Pilot Experience Groups were fairly low, indicating that none of the group differences lend much support to Function 2.

Table 12.8. *Correlation Coefficients, Standardized Function Coefficients, and Discriminant Function Means for Card Sort using dimension weights from a) the 2 Dimensions in the 2D WMDS solution as predictors and b) the 3 Dimensions in the 3D WMDS solution as predictors.*

a) 2D WMDS Solution

	Correlation Coefficients with Discriminant Function		**Standardized Function Coefficients**	
Dimension Weights	*Function 1*	*Function 2*	*Function 1*	*Function 2*
Dim 1 ("Severity" or "Strategic-Tactical")	-.265	.964	1.791	1.529
Dim2 ("Seasonal")	.649	-.761	2.271	.624

Discriminant Function Means

Pilot Experience Groups	**Function 1** *("Seasonal")*	**Function 2** *("Severity" or "Strategic-Tactical")*
Low-Time Pilots	.407	-.041
Mid-Time Pilots	-.539	.037
High-Time Pilots	.811	.003

***Note: only Function 1 was statistically significant ($p \leq .05$)*

b) 3D WMDS Solution

Dimension Weights	Correlation Coefficients with Discriminant Function		Standardized Function Coefficients	
	Function 1	*Function 2*	*Function 1*	*Function 2*
Dim 1 ("Severity")	-.337	.186	1.585	.431
Dimension 2	.555	-.754	1.535	-.506
Dimension 3	.476	.601	1.435	.896

Discriminant Function Means

Pilot Experience Groups	Function 1 *(Dimension 2)*	Function 2 *(Dim 1 "Severity" and Dimension 3)*
Low-Time Pilots	-.511	-.234
Mid-Time Pilots	.282	.290
High-Time Pilots	.679	-.048

***Note: only Function 1 was statistically significant ($p \leq .05$)*

Analysis on 3D solution, using Dim 1 ("Severity"), Dimension 2, and Dimension 3 Weights as predictors). Two functions were generated and but only one was marginally significant (Function 1), Wilks' Lambda = .733, χ^2 (6, N=38) = 10.56, p=.10, indicating that the function of predictors was marginally significant in differentiating pilots of different experience levels. Pilot Experience Group was found to account for 23.1% of Function 1 variance and 4.6% of Function 2 variance. Standardized function coefficients and correlation coefficients (Table 12.8) revealed a fair amount of equality between the three Dimensions, with Dimension 2 being slightly more associated with Function 1 and Dimensions 2 and 3 were more associated with Function 2 than was Dimension 1. Original classification results revealed that 58.3% of Low-time pilots, 25% of Mid-time pilots, and 71.4% of High-time pilots were correctly classified as Low, Mid, or High-time, respectively (Table 12.9). For the overall sample, 57.9% were correctly classified. Cross-validation derived 52.6% accuracy for the total sample. For Function 1, High-Time pilots had the largest function mean, indicating that those with high dimension weights on Dimension 2 were likely to be classified as High-Time (Table 12.8). For Function 2, Low-Time

and Mid-Time pilots had the highest function means, indicating that those with high dimension weights on Dimensions 2 and 3 were likely to be classified as either Low-Time or Mid-Time.

Table 12.9. *Classification results for the discriminant analyses conducted on Card Sort dimension weights from a) the two-dimensional WMDS solution, and b) the three-dimensional WMDS solution.*

a) Based on dimension weights from 2D WMDS Solution

		Predicted Group Membership			*Total*
		Low-time Pilots	*Mid-Time Pilots*	*High – Time Pilots*	
Original Classification	*Low-time Pilots*	25%	58.3%	16.7%	100%
	Mid-Time Pilots	8.3%	66.7%	25%	100%
	High-Time Pilots	14.3%	14.3%	71.4%	100%
Cross-Validation	*Low-time Pilots*	**8.3%**	66.7%	25%	100%
	Mid-Time Pilots	16.7%	**58.3%**	25%	100%
	High-Time Pilots	14.3%	14.3%	**71.4%**	100%

55.3% of original grouped cases correctly classified
47.4% of cross-validated grouped cases correctly classified

b) Based on dimension weights from 3D WMDS Solution

		Predicted Group Membership			*Total*
		Low-time Pilots	*Mid-Time Pilots*	*High – Time Pilots*	
Original Classification	*Low-time Pilots*	58.3%	25%	16.7%	100%
	Mid-Time Pilots	25%	41.7%	33.3%	100%
	High-Time Pilots	7.1%	21.4%	71.4%	100%
Cross-Validation	*Low-time Pilots*	**58.3%**	25%	16.7%	100%
	Mid-Time Pilots	41.7%	**25%**	33.3%	100%
	High-Time Pilots	7.1%	21.4%	**71.4%**	100%

57.9% of original grouped cases correctly classified
52.6% of cross-validated grouped cases correctly classified

Summary. Dimension weights from the Card Sort WMDS seemed to able to be used to accurately classify High-Time pilots as High-Time regardless whether the simpler 2D WMDS solution was used or the more complex 3D WMDS solution was used. The more complex 3D WMDS solution made the correct classification of Low-Time pilots more likely, but slightly reduced the correct classification of Mid-Time pilots. Using the 2D WMDS solution, Low-Time pilots were more likely to be incorrectly classified as “Mid-Time” (66.7%) and to a lesser extent “High-Time” (25%) than they were to be correctly classified as “Low-Time” (8.3%). Using the 3D

WMDS solution, however, Mid-Time pilots were the most incorrectly classified, with 41.7% incorrectly classified as “Low-Time” and 33.3% incorrectly classified as “High-Time.” Thus, the Card Sort task seems to elicit data upon which more experienced pilots can be correctly classified. However, there is some difficulty in identifying the difference between Low-Time and Mid-Time pilots based on their conceptual structures as elicited from the Card Sort task. Recall that the discriminant function based on the 2D WMDS solution was significant, whereas the discriminant function based on the 3D WMDS solution was only marginally significant.

Classification Accuracy: Prime Recognition Task

Two discriminant analyses were conducted on data from the Prime Recognition Task to determine whether the dimension weights of the 2D and 3D WMDS solutions, respectively, could predict the experience level of a pilot. Bivariate scatterplots indicated that the assumption of normality of the linear combinations of predictors was not violated for dimension weights for either the 2D or 3D solution. Box’s M test failed to reach significance ($p>.001$) in either solution, indicating that the assumption of homoscedasticity was not violated.

Analysis on 2D WMDS solution, using Dimension 1 and Dimension 2 Weights as predictors. Two functions were generated but only one function (Function 1) was marginally significant, Wilks’ Lambda = .667, χ^2 (4, N=24) = 8.31, p=.08 indicating that the function of predictors differentiated between pilots of different experience groups with marginal significance. Pilot Experience Group was found to account for 33.3% of Function 1 variance and 0.01% of Function 2 variance. Standardized function coefficients and correlation coefficients (Table 12.10) revealed that Dimension 1 was most associated with Function 1 and Dimension 2 was most associated with Function 2. Original classification results revealed that 44.4% of Low-time pilots, 50% of Mid-time pilots, and 57.1% of High-time pilots were correctly classified as Low, Mid, or High-time, respectively. For the overall sample, 62.5% were correctly classified (Table 12.11). Cross-validation derived 50% accuracy for the total sample. For Function 1, Mid-

Time pilots had the highest function means, indicating that those with high dimension weights on Dimension 1 were likely to be classified as Mid-Time (Table 12.10). For Function 2, the function means of all three Pilot Experience Groups were fairly low, indicating that none of the group differences lend much support to Function 2.

Analysis on 3D WMDS solution, using Dimension 1, Dimension 2, and Dimension 3 weights as predictors. Two functions were generated and but neither reached significance (p=.66, p=.90), indicating that the functions of predictors did not significantly differentiate between pilots with different levels of experience. Pilot Experience Group was found to account for 17.9% of Function 1 variance and 1.1% of Function 2 variance. Standardized function coefficients and correlation coefficients (Table 12.10) revealed that Dimension 2 was most associated with Function 1 and Dimensions 1 and 3 were most associated with Function 2. Original classification results revealed that 11.1% of Low-time pilots, 12.5% of Mid-time pilots, and 57.1% of High-time pilots were correctly classified as Low, Mid, or High-time, respectively (Table 12.11). For the overall sample, 50% were correctly classified. Cross-validation derived 25% accuracy for the total sample. For Function 1, Mid-Time and High-Time pilots had the highest function means, indicating that those with high dimension weights on Dimension 2 were likely to be classified as Mid or High-Time. For Function 2, the function means of all three Pilot Experience Groups were fairly low, indicating that none of the group differences lend much support to Function 2.

Table 12.10. *Correlation Coefficients, Standardized Function Coefficients, and Discriminant Function Means for Prime Recognition Task using dimension weights from a) the 2 Dimensions in the 2D WMDS solution as predictors and b) the 3 Dimensions in the 3D WMDS solution as predictors.*

a) 2D WMDS Solution

Dimension Weights	Correlation Coefficients with Discriminant Function		Standardized Function Coefficients	
	Function 1	*Function 2*	*Function 1*	*Function 2*
Dimension 1	.692	.722	2.11	-.636
Dimension 2	.289	.957	-1.59	1.52

Discriminant Function Means

Pilot Experience Groups	Function 1 *(Dimension 1)*	Function 2 *(Dimension 2)*
Low-Time Pilots	-.16	.011
Mid-Time Pilots	.865	-.005
High-Time Pilots	-.782	-.009

***Note: only Function 1 was marginally statistically significant (p=.08)*

b) 3D WMDS Solution

Dimension Weights	Correlation Coefficients with Discriminant Function		Standardized Function Coefficients	
	Function 1	*Function 2*	*Function 1*	*Function 2*
Dim 1 ("Severity")	.435	.296	-1.387	1.346
Dimension 2	.799	.244	1.855	.768
Dimension 3	.595	-.206	.205	-2.017

Discriminant Function Means

Pilot Experience Groups	Function 1 *(Dimension 2)*	Function 2 *(Dim 1 "Severity" and Dimension 3)*
Low-Time Pilots	.094	.126
Mid-Time Pilots	.453	-.095
High-Time Pilots	-.638	-.053

***Note: both functions failed to reach statistical significance (p=n.s.)*

Table 12.11. *Classification results for the discriminant analyses conducted on Prime Recognition Task dimensional weights from a) the two-dimensional WMDS solution, and b) the three-dimensional WMDS solution.*

a) Based on dimension weights from 2D WMDS Solution

		Predicted Group Membership			Total
		Low-time Pilots	*Mid-Time Pilots*	*High – Time Pilots*	
Original Classification	*Low-time Pilots*	55.5%	22.2%	22.2%	100%
	Mid-Time Pilots	25%	50%	25%	100%
	High-Time Pilots	0%	14.3%	85.7%	100%
Cross-Validation	*Low-time Pilots*	**44.4%**	22.2%	33.3%	100%
	Mid-Time Pilots	25%	**50%**	25%	100%
	High-Time Pilots	28.6%	14.3%	**57.1**%	100%

62.5% of original grouped cases correctly classified
50% of cross-validated grouped cases correctly classified

b) Based on dimension weights from 3D WMDS Solution

		Predicted Group Membership			Total
		Low-time Pilots	*Mid-Time Pilots*	*High – Time Pilots*	
Original Classification	*Low-time Pilots*	22.2%	44.4%	33.3%	100%
	Mid-Time Pilots	12.5%	62.5%	25%	100%
	High-Time Pilots	0%	28.6%	71.4%	100%
Cross-Validation	*Low-time Pilots*	**11.1%**	55.6%	33.3%	100%
	Mid-Time Pilots	50%	**12.5%**	37.5%	100%
	High-Time Pilots	14.3%	28.6%	**57.1%**	100%

50% of original grouped cases correctly classified
25% of cross-validated grouped cases correctly classified

Summary. Discriminant analysis was more successful in classifying pilot experience based on dimension weights when using the simpler 2D WMDS solution, especially with respect to classifying Low-time and Mid-time Pilots. Cross-validation percentage of correct classification decreased quite significantly for Low-time and Mid-Time pilots as well between the 2D and 3D WMDS solutions.

Summary of Classification Results

The dimensional weights from conceptual structures were moderately successful predictors of pilot experience, with some KETs outperforming others in classifying certain groups depending on the complexity of the conceptual structure (i.e., the number of Dimension

Weights used as predictors). When dimension weights were used from the more complex conceptual structures (3D WMDS solutions), Low-Time pilots were more successfully classified based on the Relationship Judgment conceptual structure while High-Time pilots were more successfully classified based on the Card Sort conceptual structure (Figure 12.2). No KET conceptual structure was able to be used to classify Mid-time pilots with any great success, although Relationship Judgment did so with slightly greater than chance accuracy. The Prime Recognition Task conceptual structure was not at all successful in classifying Low and Mid-time pilots but curiously was much more accurate when classifying High-time pilots. The cause of this drastic increase in accuracy among Pilot Experience Groups is not known, as most of the other analyses have consistently shown Prime Recognition Task to be fairly uninfluenced by pilot experience.

When dimension weights were used from the less complex conceptual structures (the 2D WMDS solutions), the Relationship Judgment conceptual structure was again most accurate for classifying Low-Time pilots and the Card Sort conceptual structure was most accurate in classifying High-Time pilots (Figure 12.3). Mid-Time pilots were more successfully classified by Card Sort and Prime Recognition Task conceptual structures when they were based on the less complex 2D WMDS solutions. Classification accuracy of Relationship Judgment for Mid-Time pilots was unaffected by complexity of the solution.

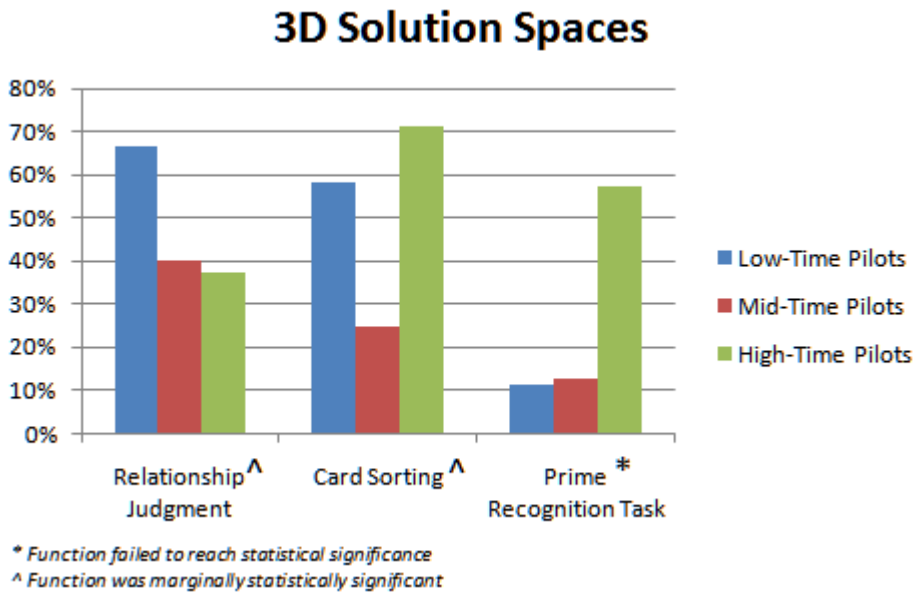


Figure 12.2. Cross-validation classification accuracy from the discriminant analysis applied to three dimension weight data for each Pilot Experience Group and KET.

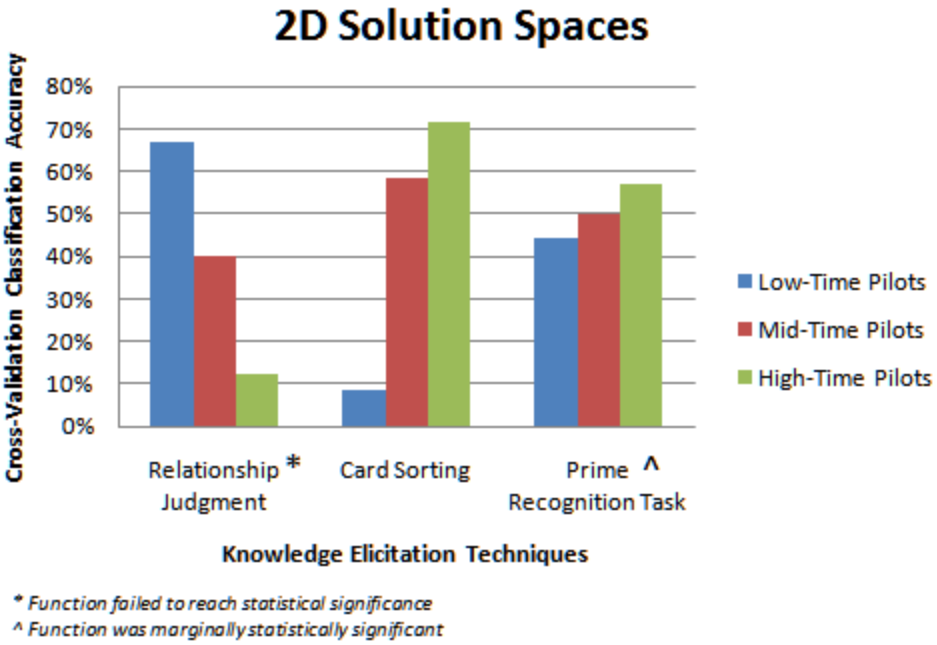


Figure 12.3. Cross-validation classification accuracy from the discriminant analysis applied to two dimension weight data for each Pilot Experience Group and KET.

Thus, classification results suggest that each of the KETs elicit proximity data that when analyzed using MDS yield conceptual structures that capture some valid structural information about how pilot experience level influences the organization of weather information in memory. The Relationship Judgment KET appears to capture more of the structural information useful for discriminating Low-Time pilots from the rest of the group while the Card Sort KET appears to capture more of the structural information useful for discriminating High-Time pilots from the rest of the group. However, these results must be taken with caution, as the sample size was small for discriminant analysis and for some of the KETs the functions failed to reach statistical significance or were only marginally statistically significant.

Summary of Validation

The focus of this section was on validating the conceptual structures as elicited from the three KETs: Relationship Judgment, Card Sort, and Prime Recognition Task. Validation was examined in different two ways. Both analyses indicated Card Sort to be the most valid of the three KETs.

First, Card Sort was only KET from which it was possible to discriminate between pilots of different experience level. High-time pilots had higher overall weights for the dimensions that underlie the “expert” conceptual structure compared to Low-Time and Mid-Time pilots, meaning that those dimensions had a greater influence on the Card Sorts of the High-Time pilots, which in turn implies that the High-Time pilots had a greater understanding of the meaning of those dimensions. Also, High-Time pilots were found to rely on more than just the “Severity” dimension to make their Card Sorts, whereas Low-Time and Mid-Time pilots had much higher weights for “Severity” than the other dimensions indicating that they relied on that dimension primarily in sorting the concepts.

Second, classification analysis indicated that the conceptual structure based on Card Sort contained more structural information useful for classifying High-Time pilots from the rest of

the group than did the other KETS. The Relationship Judgment conceptual structure maintained a relatively high cross-validation accuracy for classifying Low-Time pilots, however, the functions for Relationship Judgment failed to reach statistical significance, indicating that the function of predictors did not significantly differentiate between pilots with different levels of experience.

Chapter 13 - Practical Evaluation of the KETs

Even KETs that elicit highly valid conceptual structures may still have their usefulness questioned by those in applied fields if they do not fit well with the characteristics that describe applied environments such as tight timelines, minimal resources and personnel who may have limited formal training in Human Factors, Psychology, and/or Statistics. Hoffman et al. (1995) identified several dimensions upon which KET efficiency and effectiveness can be qualitatively evaluated (see Table 13.1). Efficiency was evaluated in terms of amount and type of necessary resources. Effectiveness was evaluated in terms of KET simplicity, artificiality, and flexibility.

Table 13.1. *Operationally defined dimensions on which the efficiency and effectiveness of the KETs were qualitatively compared (adapted from Hoffman et al., 1995, p. 142).*

Dimension of Efficiency & Effectiveness	Operational definition	Relationship Judgment	Card Sort	Prime Recognition Task
Simplicity of materials	The number of stimuli or other materials and their complexity relative to the familiar task	Fair	Excellent	Poor
Simplicity of the task	Brevity of the instructions necessary to specify precisely what the participant is expected to do	Good	Good	Fair
Brevity of the task	Total time on task, or total time relative to the duration of the familiar task	Fair (25-35 min)	Excellent (under the participants' control)	Poor (45-60 min and at most half of data collected is usable)
Flexibility of the task	Is it adaptable to different materials, different participants, variations in instructions, etc.?	Good	Excellent	Poor
Artificiality of the task	How much, and in what ways, does it depart from the familiar tasks?	Fair (has some artificiality)	Good (has little artificiality)	Poor (does not resemble familiar tasks in any way)
Data format	Do the data records come out of the task in a format ready to be represented in a computer	Good	Poor when administered by hand	Fair

Simplicity of the Materials

The same 15 weather-related information concepts were used for each KET. However, each KET required different numbers of stimuli and therefore had different set up requirements.

Relationship Judgment

The Relationship Judgment required that each of the 15 weather information concepts be judged in pairwise comparison with all other weather information concepts, resulting in 105 pairs. To simplify data collection, a computer program was constructed using E-Prime Experimental Lab Software (Schneider, 2000) that would display each of the 105 pairs in random order and collect pilots' 1-9 responses on the computer keyboard. In addition to constructing the computer program, all 105 stimuli had to be created. Also, it is recommended that the order in which an item in each pair is presented should be counterbalanced across all participants, which would have required an additional 105 stimuli to be created. However, order of presentation within each pair was not counterbalanced in the current study. Lastly, an interference task had to created and administered after every rating to ensure there were no carryover effects from previous ratings.

Card Sort

The Card Sort required minimal materials and time for set up. Each of the 15 weather-related items was printed on a small card. Stacks of 15 cards were created for each participant. Blank cards were also made available for creating duplicate cards and for applying labels to the resultant groups. Card sorts could be performed anywhere there was a tabletop surface for grouping the cards.

Prime Recognition Task

The Prime Recognition Task required the most material and set up time of all of the KETs by far. Each of the 105 pairs of the 15 weather-related items were used as prime-target

pairs. Of the 105 “Target” trials (trials where the target was part of the memory set), each concept was used 50% of the time as a prime and 50% of the time as a target. Memory sets had to be custom-created for each Target so that each item in the memory set shared the same physical characteristics or structure as the target. For example, if the target was an acronym then all items in the memory set needed to be acronyms of relatively equal length. If the target was two words, then all items in the memory set had to be two words. If at all possible, words were chosen for the memory set that were relatively equal in length as well to reduce any possibility that the participant could rely on any non-semantic information to complete the task. Also, because the response is either “yes” or “no” to whether the target was part of the memory set, the Prime Recognition Task required the set up of “Foil trials” (i.e., trials where the target was not in the memory set) in addition to the 105 Target trials. It is recommended that the number of Target and Foil trials be relatively equal to ensure the participant is responding based on memory and not based on any perceived likelihood ratios. Lastly, an interference task had to created and administered after every trial to help extinguish any memory traces from the concepts used in the previous trials.

Simplicity of the Task

Relationship Judgment

The task of rating the similarity between items was a relatively straightforward and familiar task that required little instruction. However, the experimenter must make the choice about whether or not to define similarity. Defining what “similarity” means will ensure that participants are consistent in how they are evaluating the relationship but it could unfairly bias them into evaluating the items on a relationship that is less meaningful within their overall domain understanding.

Card Sort

The task of sorting items into groups was also a relatively straightforward and familiar task that required little instruction. However, as with Relationship Judgment, the experimenter must make the choice about whether or not to provide further instruction as to the relationship or context within which the items are to evaluated and grouped.

Prime Recognition Task

The Prime Recognition Task was not a familiar task to participants and required a fair amount of initial instruction and practice trials during which they were allowed to answer any clarification questions. Once the practices trials were completed, however, pilots had no difficulty performing the task correctly.

Brevity of the Task

Relationship Judgment

Because all concepts must be evaluated in pairwise comparison with each other concept, a very modest list of 15 concepts required 105 ratings. Typically pilots could complete these 105 ratings in about 25-35 min.

Card Sort

The task of sorting 15 cards could be done rather briefly, within a matter of minutes, or the participant could take their time and revise the groupings multiple times until they were satisfied. The important point is that the duration of the task is largely under the participants' control.

Prime Recognition Task

The 105 Target Trials (105 prime-target pairs), the 15 additional "Target" trials that involved acronyms, and the additional 79 "Foil" trials resulted in 199 trials for the participants to

complete. Typically participants completed this task in 45-60 min. However, this task is extremely inefficient considering that of those 199 trials, only data from *at most* 105 trials will be used to build conceptual structures. The phrase “at most” is used because data can only be used from trials in which participants are correct in their memory that the target was in the memory set. Therefore, depending on participant accuracy, there may be missing data for some of the 105 pairs on an individual participant basis.

Flexibility of the Task

Relationship Judgment

The Relationship Judgment Task is fairly flexible. Concepts to be compared can be words, phrases, pictures, etc. Participants can be asked to judge relationships between items in any number of different ways (e.g., similarity, dissimilarity, relatedness, etc.). However, because it requires pairwise comparisons, even a small increase in the number of concepts that are of interest will result in a much larger increase in the number of pairwise comparisons needed. For example, increasing the list of 15 to 18 will increase the number of pairwise comparisons from 105 to 153. There are some alternative methods for administering rating tasks that do not require participants to rate every pairwise comparison, but those methods often require increasing the number of participants.

Card Sort

Like the Relationship Judgment, Card Sorts can be done with any number of different types of concepts, including words, phrases, pictures, etc. Instructions and protocol can be varied to allow participants to perform open card sorts (where participants create their own groups) or closed card sorts (where participants sort cards into pre-specified groups). Card sorts are very capable of handling large numbers of cards (up to 100) with less dramatic effects on total time on task compared to the Relationship Judgment (Spencer, 2009).

Prime Recognition Task

Because of all of the required procedures for setting up a Prime Recognition Task, the task itself is fairly inflexible to different materials. It would be very difficult to conduct a Prime Recognition Task (in the form used in the current study) using anything other than one- or two-word concepts. Also, the Prime Recognition Task is based on response time and requires a motor response. Since motor response time can get slower with increasing age, it makes the Prime Recognition Task difficult to justify using for reliably discriminating between groups on a certain factor when age is not controlled within each group.

Artificiality of the Task

Relationship Judgment

Participants are asked to make judgments about the relationship between two items independently, without respect or consideration for any overlapping or moderating relationships that may occur in combination with other concepts as well. This independence adds some artificiality to the task, considering that in any real-world context within which the two items co-occur, there will most likely never be a need to assess the relationship between these two items in isolation.

Card Sort

The extent to which the Card Sort is artificial depends largely on the type of concepts that are used in the sort. If the concepts are not representative of the domain or are unfamiliar to the participant, then the Card Sort could be considered to be highly artificial. However, the Card Sort, when used with appropriately representative concepts, is the least artificial of all of the KETs evaluated in the current study. Also, the physical task of spreading cards out on a table, putting similar items nearer to each other and eventually putting like items into groups is a very natural task, making it very intuitive for participants (Spencer, 2009).

Prime Recognition Task

The Prime Recognition Task does not at all resemble any familiar task. Participants often expressed confusion as to the reason they were being asked to complete this task as they saw no relevance between it and display design.

Data Format

Relationship Judgment

In the current study, the Relationship Judgment Task was administered via computer. Therefore, no by-hand data entry or coding was required. Data were able to be exported from E-Prime into Excel for further analysis. Exported data required some cleaning and formatting but overall the data format process was relatively efficient.

Card Sort

In the current study, Card Sorts were administered and conducted by hand, meaning that participants physically sorted and grouped cards. Cards that were formed into groups were physically attached (via paper clip) and placed into plastic bags. At the end of data collection, all groupings had to be hand-coded into excel spreadsheets. This process took several days to complete. However, there are some computer-based software programs that allow card sorts to be administered electronically. Spencer (2009) provides a review of card sort software. These software programs will collect the card sort data and enter it into some type of analysis tool automatically. However, some of the analysis tools are fairly limited in what they can do. Also, most software programs will not allow the use of duplicate cards and/or hierarchical groups to be created (at the time of data collection for the current study, no software program supported any of these options). Therefore, while software programs are available to make the data formatting more efficient, they tend to do so at the expense of opportunities for flexibility in how the card sort is administered.

Prime Recognition Task

Like the Relationship Judgment Task, all data are collected via computer which negates the need for any hand coding. Data can be exported from E-Prime into Excel for analysis. Again, the data file requires the same cleaning and formatting that the Relationship Judgment Task does. However, Prime Recognition Task requires the added procedures of 1) removing Foil Trials, 2) removing target trials where the participant was “incorrect” in their response, and 3) identifying when response times constitute outliers and how to deal with outliers.

Summary

Of all of the dimensions used to qualitatively evaluate efficiency and effectiveness of the KETs, the Card Sort method was identified as being the most efficient and effective. It is flexible in its ability to handle a lot of concepts without drastic increases in task completion time. Sorting and grouping related items, especially physical representations of them, comes very naturally to most participants. It provides a more realistic opportunity for participants to evaluate each concept within the context of its relationship to multiple other concepts and allows the groupings to be revised in iterative fashion until the participant is satisfied that it adequately represents their view. The major drawback to card sorts is the need for manual coding if the card sorts are administered manually.

Chapter 14 - General Discussion

The main goal of this research was to compare and contrast three different methods for eliciting and representing pilots' knowledge for weather-related information. Each of the three KETs is grounded in psychological theory and each has a history of employment in the fields of Human Factors and HCI. This study sought to increase the understanding of the strengths and weaknesses and similarities and differences of these techniques in terms of the data they yield, the information about knowledge structure they elicit, and the resources they require. Once the strengths and weaknesses have been assessed, guidelines can be established for selecting the appropriate KET to arrive at the desired information.

The approach taken to compare the three KETs was based on one fundamental tenet – that individuals who differ in skill level will also differ in knowledge structure – for which previous research in Cognitive Psychology provides support. Thus, each of the KETs was evaluated in terms of how well the data and the information derived from that data reflected differences associated with skill level of the pilot participants. KET(s) that represented or maintained differences in output as a function of skill level were considered to be more valid conceptual structures than KET(s) that could not discriminate between pilots of different experience.

Results Review

Data Collected

The first analysis was conducted to understand the similarities and differences between proximity data elicited by each of the techniques. Proximity data from the Relationship Judgment and Card Sort Tasks were significantly correlated, meaning that items that were judged as being similar in the Relationship Judgment task tended to be grouped with each other in the Card Sort. This significant correlation between the techniques maintained for all three Pilot Experience Groups, with the highest correlation occurring on proximity data for High-Time

pilots. However, proximity data from the Prime Recognition Task was not correlated with proximity data from the Relationship Judgment Task and only marginally correlated to proximity data from the Card Sort Task. This suggests that the Relationship Judgment and Card Sort Techniques tap in to a similar type of knowledge structure relative to the Prime Recognition Task. This finding was as hypothesized, since the Relationship Judgment and Card Sort tasks were both explicit KETs that specifically asked pilots to consider similarity or relationships between items, whereas the Prime Recognition Task was an implicit KET in which relationship between items is inferred based on response times and memory performance.

Information about Knowledge Structure

Several analyses were conducted to examine how the information gained about knowledge structure was influenced by the type of technique used to elicit it. Special focus was on evaluating which technique elicited proximity data and ultimately conceptual structures that differed with pilot experience. Card Sort was the only KET to reliably support and represent differences between pilots of different experience across all analyses. The Mantel Test showed that Card Sort was the only KET to elicit proximity data on which pilots with the same level of experience were more correlated with each other than they were with pilots who had different levels of experience.

Proximity data from each KET were submitted to WMDS and the analyses on the resulting conceptual structures further supported the greater validity of the Card Sort technique. WMDS achieved at least a “Fair” fit with the least complex solution for the Card Sort data. Card Sort data were fit with a 2D solution, whereas more complex solutions were required for Relationship Judgment (3D) and Prime Recognition Task (4D) to achieve a “Fair” model fit. This low complexity (2D) solution was an even better fit for Card Sort data from High-Time pilots, whereas for the other two KETs the fit did not improve with experience level. Previous research suggests that knowledge organization actually simplifies with increased experience.

Experience allows one to identify and maintain the important information and associations, yielding a simpler organization (e.g., Schvaneveldt et al., 1985; Ye & Salvendy, 1994).

SMEs were only able to ascertain meaning for the dimensions underlying the Card Sort and Relationship Judgment 2D conceptual structures. Further, in the Card Sort conceptual structure they were able to identify differences in how pilots of various levels of experience seem to interpret those dimensions based on how the weather information concepts were located along those dimensions as a function of Pilot Experience Group. However, they were unable to identify any such differences among the Pilot Experience Groups for their Relationship Judgment conceptual structures. Further, they were unable to interpret any meaning for the dimensions that defined the conceptual structures from the Prime Recognition Task.

"Expert" conceptual structures were defined for each KET (based on data from High-Time pilots) and analyses were conducted to understand how important each of the dimensions of the expert conceptual structure was to the overall judgments/groupings/response times of pilots within each Pilot Experience Group. Given that experience influences knowledge structure, it was assumed that the dimensions for the expert conceptual structure should be more important (weighted higher) for High-Time pilots than for pilots with less experience. Results showed Pilot Experience Groups differed in the importance perceived for the dimensions for Card Sort conceptual structures only. High-Time pilots had overall higher dimension weights than the other two experience groups and tended to rely on the dimensions more equally while Low- and Mid-Time pilots seemed to rely more heavily on the "Severity" dimension when completing their card sorts. Pilot Experience Groups did not differ in their importance of the dimensions that defined the expert conceptual structures for either the Relationship Judgment or Prime Recognition Task.

Lastly, Discriminant Analyses performed on the dimension weights of respective KET conceptual structures showed that the conceptual structure based on Card Sort contained more

structural information useful for accurately classifying High-Time pilots from the rest of the group than the other KETs. The Relationship Judgment conceptual structure showed greatest accuracy for classifying Low-Time pilots from the rest of the group than did the other KETs; however, the functions for the Relationship Judgment failed to reach statistical significance, calling into question the reliability of those results.

Type of Knowledge Elicited

Up to this point, KET validity has been evaluated in terms of how well each of the KETs result in conceptual structure that reflect differences in knowledge structure presumed to occur as a result of different levels of experience. However, another important consideration when evaluating KETs is to understand what type of knowledge is being elicited by each technique. The “differential access hypothesis” is a belief held in the knowledge elicitation domain that different KETs may elicit different types of knowledge (e.g., declarative vs. procedural) and/or evoke different kinds of strategies (e.g., top-down vs. bottom-up reasoning) (Hoffman et al., 1995). For example, Gammack & Young (1985) found that sorting and scaling tasks were best for eliciting interactions between domain concepts and think aloud problem solving and task analysis techniques were best for eliciting procedures and heuristics. Further, some KETs may tap knowledge that is more predictive of performance than others. For example, Rowe et al. (1996) found that relatedness ratings and hierarchical concept listing interviews were better than diagramming and think aloud methods in terms of eliciting knowledge that corresponded to avionics troubleshooting performance. Therefore, the results of the current study taken in combination with an understanding of the characteristics of the KETs should provide some insight into how the KETs may differ in the type of knowledge they attempt to elicit.

Review of KET Task Characteristics

The three KETs differ in several key ways that may impact the type of knowledge and/or the validity of the data that are elicited (see Table 14.1). First, as previously noted, the KETs differed in the amount of awareness that the participant has about the *true purpose of the task*. Participants were aware of the need to evaluate relationships between concepts in the Card Sort and Relationship Judgment tasks (explicit techniques) but were not aware that relationships were being evaluated in the Prime Recognition Task (implicit technique). Second, the KETs differed in terms of *how the concepts were presented for evaluation.* The Card Sort allows the evaluation to occur within the context of all other concepts, the Relationship Judgment Task requires concepts to be evaluated in pairs independently of other pairs, and Prime Recognition Task presents concepts in implicit prime-target pairs but within the context of a memory task rather than an evaluation.

KETs also differed in the amount of *time pressure* the participant felt when completing the task. Card Sort participants had no time pressure (told to take as much time as necessary), Prime Recognition Task participants had high time pressure (told to work quickly and accurately and were aware that response time was being collected). Relationship Judgment participants were not given expectations or instructions regarding time but they may have been motivated to make their judgments more quickly in order to be finished with the rather mundane task. Consequently, only the Card Sort task provided participants with the opportunity to *mentally simulate situations where the concepts would be used* in order gauge relationships between concepts. They could have taken the time during the Relationship Judgment as well, but response times for their similarity judgments suggest that most did not evaluate the relationships to that degree. Only the Card Sort allowed participants the opportunity to *rethink and revise their judgments* to *visually* evaluate, represent, and revise the relationships between concepts.

Table 14.1. *Summary of how the key differences in task characteristics vary among the three KETs.*

Task Characteristics	**Knowledge Elicitation Tasks (KETs)**		
	Card Sort	*Relationship Judgment*	*Prime Recognition Task*
Awareness of the true task purpose (i.e., evaluate relationships between concepts)	Explicit (fully aware)	Explicit (fully aware)	Implicit (not aware)
Presentation of concepts for evaluation	Within the context of other concepts	In pairs, independently evaluated (given an interference task to remove memory traces of previous ratings)	Concepts presented within context of memory task (not for evaluation)
Time pressure	None	Some (implied)	High (aware performance is being timed)
Opportunity to mentally simulate situations that involve the information concept(s) to help make judgments	Yes	Some (but implied time pressure may make participants reluctant to make much effort to simulate)	No (participants are unaware that relationships are being evaluated)
Opportunity to rethink and revise judgment	Yes	No	No
Visual feedback about relationships	Yes – can visually see and tactually manipulate groupings	No	No

Generalizing Results to the CI Model

The Construction Integration Model (Kintsch, 1988) postulates that information can be processed at three different levels of representation. The *surface level* is the minimally processed text information that is initially coded. For example, given "*Nebraska really pounded K-State in football last fall*," the surface level processing reveals the syntactic and orthographic structure of the sentence and is short-lived. At the *text-based level*, information is parsed into predicates (e.g., verbs, adjectives, adverbs) and arguments that are processed in terms of their semantic meaning with minimal inferences (e.g., Nebraska scored more points than K-State).

The *situation model level* activates relevant and non-relevant information stored in LTM and engages elaborative processing and inference generation beyond the semantic meaning of the information. For example, when processing the sentence above, other concepts and memories may be activated that are beyond semantic meaning and are based on previous experiences, like the sad looks on the faces of the K-State fans watching the game on TV at the bar with you, the fun times you had with your friends at the last game you attended at Bill Snyder Family Stadium, and the smell of the funnel cakes that are sold at the games. Thinking about the KET task characteristics within the context of the CI model provides some insight into the reasons why the card sort was the only technique for which experience seemed to play a role.

The Card Sort task was the only technique that explicitly asked pilots to consider each weather-related item within the context of all of the other weather-related items when making their assessments of relationship. The technique also applied the least amount of time pressure on the sorting judgments and allowed pilots to reevaluate their groupings continuously within the context of other items they were sorting. Perhaps it was this opportunity for deliberation and revision that allowed the pilot's experience level to factor into the task. In other words, the Card Sort technique allowed for the pilots to engage in a deeper level of processing of the weather-related information and/or simulate the various scenarios for how these concepts are related in flight (i.e., situation model). It was this deeper level of processing or mental simulation that triggered the associations and usage instances on the basis of the pilot's past experience that ultimately influenced the Card Sort groupings.

Although the Relationship Judgment task did ask pilots to make similarity judgments on aviation-related items, pilots made each relationship judgment between two items individually and were not allowed to revise their judgments like they could in the Card Sort. Pilots had to complete a masking task between trials to further ensure that the judgments were made independently. The fact that the judgments were made individually, independently, and perhaps

under some implicit time pressure may have resulted in pilots relying on text-based level of processing of items (i.e., using the general semantic meaning or textbook definition of the terms) to make their judgments rather than evoking the situation model level of processing at which experience may play a stronger role (Kintsch, 1988). In other words, shallow processing was sufficient to complete the task (Graesser, Singer & Trabasso, 1999).

Less response variability would be expected among Relationship Judgments if they were based on a general "textbook definition" of a term because pilots should have the same general understanding of the semantic meaning of each of the terms regardless of their experience level. Less variability in responses would, in turn, lead to fairly high correlations between all pilots regardless of experience level. Recall that correlations between all pairs of pilots were overall higher for Relationship Judgment proximity data than for data from the other two KETs, further supporting this notion that Relationship Judgments are based on a text-based level of processing of the items. In other words, task demands and participant strategy could have combined to motivate only processing of the "textbook definition" of the weather-related items, resulting in higher average correlations among pilots compared to the other KETs but also a lack of effect of pilot experience on the responses.

The decision to examine the Prime Recognition Task as an alternative technique for knowledge elicitation arose out of concern that the cognitive processing involved with having participants explicitly judge or evaluate relationship introduces unnecessary variability into the data. Specifically, responses could be biased by transient environmental and situational factors (e.g., recent events may stand out in memory as participants think about situations in which the concepts may occur but those events may not be representative of what is typical). The Prime Recognition Task was hypothesized to be able to provide information about knowledge structure that is unbiased by these transient potential influencers. However, results from the current study showed that the Prime Recognition task yielded proximity data that was very noisy and not at all

affected by experience level of the pilot. WMDS required a 4D solution in order to provide a "Fair" model fit to the data and even then, the R^2 value was not nearly as high as it was for "Fair" fits to the data from the other KETs. In addition, the dimensions underlying the conceptual structure based on the 2D MDS solution were uninterpretable by SME pilots. In sum, the Prime Recognition Task did not lend itself well to being a KET used in this context.

The task characteristics of the Prime Recognition Task provide at least a partial explanation of its shortcomings as a KET. At its core, the Prime Recognition Task is a memory task. Perhaps the fact that the task was presented to the pilots as a memory task with both response time and accuracy as primary goals motivated pilots to develop and employ memory retrieval strategies (e.g., mnemonics, location based cues on the screen, visual appearance of the words, etc.) that effectively removed any aviation-related context and processing for the weather items, including the prime and the target. Thus, the priming effect (and therefore the prime-target relationship on which the conceptual structures were ultimately supposed to be based) may not have been based on any aviation-related context, knowledge or experience. This may explain why there was no evidence that pilot experience was a factor influencing the conceptual structures elicited by the Prime Recognition Task.

Another potential limitation of the Prime Recognition Task and its use of response time as proximity data is that it requires a motor response. Motor responses have been shown to decrease with increasing age (Falkenstein, Yordanova & Kolev, 2006). The High-Time pilot group had the highest average age (49 yrs) compared to Mid-Time (41 yrs) and Low-Time (26 yrs) pilots. Therefore, the requirement of a motor response may have confounded the use of response time to assess relatedness of items in knowledge structure. Low-Time Pilots did have significantly quicker response times overall compared to Mid-Time and High-Time Pilots (see Appendix B). If the prime were activating global structure, one would expect High-Time pilots to be overall quicker in response time since highly experienced people tend to have denser

knowledge structures that should facilitate the priming. Perhaps any gain in response time due to experience and priming was confounded by a degradation in motor response time of the older High-Time pilots relative to the other Pilot Experience Groups.

Limitations

The current study had several limitations. Also, decisions were made regarding the methodology that may be worth revisiting if this study was to be replicated and extended.

First, the number of weather-related information concepts on which the conceptual structures were derived was very small (15 concepts). Further, previous research has shown that the strength of relationship between any two concepts is affected by the context in which the judgment is made. One of the factors that shapes the context is the combination of concepts that comprise the stimulus set (Dorsey et al., 1999). Efforts were taken in Phases I and II of the current study to identify the most important weather-related concepts for inclusion in the stimulus set. However, the desire to maintain the same stimulus set across the three KETs combined with the procedural and stimulus requirements of the Prime Recognition Task resulted in a list of 15 concepts that may not have been as comprehensive in their representation of aviation weather-related knowledge structure as would have been desired. Further, since validity was defined as the ability to distinguish between different levels of pilot experience, it may have been beneficial to include different groupings of concepts that would be presumed to be differentiators of experienced-based knowledge (Dorsey et al., 1999).

Second, experience was defined as the total number of flight hours. Previous studies have used total flight hours as an indication of experience (e.g., Goh & Wiegmann, 2002). However, pilots were not specifically recruited based on their total number of hours flown. Instead, the Pilot Experience Groups were identified at the conclusion of the data collection and were designed to maintain relatively equal numbers of pilots per group while also taking into consideration any natural breaks in the distribution of hours flown and the range of hours within

which pilots could logically be expected to have relatively similar flight experiences. Perhaps the KET validity results would have been different if pilots were recruited based on total flight hours such that greater delineations between experience groups could have been maintained. For example, perhaps the Relationship Judgment and Prime Recognition Task would were not sensitive enough to detect conceptual structure differences given the high variability of total hours within each Pilot Experience Group (see Table 9.2). Perhaps if Pilot Experience Groups were to differ more significantly and with less variability in terms of total hours flown, then Relationship Judgment and Prime Recognition Task validity may have been improved.

It is also possible, however, that total hours may not have been the most appropriate parameter on which to base experience, especially with respect to the use of weather information. For example, a High-Time pilot may have been able to accumulate many hours of flight time but decide to never fly in anything but VFR conditions (e.g., sunny) even though he or she may be rated to do so. Also, a pilot may have accumulated many total hours but has maybe flown just enough to stay current in the last several years. Other factors that may be useful to consider when trying to classify experience include how many seasons they have flown through and in which seasons they have accumulated a certain number of hours (seasonal flying is emphasized in FAA flight training), type of airplane operation (recreational or commercial), and type of aircraft flown (low performance, high performance) (G. Shetterly, personal communication, March 18, 2011).

Another potential limitation of the current study is the subjectivity with which the conceptual structure dimensions were identified. SMEs were asked to provide their interpretation of meaning for the underlying dimensions of the conceptual structures derived from each of the KETs. While the SMEs were able to confidently identify meaning in some of the dimensions, they were unable to identify meaning in all of the conceptual structures. Also, even though they felt confident in their interpretations, these interpretations were still very subjective.

Perhaps other types of data could have been collected to help in identifying dimension meaning. For example, pilots could have also been asked to complete a general knowledge measure, such as a general comprehension test (e.g., defining terms, recognizing information) and/or a skilled performance tests (e.g., identify steps needed to accomplish an aviation-related task), the differences in their performance on these tests could have provided insight into the meaning of the dimensions in the conceptual structures (Dorsey et al., 1999; Schiffman et al., 1981). Other types of data that could have provided quantitative help or objective insight into identifying dimension meaning include 1) importance/severity ratings for each of the concepts, 2) frequency of use ratings for each concept, 3) phase of flight within which each concept is used, and 4) familiarity with each concept.

Altering the procedures for administering the KETs could also have provided better insight into the meaning of the dimensions underlying the conceptual structures. For example, pilots performed Relationship Judgments without specific instruction as to what “relatedness” really meant. It is commonly assumed that raters will choose the most relevant features on which to make their comparisons. However, this lack of instruction can lead to difficulty in interpreting dimensions using MDS (Dorsey et al., 1999). However, as discussed in Chapter 9, providing explicit guidance on how relatedness should be judged could lead to biased responses and result in the identification of dimensions and that are not truly meaningful in the actual knowledge structure. When completing the Card Sort, participants were not explicitly told to talk out loud while making their groupings, nor were any notes taken about observations of participant behaviors or comments during the sorts. If participants would have been asked to think aloud while performing the card sorts, their comments and behaviors (e.g., deliberations for specific card placement) could have given some insight into the factors that were influencing their groupings. However, the mere act of talking outloud may have impaired or influenced their

ability to engage in the deeper level of processing that the current results suggest they achieved when completing their sorts.

Lastly, while several attempts were made to evaluate the validity of each of the tasks as elicitation techniques, no analyses were conducted to assess the reliability of each technique. In retrospect, reliability could have been assessed in a couple of different ways. However, each option for assessing reliability also presents some drawbacks as well.

First, trial replications could have been used to examine reliability for each technique (i.e., whether the elicited data were comparable across replication). Ideally, all trials would be replicated, but practically speaking this is rarely an option in applied domains like aviation. Therefore, even replicating a proportion of the total trials could give some indication of the reliability of each technique. For example, pilots could have been asked to rate 30% of the 105 concept pairs twice in the Relationship Judgment, respond to 30% 105 prime-target pairs twice in the Prime Recognition Task and asked to sort the same 15 cards into groups again after some time had passed.

However, adding trials to each of the tasks also adds to the amount of time required to complete each technique (e.g., increasing the number of Relationship Judgments to 137 compared to 105). The addition of extra trials would be even more extreme for the Prime Recognition Task, given that any additional Target trials must come with the addition of Foil trials to maintain roughly the same proportion of Target and Foil trials (important for preventing participants from responding based on any internal calculation of the likelihood of Target Trial presentation). Thus, a 30% increase in prime-target trials would actually result in the addition of 50 trials or more. Considering that the pilots were volunteering their time for no reimbursement, the decision was made to keep the number of necessary trials to a minimum for each technique, thereby sacrificing the ability to assess reliability.

Also, analyses could have been conducted on half of the participants within a technique with results confirmed by analyzing the other half of the participants who completed that technique (i.e., split-half technique). However, given that the sample size for each technique in some cases was already less than ideal, the informativeness of a split-half reliability analysis would have to be questioned. Thus, reliability assessment remains a limitation of this study.

Chapter 15 - Implications and Future Research

Implications

The focus of much of the research regarding display design has been on optimizing display image quality (e.g., luminance, contrast, resolution, legibility, etc.) or on other perceptual factors of display formatting (e.g., color coding, map orientation, etc.). Within the design lifecycle, higher-order issues such as content and information organization typically are assessed later in the design process, when changes are more costly to make and there is less motivation to make the changes. Further, the analysis is usually based on *expert opinion* applied in a relatively informal manner. Thus, the major practical application of this research was to support the use of objectivity (not subjectivity) in evaluating the effectiveness of MFD menu structure and terminology during FAA Avionics certification.

Specifically, the current research study was meant to lay the foundation upon which a tool could be developed to assess the extent to which re-learning a candidate system's menu structure would be required after an extended period of absence from the system. This tool would provide the FAA with data regarding the degree of mapping between the candidate system's menu structure for weather information and the knowledge structure for weather information that the targeted pilot population can be assumed to have. If the degree of mapping between the system's menu structure is high, then the time needed to re-learn a system should be low. If the degree of mapping is low, then more time will be required to re-learn the system and the interaction may be more prone to errors, confusion, and/or the inability to find the necessary information when they need it. Such a metric is especially important for Low-Time GA pilots (e.g., "weekend" flyers) or pilots who routinely engage in "plane switching." Low-Time pilots may need to plane switch because they do not own their own aircraft and therefore need to rent. More experienced pilots also may be plane-switching as they may hold ratings in several different types of aircraft equipped with various types of display technology. Thus, it is important

to understand how the MFD menu structure can be organized to support efficient and effective information access by pilots of different levels of aviation experience.

The first step to developing this menu structure assessment tool is to understand how pilots of different levels of experience think about flying and the type of information on which they rely to maintain safe flight. To that end, the current study has implications for how best to achieve that understanding of pilots' knowledge and information use.

First, when the goal is to define and understand differences in knowledge structure as a function of experience, the Card Sort task provides a valid technique for providing insight into those differences, as it was the only KET to identify any differences in conceptual structure with pilot experience. Further, use of MDS allowed the identification of underlying factors for how knowledge was structured and how experience level affected the importance of those factors. With future research, it may be possible to define different "modes" for MFDs that correspond to different levels of pilot experience. Each mode may be able to highlight and prioritize types of information that are most relevant and important for that Pilot Experience Group for certain key flight situations (e.g., phase of flight).

The current study also provided some insight into the commonalities across pilots of all levels of experience in how they think about weather. All pilots, regardless of experience, think about weather in terms of severity. However, the meaning of severity was interpreted differently by pilots of different experience. For example, Low-Time pilots may evaluate hazardous weather in terms of how it will affect their ability to maintain safe flight. High-Time pilots tend to have the advanced training, the experience, and the type of aircraft that make weather less of a concern in terms of maintaining safe flight. However, they still want to avoid severe weather because they do not want to alarm or inconvenience their passengers.

Third, the current study showed that total flight hours can be used as a valid indicator of pilot experience level when other information about skilled vs. unskilled performance is

unavailable (e.g., performance tests, type of aircraft experience, memory tests, problem-solving tests, etc.). The Card Sort technique was able to identify differences in conceptual structure among different Pilot Experience Groups when experience was defined by total hours flown. Also, in most of the analyses where Pilot Experience Levels were compared, High-Time pilots showed the least amount of variance in their judgments. High-Time pilots also showed more consistency in their knowledge structures compared to Low-Time or Mid-Time pilots. This finding is consistent with previous research demonstrating that variance in knowledge representation decreases with increasing skill level (e.g., Goldsmith & Johnson, 1990). The major benefit of this finding is that total flight hours is a measure that is a consistently tracked industry variable. Thus, while there are certainly other types of indicators that future research may identify as even better indicators of pilot experience, the current study did find value in using total flight hours in the absence of other information.

The current study was designed within the context of an applied environment. However, the implications of this study are not necessarily constrained to the field of Aviation. There are many different domains and product designs that are characterized by the need to display and/or provide access to a large amount of information on a relatively small screen to be used by people of different skill levels. Website design is one such field. The one major issue to consider about applications to other domains is the extent to which experience is expected to affect knowledge structure. In aviation, experience plays a key role in shaping and organizing domain knowledge. Pilots are trained, spend many hours honing their skills, are tested on their skills, and can experience very extreme consequences for poor skilled performance. However, users of websites may vary widely in their experience with the content, their knowledge of the domain, and their task goals. Also, there are less severe consequences for poorly skilled performance in a website compared to in flying a plane. It would be interesting to examine whether Card Sort would still be identified as the most valid KET for eliciting knowledge

structure when the task goals can vary as widely as they may with website usage. Perhaps the experience of using a website requires mostly shallow processing of the presented information, and thus the Relationship Judgment may be the more appropriate elicitation technique. However, any extrapolation of the current results to other domains is highly speculative.

Future Research

The current study provides some insight into the similarities and differences between three KETs. Strengths and weaknesses of each technique have been highlighted. However, more research is needed to truly provide guidance into the most appropriate KET to use to satisfy specific research and design needs.

Making the Link to Performance

This study defined KET validity in terms of the ability to use derived conceptual structure to distinguish between levels of experience. However, just because a conceptual structure is able to identify the knowledge that distinguishes groups from one another does not mean that that knowledge actually correlates with performance (Rowe et al., 1996). In order for validity to truly be established, the correlation between task performance and knowledge representation should be assessed. If a particular KET derives a conceptual structure that is found to be associated with successful performance, then the insights about knowledge gained from that conceptual structure can be used to affect training and/or display design to improve performance.

Thus, the next step in this line of research is to confirm the results of the current study by assessing the validity of each KET with respect to task performance. For example, Card Sort was the only technique for which evidence was found for conceptual structures differing as a function of pilot experience. Thus, one way to validate whether or not the results of this study are truly indicative of performance is to build a menu structure based on the conceptual

structures identified for the High-Time pilots via each KET. Then a usability evaluation should be conducted during which pilots of various levels of experience (i.e., total number of hours) are asked to perform different tasks using those menu structures. Figure 15.1 provides an illustration of how the representations of the conceptual structures from the MDS analysis on Card Sort data could be transformed into groupings in MFD menu structure.[15] Figure 15.2 provides a simple mock-up of what the MFD prototype interface may look like to support usability testing of menu structure.

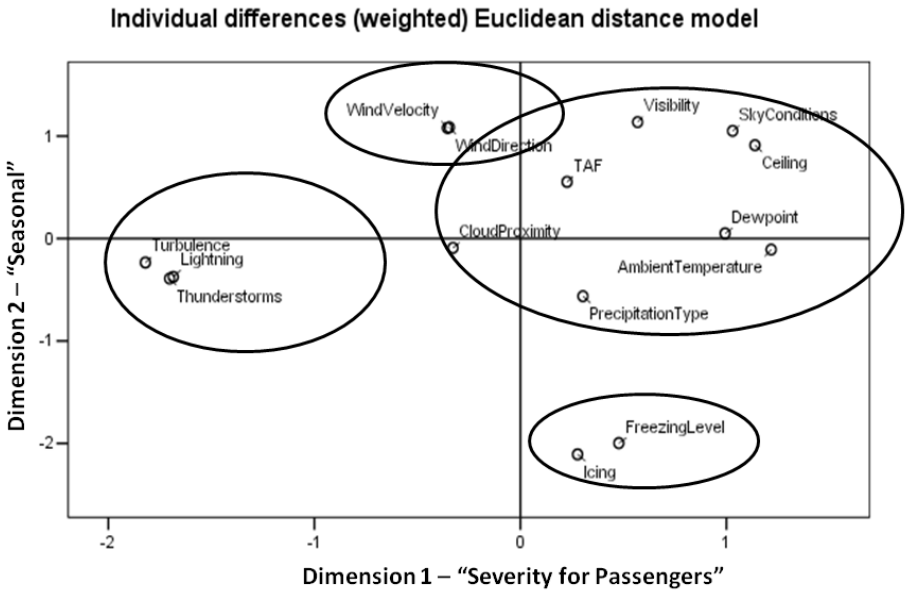


Figure 15.1. Example illustration of how conceptual structure representation from WMDS analysis may be transformed into groupings for MFD menu structure. Example is based on conceptual structures derived by the Card Sort task performed by High-Time Pilots only.

[15] Note that the use of only 15 information concepts in the current study results in a rather small and simplistic menu structure for usability testing (as discussed in the limitations section in Chapter 14). Ideally, future research would also include additional information concepts for deriving pilots' knowledge structures for weather information and then testing the usability of menu structure(s) based on those resulting conceptual structures.

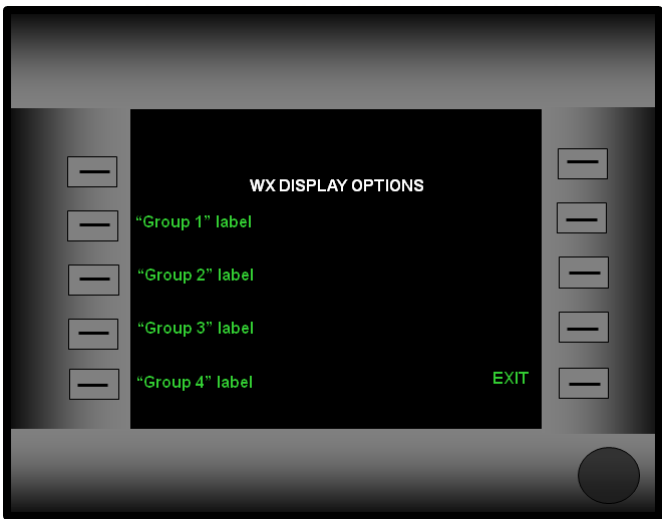


Figure 15.2. Simple visual illustration of what a computer-based MFD prototype interface may look like to support usability testing of menu structure.

In theory, if the conceptual structures from each KET are valid, Card Sort menu structure usability (e.g., time on task, number of steps to complete, satisfaction ratings) would be expected to be best when used by the High-Time pilots since the menu structure organization would have been based on the conceptual structure derived from High-Time pilots which were shown in the current study to differ from the conceptual structures of the other Pilot Experience Groups. Usability for menu structures based on conceptual structures of High-Time pilots from the Relationship Judgment and Prime Recognition Tasks should not differ between pilots of different experience levels because their conceptual structures were not shown to differ based on the results of the current study.

Note that in order to create the menu structures, other techniques would have to be used to resolve the local structure of the relationship between concepts, as MDS is less accurate at depicting the local relationships compared to global relationships. Pathfinder is one technique that could be used (see Chapter 4). Pathfinder analysis may also be of use to further analyze the results of the current study. For example, Pathfinder analysis may be used to understand how the presence or absence of specific links between concepts differs with respect to pilot

experience. This information could be very useful to the design of training curricula and protocols because it could help identify what information to highlight when the desire is to train less experienced pilots to think about weather information in ways analogous to the more experienced pilots. However, limitations noted about the type of number of information concepts may compromise the informativeness of a Pathfinder analysis applied to the current results.

Further Evaluation of the Card Sort Technique

The current study identified the Card Sort as the most valid of the three KETs used to elicit pilots' knowledge structure for weather information. The Card Sort was performed manually, with pilots physically grouping cards on a table, which allowed pilots the maximum flexibility with which to create their groups (unlimited time, ability to revise, ability to create outlier piles, duplicates cards, subgroups, and alternative terminology, etc.). Pilots in the current study did, in fact, take advantage of this flexibility with some Pilot Experience Groups exhibiting some behaviors more than other groups. For example, 50% of the Low-Time pilots in created sub-groups within parent groups, whereas only 25% of Mid-Time and 21% of High-Time pilots created sub-groups. However, more Mid-Time pilots (58%) and High-Time pilots (50%) created duplicate cards than did Low-Time pilots (33%). Further, more Mid-Time pilots identified more concepts as outliers (42%) than did Low-Time pilots (25%) or High-Time pilots (21%). While none of these differences reached statistical significance, it does warrant the need for more research specifically designed to understand whether or not this increased flexibility is necessary to achieve valid conceptual structures, especially when it is desirable to identify differences in conceptual structures as a function of experience level. This increased flexibility comes at a significant cost in terms of the time and resources necessary to collect, code and format the data. Therefore, it is important to understand just how important this flexibility is to the validity of the Card Sort as a knowledge elicitation task.

Also, there are computer programs available that will allow card sort data to be collected electronically and therefore negate the need for a lot of the hand coding and formatting work (see Spencer, 2009 and Chaparro, Hinkle & Riley, 2008 for a partial review of available software programs). In most cases, the software programs try to simulate the look and feel of a manual card sort (e.g., concepts written on graphical representations of “cards” that can be dragged across the screen into categories). While the technology advances have improved the functionality of these computer-based sorts, limitations still may exist in terms of the flexibility with which participants may group their concepts (e.g., inability to create sub-groups, inability to create duplicate cards). Therefore, understanding the impact of that increased flexibility to the overall validity and reliability of card sorting as a knowledge elicitation technique will help practitioners understand the tradeoffs they may be making in terms of data informativeness if they opt for the quicker but less flexible computer-based card sort methodology.

More research is also needed to understand whether computer-based card sorts yield different information about knowledge structures compared to paper-based card sorts. In other words, does the physicality of holding cards in hand and placing them into piles add to the validity of the card sort as an elicitation technique? Further, the comparison between the modalities should be conducted on domain-specific concepts for which there is a logical expectation that participants may differ in the structure of that information based on their level of experience, as that difference in physicality may facilitate more elaborative processing about the relationships between the concepts in the paper-based method. However, if there is no difference between paper and computer-based card sorts, the computerized card sorts could provide a lot of benefit over paper-based card sorts, especially in terms of automating the data collection and formatting. Computerized card sorts could also provide the ability to collect other process-related data about the sorting behavior that would be very difficult if not impossible to collect during a paper-based card sort (e.g., number of times participants revise their groupings,

deliberation time – overall and for placement of each concept within a group, sequences with which the items are grouped, etc.).

Conclusions

Display technology is increasing at very rapid rates. With the advent of touch screen interfaces and increases in software development, the amount and type of information that can be made available to the user is virtually limitless. However, the one component of the user experience that has not changed is basic human processing ability and capacity. The only way to truly leverage these advances in technology is to build the experience with the human in mind. Understanding how the human thinks about the information and identifying the relationships that are important and relevant in a given domain are important first steps to ensuring usable design. Valid knowledge elicitation is key to the success of those first steps.

References

Adelson, B. (1981). Problem solving and the development of abstract categories in programming languages. *Memory & Cognition, 9*, 422-433.

Anderson, J.R. & Bower, G.H. (1973). *Human associative memory*. Washington: Winston and Sons, 1973.

Anderson, J.R. & Lebiere, C. (1998). *The atomic components of thought.* Mahwah, NJ: Erlbaum.

Anderson, J.R. (1974). Retrieval of prepositional information from long-term memory. *Cognitive Psychology, 6*, 451-474.

Anderson, J.R. (1983). *The architecture of cognition.* Cambridge, MA: Harvard University Press.

Arabie, P. (1993). Methodology neither new nor improved. *PsychCRITIQUES, 38*(1), 66-67.

Baxter, G.P., Elder, A.D., & Glaser, R. (1996). Knowledge-based cognition and performance assessment in the science classroom. *Educational Psychologist, 31*(2), 133-140.

Beilock, S.L. & Carr, T.H. (2004). From novice to expert performance: Memory, attention and the control of complex sensori-motor skills. In A.M. Williams & N.J. Hodges (Eds.), Skill *acquisition in sport: Research, theory and practice* (pp. 309-327). New York: Taylor & Francis.

Bijmolt, T.H.A. & Wedel, M. (1995). The effects of alternative methods of collecting similarity data for multidimensional scaling. *International Journal of Research in Marketing, 12*, 363-371.

Borg, I. & Groenen, P. (1997). *Modern multidimensional scaling: Theory and applications.* New York: Springer-Verlag.

Branaghan, R.J. (1990). Pathfinder networks in multidimensional spaces: Relative strengths in representing strong associates. In R.W. Schvaneveldt (Ed). *Pathfinder associative networks: Studies in knowledge organization* (pp. 111-120). Norwood, NJ: Ablex.

Burian, B., Orasanu, J. & Hitt, J. (2000). Weather-related decision errors: Differences across flight types. *Proceedings of the IEA2000 / Human Factors and Ergonomics Society Congress, 22-25.*

Burton, A.M. & Shadbolt, N.R. (1987). Knowledge Engineering. In N. Williams & P. Holt (Eds.), *Expert systems for users.* London: McGraw Hill.

Canas, J.J, Antoli, A. & Quesada, J.F. (2001). The role of working memory on measuring mental models of physical systems. *Psicologica, 22*, 25-42.

Capra, M. (2005). Factor analysis of card sort data: An alternative to Hierarchical Cluster Analysis. *Proceedings of the 49th Annual Meeting of the Human Factors and Ergonomics Society,* 691-695.

Carroll, J. D. & Chang, J. J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. *Psychometrika, 35*, 238-319.

Carroll, J.M. & Olsen, J.R. (1988). Mental models in human-computer interaction. In M. Helander (Ed.), *Handbook of human-computer interaction* (pp. 45-65). North Holland: Elsevier Science Publishers, B.V.

Chaparro, B.S., Hinkle, V.D., & Riley, S.K. (2008). The usability of computerized card sorting: A comparison of three applications by researchers and end users. *Journal of Usability Studies, 4*(1), 31-48.

Chase, W.G. & Simon, H.A. (1973). Perception in chess. *Cognitive Psychology, 4*, 55-81.

Chi, M.T.H., Feltovich, P.J. & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*, 121-152.

Collins, A.M. & Loftus, E. (1975). A spreading activation theory of semantic memory. *Psychological Review, 82*, 407-428.

Collins, A.M. & Quillian, M.R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior, 8*, 240-247.

Cooke, N. M., & Schvaneveldt, R. W. (1988). Effects of computer programming experience on network representations of abstract programming concepts. *International Journal of Man-Machine Studies, 29*, 407-427.

Cooke, N.J. (1994). Varieties of knowledge elicitation techniques. *International Journal of Human-Computer Studies, 41*, 801-849.

Cooke, N.J. (1999). Knowledge elicitation. In F.T.Durso, R.S. Nickerson, R.W. Schvaneveldt, S.T. Dumais, D.S. Lindsay, & M.T.H. Chi (Eds.), *Handbook of applied cognition* (p. 479-509). Chichester: John Wiley.

Cooke, N.M., Durso, F.T., & Schvaneveldt, R.W. (1986). Recall and measures of memory organization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12*, 538-549.

Davidson, M.L. (1983). *Multidimensional scaling*. New York: John Wiley & Sons.

de Groot (1965). *Thought and choices in chess.* The Hague, Netherlands: Mouton.

Dorsey, D.W., Campbell, G.E., Foster, L.L. & Miles, D.E. (1999). Assessing knowledge structures: Relations with experience and posttraining performance. *Human Performance, 12*(1), 31-57.

Falkenstein, M., Yordanova, J., & Kolev, V. (2006). Effects of aging on slowing of motor-response generation. *International Journal of Psychophysiology, 59*(1), 22-29.

Fenker, R.M. (1975). The organization of conceptual materials: A methodology for measuring ideal and actual cognitive structures. *Instructional Science, 4*, 33-57.

Fiore, S.M., Fowlkes, J., Martin-Milham, L., & Oser, R.L. (2000). Convergence or divergence of expert models: On the utility of knowledge structure assessment in training research. *Proceedings of the 44th Annual Meeting of the Human Factors and Ergonomics Society,* 427-430.

Fitts, P.M. & Posner, M.I. (1967). *Human performance.* Belmont, CA: Brooks/Cole.

Gammack, J.G. & Young, R.M. (1985). Psychological techniques for eliciting expert knowledge. In M.A. Bramer (Ed.), *Research and development in expert systems.* Cambridge: Cambridge University Press.

Gentner, D. & Stevens, A. (1983). *Mental models* (pp.1-6). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

George, D. & Mallery, P. (2009). *SPSS for Windows step by step: A simple guide and reference 16.0 Update.* (9th Ed.). Boston: Pearson Education, Inc.

Giguere, G. (2006). Collecting and analyzing data in multidimensional scaling experiments: A guide for psychologists using SPSS. *Tutorials in Quantitative Psychology, 2*(1), 27-38.

Goh, J. & Wiegmann, D.A. (2002). Relating flight experience and pilots' perceptions of decision-making skill. *Proceedings of the 46th Annual Meeting of the Human Factors and Ergonomics Society.*

Goldsmith, T.E. & Johnson, P.J. (1990). A structural assessment of classroom learning. In R.W. Schvaneveldt (Ed.), *Pathfinder associative networks: Studies in knowledge organization* (pp. 241-254). Norwood, NJ: Ablex.

Goldsmith, T.E., Johnson, P.J., & Acton, W.H. (1991). Assessing structural knowledge. *Journal of Educational Psychology, 83*, 88-96.

Gonzalvo, P., Canas, J.J., & Bajo, M.T. (1994). Structural representations in knowledge acquisition. *Journal of Educational Psychology, 86*(4), 601-616.

Graesser, A., Singer, M. & Trabasso, T. (1999). Constructing inferences during narrative text comprehension. In R.J. Sternberg & R.K. Wagner (Eds.), *Readings in Cognitive Psychology* (pp. 318-359). Belmont, CA: Wadsworth Group / Thomson Learning.

Halgren, S.L. & Cooke, N.J. (1993). Towards ecological validity in menu research. *International Journal of Man-Machine Studies, 39*, 51-71.

Hoffman, R. R., Shadbolt, N.R., Burton, A.M., & Klein, G. (1995). Eliciting knowledge from experts: A methodological analysis. *Organizational Behavior and Human Decision Processes, 62*(2), 129-158.

Hoffman, R.R. (1998). Human factors contributions to knowledge elicitation. *Human Factors, 50*(3), 481-488.

Hubert, L.J., Golledge, R.G. & Costanzo, C.M. (1982). Analysis of variance procedures based on a proximity measure between subjects. *Psychological Bulletin, 91*(2), 424-430.

Jaccard, P. (1912). The distribution of flora in the alpine zone. *The New Phytologist, 11*(2), 37-50.

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

Jonassen, D.H., Beissner, K., & Yacci, M. (1993*). Structural knowledge: Techniques for representing, conveying, and acquiring structural knowledge.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Jones, D.G. & Endsley, M.R. (2000). Overcoming representational errors in complex environments. *Human Factors, 42*(3), 367-378.

Jones, L.E. & Young, F.W. (1972). Structure of a social environment: Longitudinal individual differences scaling of an intact group. *Journal of Personality and Social Psychology, 24*(1), 108-121.

Jonsson, J.E. & Ricks, W.R. (1995). *Cognitive models of pilot categorization and prioritization of flight-deck information.* NASA Technical Paper 3528. Hampton, VA: National Aeronautics and Space Administration (NASA) Langley Research Center.

Kellogg, W. A. & Breen, T. J. (1990). Using Pathfinder to evaluate user and system models. In R. W. Schvaneveldt (Ed.), *Pathfinder Associative Networks: Studies in Knowledge Organization* (pp. 179-195). Norwood, NJ: Ablex Publishing Corporation.

Kieras, D.E. & Bovair, S. (1984). The role of a mental model in learning to operate a device. *Cognitive Science, 8*(3), 255-273.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review, 95*, 163-182.

Kraiger, K., Ford, J.K. & Salas, E. (1993). Application of cognitive, skill-based, and affective theories of learning outcomes to new methods of training evaluation. *Journal of Applied Psychology, 78*, 311-328.

Kruskal, J.B. (1964). Multidimensional scaling by optimizing goodness-of-fit to a nonmetric hypothesis. *Psychometrika, 29*(1), 1-27.

Kruskal, J.B. & Wish, M. (1978). *Multidimensional scaling.* Beverly Hills, CA: Sage.

LaFrance, M. (1992). Excavation, capture, collection, and creation: Computer scientists' metaphors for eliciting human expertise. *Metaphor and Symbolic Activity, 7*, 135-156.

Latorella, K.A., Lane, S., & Garland, D. (2002). *General aviation pilots' perceived usage and valuation of aviation weather information sources.* NASA Technical Memorandum 2002-211443. Hampton, VA: National Aeronautics and Space Administration.

Latorella, K.A., Pliske, R., Hutton, R. & Chrenka, J. (2001). *Cognitive task analysis of business jet pilots' weather flying behaviors: Preliminary results.* NASA Technical Memorandum 2001-211034. Hampton, VA: National Aeronautics and Space Administration.

Legandre, P. & Legandre, L. (1998). *Numerical ecology* (2nd ed.). Amsterdam: Elsevier.

Lewis, S. (1991). Cluster analysis as a technique to guide interface design. *International Journal of Man-Machine Studies, 35*, 251-265.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research, 27*, 209-220.

McDonald, J. E., Dearholt, D. W., Paap, K. R., & Schvaneveldt, R. W. (1986). A formal interface design methodology based on user knowledge. *Proceedings of Computer Human Iinteraction (CHI)*, 285-290.

McDonald, J., Stone, J.D., Liebelt, L.S., & Karat, J. (1982). Evaluation of methods for structuring the user system interface. *Proceedings of the 41st Annual Meeting of the Human Factors Society,* 551-554.

McDonald, J.E., Stone, J.D., & Liebelt, L.S. (1983). Searching for items in menus: The effects of organization and type of target. *Proceedings of the 27th Annual Meeting of the Human Factors Society.*

Mertler, C.A. & Vannatta, R.A. (2001). *Advanced and multivariate statistical methods: Practical application and interpretation.* Los Angeles, CA: Pyrczak Publishing.

Miller, G.A. (1969). A psychological method to investigate verbal concepts. *Journal of Mathematical Psychology, 6*, 169-191.

Navarro-Prieto, R. & Canas, J.J. (2001). Are visual programming languages better? The role of imagery in program comprehension. *International Journal of Human-Computer Studies, 56*(6), 799-829.

Norman, D.A. (1983). Some observations on mental models. In D. Gentner & A.L. Stevens (Eds.), *Mental models* (pp. 7-14). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Norman, D.A. (1988). *The psychology of everyday things.* New York: Harper & Row.

O'Hare, D., Owens, D., & Wiegmann, D. (2001). The "where" and "why" of cross-country VFR crashes: Database and simulation analyses. *Proceedings of the 45th Annual Meeting of the Human Factors and Ergonomics Society*.

Parkinson, S.R., Sisson, N. & Snowberry, K. (1985). Organization of broad computer menu displays. *International Journal of Man-Machine Studies, 23*, 689-697.

Pennington, N. (1987). Stimulus structures and mental representation in expert computer programmers. *Cognitive Psychology, 19*, 295-341.

Preece, P.F.W. (1976). Mapping cognitive structure: A comparison of methods. *Journal of Educational Psychology, 68*(1), 1-8.

Rasmussen, J. (1983). Skills, rules, and knowledge; Signals, signs, and symbols, and other distinctions in human performance models. *IEEE Transactions on Systems, Man and Cybernetics, SMC* 13, 257-266.

Reitman, J.S. (1976). Skill perception in Go: Deducing memory structure from inter-response times. *Cognitive Psychology, 8*, 336-356.

Richard, C.M., Kleiss, J.A., & Bittner, A.C. (2004). Comparison of cluster analysis and structural analysis methods for identifying user mental models: An integrated in-vehicle telematics systems illustration. In *Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting* (pp. 2471-2475).

Rickheit, G. & Sichelschmidt, L. (1999). Mental models: Some answers, some questions, some suggestions. In G. Rickheit & C. Habel (Eds.) *Mental models in discourse processing and reasoning.* (pp. 9-40). New York, NY: North-Holland Elsevier Science B.V.

Rietman, J.S. & Rueter, H.R. (1980). Organization revealed by recall orders and confirmed by pauses. *Cognitive Psychology, 12*, 554-581.

Rips, L.J., Shoben, E.J. & Smith, E.E. (1973). Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior, 12*, 1-20.

Roske-Hofstrand, R.J. & Paap, K.R. (1986). Cognitive networks as a guide to menu organization: An application in the automated cockpit. *Ergonomics, 29*(11), 1301-1311.

Roske-Hofstrand, R.J. & Paap, K.R. (1990). Discriminating between degrees of low and high similarity: Implications for scaling techniques using semantic judgments. In R.W. Schvaneveldt (Ed.), *Pathfinder associated networks: Studies in knowledge organization* (pp. 61-73). Norwood, NJ: Ablex Publishing.

Rowe, A.L, Cook, N.J., Hall, E.P., & Halgren, T.L. (1996). Toward an on-line knowledge assessment methodology: Building on the relationship between knowing and doing. *Journal of Experimental Psychology: Applied, 2*(1), 31-47.

Rumelhart, D. E. & Norman, D.A. (1985). Representation of knowledge. In A.M. Aitkenhead & J.M. Slack (Eds.), *Issues in cognitive modeling (pp. 15-62).* Hillsdale, NJ: Lawrence Erlbaum & Associates, Inc.

Rumelhart, D.E. & Abrahamson, A.A. (1973). A model for analogical reasoning. *Cognitive Psychology, 5*, 1-28.

Salmeron, L., Canas, J.J. & Fajardo, I. (2005). Are expert users always better searchers? Interaction of expertise and semantic grouping in hypertext search tasks. *Behaviour & Information Technology, 24*(6), 471-475.

Schiffman, S. S., Reynolds, M. L. & Young, F. W. (1981). *Introduction to multidimensional scaling.* New York: Academic Press.

Schneider, W. (2000). *E Prime* (Beta Version 5.0) [Computer software]. Pittsburgh: Psychological Software Tools, Inc.

Schvaneveldt, R.W. (1990). *Pathfinder associative networks: Studies in knowledge organization.* Norwood, NJ: Ablex Publishing.

Schvaneveldt, R.W., Beringer, D.B., Lamonica, J., Tucker, R., & Nance, C. (2000*). Priorities, organization, and sources of information accessed by pilots in various phases of flight.* DOT/FAA/AM/00-26. Washington, D.C.: Office of Aviation Medicine.

Schvaneveldt, R.W., Durso, F.T., Goldsmith, T.E., Breen, T.J., & Cooke, N.M. (1985). Measuring the structure of expertise. *International Journal of Man-Machine Studies, 23*, 699-728.

Shavelson, R.J. (1972). Some aspects of the correspondence between content structure and cognitive structure in physics instruction. *Journal of Educational Psychology, 63*, 225-234.

Sokal, R.R., & Rohlf, F.J. (1995). *Biometry: The principles and practice of statistics in biological research* (3rd ed.). New York: Freeman.

Spence, I. & Domoney, D. W. (1974). Single subject incomplete designs for nonmetric multidimensional scaling, *Psychometrika, 39*, 469-470.

Spencer, D. (2009). *Card sorting: Designing usable categories.* Brooklyn, NY: Rosenfeld Media.

Takane, Y., Young, F.W. & de Leeuw, J. (1977). Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features. *Psychometrika, 42*, 7-67.

Torgerson, W. S. (1952). Multidimensional scaling: I. Theory and method. *Psychometrika, 17*(4), p.401-419.

Trafton, J.G. (2004). Dynamic mental models in weather forecasting. *Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting,* 311-314.

Valero, P. & Sanmartin, J. (1999). Methods for defining user groups and user-adjusted information structures. *Behaviour & Information Technology, 18* (4), 245-259.

Van der Veer, G.C. & Melguizo, M.C.P. (2002). Mental models (pp. 52-80). In J.A. Jacko & A. Sears (Eds.), *The human-computer interaction handbook: Fundamentals, evolving technologies, and emerging applications* (pp. 52-80). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Vigeant-Langlois, L. & Hansman, R.J. (2002). Trajectory-based performance assessment of aviation weather information. *10th AMS Conference on Aviation, Range, and Aerospace Meteorology.*

Whitener, E.M. & Brodt, S.E. (1994). When is the “ko” ok? Capitalizing on existing knowledge structures to facilitate pretraining transfer. *Human Resource Management Review, 4*, 363-381.

Wickens, C.D. & Hollands, J.G. (2000). *Engineering psychology and human performance* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Wickens, C.D., Gordon, S.E., & Liu, Y. (1998). *An introduction to human factors engineering.* New York: Addison Wesley Longman.

Wiggins, M. W., & O'Hare D. (1995). Expertise in aeronautical weather-related decision-making: A cross-sectional analysis of general aviation pilots. *Journal of Experimental Psychology: Applied, 1*, 305-320.

Williams, K.W. & Joseph, K.M. (1998). Developing data link user-interface designs using pilot conceptual networks. In *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting* (pp. 77-81). Santa Monica, CA: The Human Factors and Ergonomics Society.

XLSTAT (Version 2010.5.03) [Computer software]. New York, NY: Addinsoft.

Yates, J. (1985). The content of awareness is a model of the world. *Psychological Review, 92*(2), 249-284.

Ye, N. & Salvendy, G. (1994). Quantitative and qualitative differences between experts and novices in chunking computer software knowledge. *International Journal of Human-Computer Interaction, 6*, 105-118.

Ye, N. (1997). Objective and consistent analysis of group differences in knowledge representation. *International Journal of Cognitive Ergonomics, 1* (2), 169-187.

Ye, N. (1998). The MDS-ANAVA technique for assessing knowledge representation differences between skill groups. *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans, 28*(5), 586-600.

Young, F. W. (1985). Multidimensional scaling. In Kotz, S., Johnson, N. L., & Read, C. B. (Eds), *Encyclopedia of Statistical Sciences*, Volume 5 (pp. 649-659). Wiley, New York, NY.

Young, F.W. & Hamer, R.M. (1987). *Multidimensional scaling: History, theory and applications.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Appendix A - Forward Scenario Simulation Materials

Scenario #1

Aircraft:	Beechcraft King Air 200 •Turboprop with de-cing equipment; can be flown by single pilot, but usually flown by 2
Pilot(s):	2 pilots onboard
Departure APT:	Boca Raton, FL (BCT)
Destination APT:	Atlanta Peachtree Dekalb (PDK)
Time Enroute:	2 hrs 10 min
Departure Time:	1:00 p.m. EST
Flight Plan:	IFR
Wx Conditions:	•Convective Activity over the state (clear off-shore) •Moderate to severe turbulence between 5,000ft and 20,000ft
Route of flight:	Up the middle of the state (over Orlando)

- 10 minutes after takeoff, you are cleared to climb to 7000ft where you experience moderate to severe turbulence. You make repeated calls to Miami Center to ask for higher but you are either ignored or are unable to make contact for some reason. The aircraft is very difficult to control (even with both pilots' hands on the controls). The aircraft is flying in and through cumulonimbus clouds (no icing) and passengers are starting to panic.

Figure A.1. Description of Scenario #1 presented to pilots as part of the Forward Scenario Simulation conducted during Phase I data collection.

Scenario #2

Aircraft:	Cessna 172 •Single engine reciprocating aircraft with no anti-ice or de-icing equipment
Pilot(s):	Single pilot
Departure APT:	Daytona Beach, FL (DAB)
Destination APT:	Bartow Municipal (BOW)
Time Enroute:	1 hr
Departure Time:	5:00 p.m. EST
Flight Plan:	VFR
Wx Conditions:	•Skies 2000ft overcast decreasing to marginal VFR along the route of flight

- 30 minutes after take-off you have descended to 1000ft and are scud-running. You've checked and there aren't any obstacles between you and your destination above 500ft, but you're slowly descending with the overcast cloud layer.

Figure A.2. Description of Scenario #2 presented to pilots as part of the Forward Scenario Simulation conducted during Phase I data collection.

Scenario #3

Aircraft:	Cessna 421 •Twin engine reciprocating aircraft with no anti-ice or de-icing equipment
Pilot(s):	Single pilot
Departure APT:	Great Barrington, MA (GBR)
Destination APT:	Charlotte, NC (CLT)
Time Enroute:	4 hrs
Departure Time:	10:00 a.m.
Flight Plan:	IFR
Wx Conditions:	•Icing between 8000ft and 18000ft

- You file an IFR flight plan at 6000ft. There are forecast layers between 3000 and 5000ft and 8000 and 18000ft. It is clear above FL180 but you do not have oxygen equipment aboard so you cannot go above 14000ft. The ceiling is at 3000ft
- After departure you climb to 6000ft and you're still in the clouds. The PIREPs have reported that there are breaks in the clouds near your location but you are starting to pick up ice on the leading edge of the wings.

Figure A.3. Description of Scenario #3 presented to pilots as part of the Forward Scenario Simulation conducted during Phase I data collection.

Scenario #4

Aircraft:	Beechcraft Bonanza •Single engine reciprocating aircraft with no anti-ice or de-icing equipment
Pilot(s):	Single Pilot
Departure APT:	Destin – Ft. Walton Beach (DTS)
Destination APT:	Hollywood (N. Perry), FL (HWO)
Time Enroute:	3 hrs, 30min
Departure Time:	7:00 a.m.
Flight Plan:	None (although the pilot is IFR-rated)
Wx Conditions:	•Marginal VFR •Ceiling at 1000ft and visibility is 1mi in rain and mist. •There is a Low pressure over the Gulf of Mexico. •Weather is forecast to be raining to severe thunderstorms along your route of flight until after you pass Sarasota, FL (approx 300mi).

- The route of flight you have planned is over the Gulf of Mexico. 30 minute after departure, you are over the water at 800ft to maintain cloud clearance and you are now too low to receive Departure or Center Control. If you climb to receive the flight service station or any other VHF frequency to get an IFR clearance, you will enter the clouds.

Figure A.4. Description of Scenario #4 presented to pilots as part of the Forward Scenario Simulation conducted during Phase I data collection.

Appendix B - Assessment of Pilot Experience Group Performance in each KET

Raw data collected from each of the three KETs were submitted to an initial analysis that served two major purposes. First the analyses helped determine the best approaches to coding and/or trimming the data. Second, the analyses were used to explore whether the pilot experience groups differed in their general ability to perform the KETs. Note: Because of the exploratory nature of this study and the small sample sizes, no adjustments have been made for family-wise error in any of the following analyses.

Relationship Judgment

Participants were asked to rate the similarity of 105 concept pairs on a 1-9 scale (1=not similar, 9= very similar). These data were recoded to represent dissimilarity judgments (dissimilarity = 10-similarity).

Results

Dissimilarity Ratings

The dissimilarity ratings for each item pair were averaged for each participant. Analysis revealed a marginally significant negative correlation between total hours flown and average dissimilarity rating (r = -.41, p<.10). Pilots with less hours flown tended to have higher average dissimilarity ratings.

A one-factor (Pilot Experience Group) ANOVA was conducted on the averaged dissimilarity data. The effect of Pilot Experience Group was not significant (p >.05). Dissimilarity judgments for Low-Time Pilots tended to be higher (M=5.06) than Mid-Time Pilots (M=4.73) or High-Time Pilots (M=4.38) but the difference did not reach statistical significance (see Figure

B.1). Also note that the High-Time pilots showed much less variability in their Relationship Judgments (*SD*=0.45) compared to Low-Time pilots (*SD*=1.24) and Mid-Time pilots (*SD*=0.8).

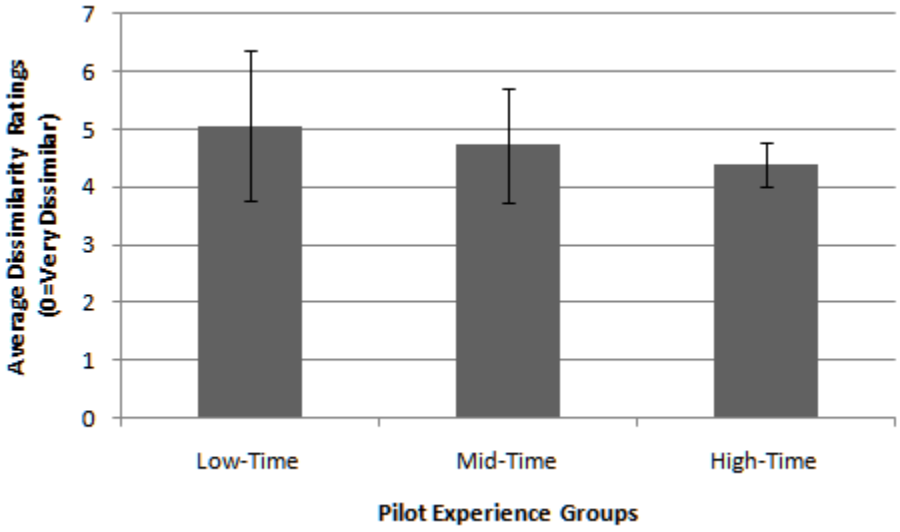


Figure B.1. Effect of Pilot Experience on Relationship Judgment dissimilarity ratings (95% confidence intervals depicted).

Response times

A one-factor (Pilot Experience Group) ANOVA was conducted on the averaged time it took each pilot to make the dissimilarity ratings. The effect of Pilot Experience Group was not significant (*p*=*n.s.*). Mid-Time pilots took an average of almost 2 sec longer to make their ratings (*M*=8087 ms) than did High-Time pilots (*M*=6128 ms) and almost 1.5 sec longer than Low-Time pilots (*M*=6664 ms) but the differences did not reach statistical significance. It is important to note, however, that pilots were told to focus on making their judgments and were not explicitly told to make their ratings as quickly as possible. Therefore, there was a lot of variability in the time pilots took to make their ratings and any interpretation of the response time data for relationship judgments should be made with extreme caution.

Discussion

Pilots with fewer hours flown tended to rate weather concept pairs as more dissimilar than pilots with more experience, although the negative correlation was only marginally significant and the differences between the Pilot Experience Groups failed to reach statistical significance. High-Time pilots did show more overall consistency in their Relationships Judgments, which would be expected if knowledge structures become more efficiently organized with experience. Mid-time pilots took an average of 2 sec longer to make their ratings than the High-Time pilots and 1.5 sec longer than the Low-Time pilots. This difference, while not statistically significant, could provide additional evidence to suggest that Low-time pilots were tapping in to declarative knowledge that was more accessible in memory for them than it was for Mid-Time pilots, and High-Time pilots were trying to tap into the a knowledge structure that, though extensive knowledge and experience, has become much more organized than the Mid-Time pilots, allowing them to make their ratings more quickly.

Prime Recognition Task

The Prime Recognition Task presented the participant with a memory set of four concepts on one screen. After a short duration, the screen was replaced with the prime (another weather concept) followed quickly by the target (in red). Participants were told to answer whether or not the target word was part of the memory set. The pilot was told that the prime was a 5^{th} memory set item that could not fit correctly on the screen with the first four concepts.

Target trials were defined as trials when the target was part of the memory set and Foil trials were defined as trials when the target was <u>not</u> part of the memory set. There were 120 total Target trials, comprised of 105 pairings of the 15 weather-related concepts and 15 additional trials in which the probe and/or prime consisted of an acronym. There were 79 Foil trials. Mixed repeated-measures ANOVAs were conducted with Pilot Experience Group (3

levels) as the between-subjects variable and Trial Type (Target or Foil) as the within-subjects variable on both accuracy and response time results.

Results

Accuracy Results

Overall, pilots were very accurate in their responses, achieving over 94% correct across both Target and Foil trials. The effect of Trial Type was significant [$F(1,21) = 25.75$, $p \leq .05$, *partial* $\eta^2 = .55$]. Pilots were more accurate in recognizing when the target was part of the memory set (M=96%) than recognizing when it was not part of the memory set (M=91%). There was no main effect of Pilot Experience Group nor was there any interaction between Pilot Experience Group and Trial Type (p=*n.s.*). Low-Time pilots more accurate in their judgments on Foil trials (M=94%) compared to High-Time pilots (M=90%) and Mid-Time pilots (M=88%) but the difference between the groups failed to reach statistical significance (see Figure B.2). Pilot experience had no effect on accuracy when the analysis was restricted only to the 105 Target trials (p>0.5).

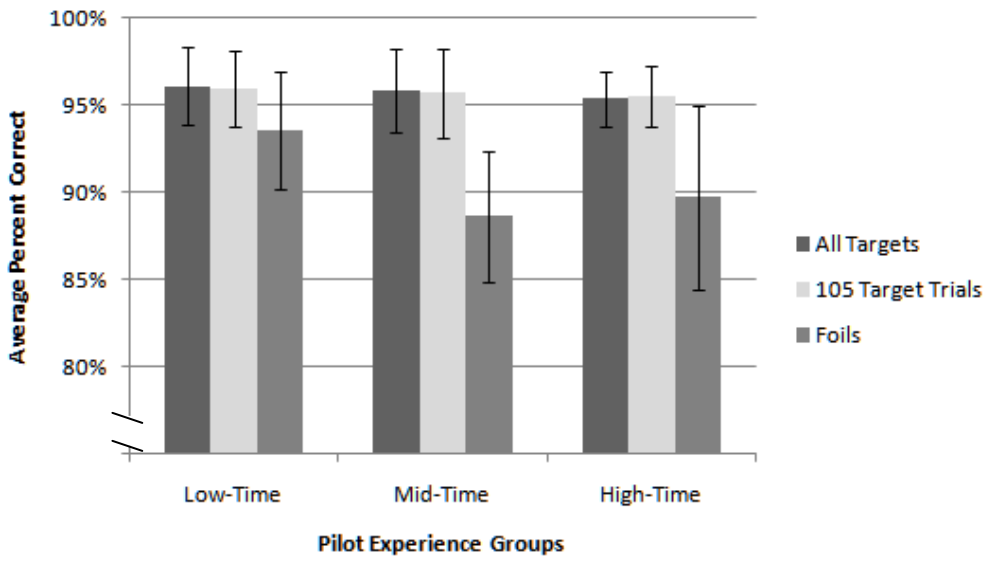


Figure B.2. The effect of Pilot Experience Group on accuracy for Target and Foil trials (with 95% confidence intervals).

Response Time Results

Only response times for correct trials were included in the response time analysis. Correlation analysis revealed no correlation between Total Hours Flown and average Response Time (*p*=*n.s.*).

The average Response Time across both Target and Foil trials was 922.4 msec (*SD*=171.1). The effect of Trial Type was significant [$F(1,21) = 47.32$, $p \leq .05$, *partial* $\eta^2 = .69$]. Pilots were quicker to recognize when the target was part of the memory set (*M*=860.9 ms) than when it was not part of the memory set (*M*=1023.2 ms). There was no significant effect of Pilot Experience Group nor was there a significant interaction between Pilot Experience Group and Target Type (*p*=*n.s.*). Low-Time pilots tended to have quicker response times for both Target and Foil trials, but the difference between the Experience Groups was not significant (Figure B.3). Pilot Experience Group also had no effect on Response Time when the analysis was restricted only to the 105 Target trials (*p*=n.s.). Again, Low-Time pilots were quickest to respond in the 105 target trials but the difference failed to reach statistical significance (*p*=*n.s.*).

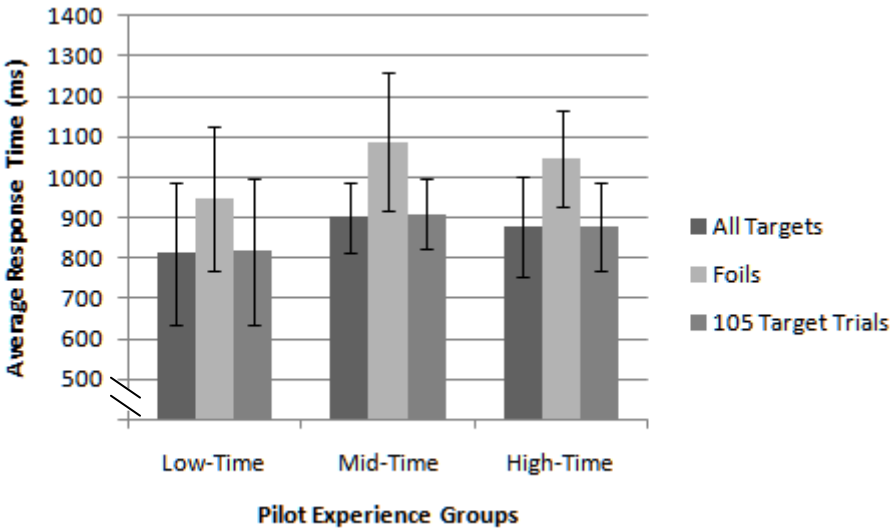


Figure B.3. The effect of Pilot Experience Group on response times for Target and Foil Trials (with 95% confidence intervals depicted).

Age may have played a role in the quickness with which one is able to perform this task. On average, Low-Time pilots were 14 years younger than Mid-Time pilots and Mid-Time pilots were, on average, 10 years younger than High-Time pilots. Response time did tend to increase between Low-Time and Mid-Time pilots, although the difference was not statistically significant. However, response time was no different between Mid-Time and High-Time pilots so perhaps there is a threshold beyond which younger age is no longer a facilitator to the task.

Upon further examination of the dataset, it became apparent that one Low-Time pilot (#54) had a significantly higher average response time (*M*=1376 ms) than two standard deviations above the mean for the Low-Time group (outer-fence = 1279 ms). None of the Mid-Time or High-Time pilots had mean response times greater than 2 standard deviations above their respective group means. Low-Time pilot #54 was not an outlier in terms of any demographic characteristic (i.e., he was 20 years old and had 133 total hours) but his response times for most of the 105 item pairs were longer than any other Low-Time pilot. Pilot #54 was removed from the Low-Time group dataset and the mixed repeated measures ANOVA was conducted once again (with the Low-Time group consisting of 8 participants instead of 9) to examine the effect of this outlier.

With pilot #54 removed, the effect of Trial Type was still significant, [$F(1,20) = 46.97$, $p \leq .05$, *partial* $\eta^2 = .70$], with pilots quicker to recognize when the target was part of the memory set (*M*=839 ms) than when it was not part of the memory set (*M*=1006 ms). The effect of Pilot Experience Group was significant, $F(2,20) = 4.19$, $p \leq .05$, *partial* $\eta^2 = .30$) but there was no interaction between Pilot Experience Group and Trial Type. Post-Hoc Tukey HSD tests revealed that Low-Time pilots were significantly quicker to respond to Target Trials and Foil Trials than Mid-Time pilots ($p \leq .05$). The difference between Low-Time and High-Time pilots was marginally significant ($p \leq .10$). The difference between Mid-Time and High-Time pilots was not significant

(*p=n.s.*). The correlation between total number of hours and average response time was still not significant even with the removal of participant #54 (*p=n.s.*).

Even though pilot #54 was more than 2 standard deviations slower than the rest of the Low-Time pilots in judging whether the target was in the memory set, it was decided to include this pilot in the rest of the data analysis because the pilot's slow response time was relatively consistent across the 105 target trials and therefore the correspondence between response time and relationships between information concepts should be maintained. In other words, because this pilot was slower to respond across *all* trials, prime-target pairs that are highly related in his memory should still be expected to have quicker response times (relatively speaking) than prime-target pairs that are not related.

In sum, pilots were more accurate in recognizing when the target was in the memory set than they were in recognizing when the target was not part of the memory set. Low-Time Pilots were the most accurate in recognizing when the Target was not part of the memory set (i.e., Foil trials), although the difference among the experience groups was not statistically significant. Accuracy on Target trials did not differ across Pilot Experience Group. Low-Time pilots were quicker to respond whether the target was part of the memory set but because the Prime Recognition Task relies on quick motor response, this quicker response time may have been partially due to the overall younger age of the Low-Time group rather than anything having to do with their knowledge structure organization.

Card Sort

Pilots were asked to organize 15 concepts (each presented on a card) into groups that make sense to them. They were allowed to create duplicate cards and to create hierarchies of groups when necessary. Jaccard scoring was used to represent the groups created through the

Card Sort. Jaccard scoring[16] ranges from 0 (meaning items were never placed together in a group) to 1 (meaning two items were always placed in the same group). Jaccard scoring accounts for hierarchal groups in that it places different weights on items depending on whether they occur within the same group or within different subgroups under the same parent group. Items occurring in different subgroups under the same parent group are given scores greater than 0 (meaning there is some relationship between the items) but less than 1 (meaning they did not occur in the same immediate group). Another way to examine Card Sort data is to look at the total number of links participants created as a function of their card sort. Of course, this metric should be highly correlated with the Jaccard similarity scores.

Total Hours Flown was not significantly correlated with the average Jaccard scores (r = -.08, p=*n.s.*) or the average number of links created between concepts (r = -.15, p=*n.s.*). As expected, Jaccard similarity scores were highly correlated with the average number of links created by pilots (r=.78, $p \leq .05$). Participants who created a lot of links within their card sorts resulted in higher similarity scores.

Results

Jaccard Similarity Score Results

A one-factor (Pilot Experience Group) ANOVA was conducted on the averaged Jaccard Similarity scores. The effect of Pilot Experience Group was not significant (p=*n.s.*). As Figure B.4 indicates, similarity scores tended to be higher for Mid-Time pilots (M=.35) compared to Low-Time (M=.27) or High-Time pilots (M=.28) but the difference did not reach statistical significance (p=*n.s.*).

[16] Note that Jaccard scores were maintained as similarity scores for this analysis rather than dissimilarity scores for the other parts of this study to make this initial comparison and interpretation simpler within the context of Number of Links.

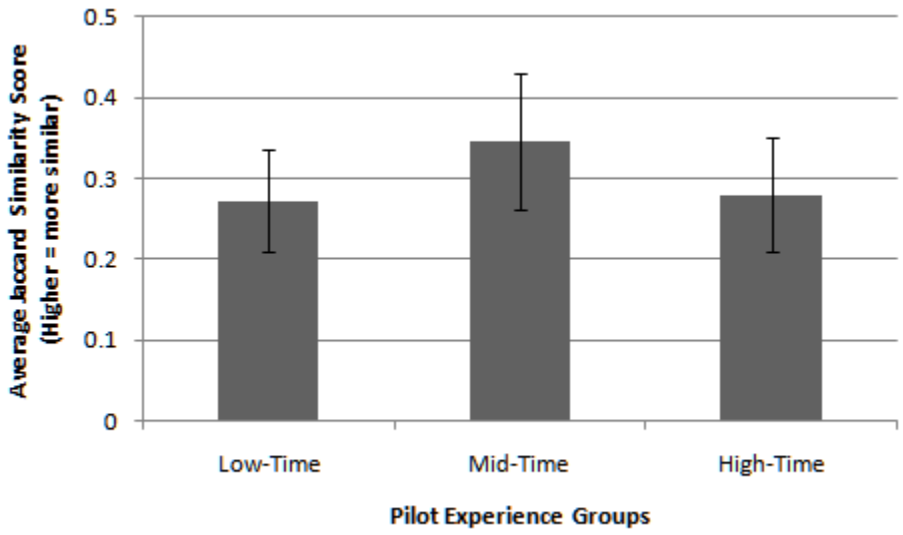


Figure B.4. Average Jaccard Similarity Score as a function of Pilot Experience Group (95% confidence intervals depicted).

Number of Links Results

Mid-Time pilots had more average links (*M*=47) in their Card Sorts than did Low-Time pilots (*M*=40) or High-Time pilots (*M*=37), although the difference between the three groups was not statistically significant (*p*=*n.s.*). To give some context to these numbers, if all 15 concepts were to be placed into the same group (meaning they would all be equally related), that Card Sort would result in the creation of 105 links. However, note that 105 does not represent the maximum number of links possible. The creation of duplicate cards would result in more than 105 links. The average number of links for each group does indeed include some duplicate cards. However, if 105 was considered a theoretical maximum number of links, then the average Card Sort for Mid-Time Pilots contained approximately 44% of the possible links between concepts, Low-Time pilot Card Sorts contained approximately 38% of the possible links, and High-Time pilot Card Sorts contained approximately 35% of the possible links. The fact that Mid-Time pilot Card Sorts had the most links was not unexpected, given that Mid-Time pilots tended to have higher Jaccard similarity scores as well.

Card Sort Behaviors

Several different Card Sort behaviors were identified, inventoried and compared across the three Pilot Experience Groups. Many of these group differences failed to reach statistical significance (evaluated at an alpha level of .05), most likely because of the small sample size. However, they do suggest potential trends in behavior. Table B.1 provides a summary of these behaviors across the three groups.

Subgroups. Half (50%) of the Low-Time pilots created at least one subgroup, compared to 25% of Mid-Time pilots and 25% of High-Time pilots. However, the difference between the groups was not statistically significant. Of those pilots who did create subgroups, Card Sorts created by Low-Time pilots had more subgroups (M=3.5), on average, than did Card Sorts created by Mid-Time (M=2) or High-Time (M=2.3) pilots. Again, however, the difference between the groups was not statistically significant.

Duplicate cards. More than half of the Mid-Time (58%) and High-Time pilots (50%) created duplicate cards of at least one of the 15 original concepts. Only 33% of Low-Time pilots created duplicate cards. While the difference between the groups was not statistically significant, duplicate cards appeared to be more necessary for the experienced pilots to properly represent the relationship between all of the concepts within the sort. The number of duplicate cards created did not vary with Pilot experience (see Table B.1).

Parent Groups. Card Sorts typically consisted of 3-4 parent groups, regardless of Pilot Experience Level. Not surprisingly, there was a significant effect of Subgroup Creation on the number of parent groups created, $F(1,32) = 4.74$, $p \leq .05$, *partial* $\eta^2 = .13$. Those who did not create subgroups tended to create more Parent groups (M=4) than did those people who created subgroups (M=2.9). Thus, pilots who did not create subgroups necessarily created more shallow but broader structures with their cards. Those who used subgroups tended to create deeper but narrower structures with their cards.

There was also a significant Pilot Experience Group x Duplicate card creation interaction, $F(2,32) = 3.93$, $p \leq .05$. High-time pilots who created duplicates cards typically created more Parent groups (M=4.7) than those who did not create duplicate cards (M=3.1). Low-time pilots who created duplicate cards typically created fewer Parent groups (M=2.5) than did those who did not create duplicate cards (M=4.0). Mid-Time pilots who created duplicate cards differed only slightly in the amount of Parent groups they created (M=3.6) compared to those who did not create duplicate cards (M=3.0).

Outliers. More Mid-time pilots created an outlier pile (42%) than Low-Time (25%) or High-Time pilots (29%). However, the difference between the groups was not statistically significant.

Created cards with new concepts in addition to original 15 concepts. Only pilots with higher levels of experience felt they needed to create additional concepts not already included in the original 15 to adequately express how the items are related (8% of Mid-Time pilots and 21% of High-Time pilots). The difference between the groups was not significant (p=n.s.).

Added description to original/duplicated concepts. Some pilots added descriptors or extra words to how the original concepts were worded. Most of the time it was to make a distinction between Enroute or Ground weather concepts (e.g., adding “Enroute” to the “Wind Speed” concept to distinguish it from “Ground” Wind Speed) or between Current or Forecasted weather (e.g., adding “Current” to the “Icing” concept to distinguish it from “Forecasted”). While all pilots were told they could change the wording on the cards if necessary, only a few pilots did this and all of them were either Mid-Time (17%) or High-Time pilots (14%).

Table B.1. *Summary table of various Card Sort Behaviors across the three Pilot Experience Groups.*

Card Sort Behaviors	**Pilot Experience Groups**		
	Low-Time	*Mid-Time*	*High-Time*
Created Subgroups	50%	25%	21%
Mean number of Subgroups (those that created them)	3.5	2	2.3
Created duplicate cards	33%	58%	50%
Ave number of duplicate cards (those that created them)	3.75	4.27	4
Ave number of Parent groups created	3.5	3.33	3.92
Created Outlier pile	25%	42%	29%
Add concepts in addition to original 15 concepts	0%	8%	21%
Added additional description to original and/or duplicated cards (e.g., added "surface" to Wind Direction card)	0%	17%	14%

Discussion

Pilot experience had no significant effect on Jaccard similarity scores, although Mid-Time pilots' scores tended to be slightly higher than the other pilot groups. Similarly, Mid-Time pilots had the most average links and were more likely to create duplicate cards in their Card Sorts although the differences were not statistically significant. These findings do suggest that Mid-Time pilots have more complex (or perhaps even less organized) knowledge structures than Low-Time pilots who may have relied on information memorized from textbooks and classes to make their groupings. However, Mid-Time pilots may not have the extent of experience that the High-Time pilots have to be able to focus on the more important relationships between items that allowed the High-Time pilots to generate less complex Card Sorts. Instead, Mid-Time pilots tended to build extra links between concepts (or create more duplicate cards) to make sure they "covered their tracks," so to speak, in representing the important relationships between concepts. Low-Time and High-Time pilots may be been more judicious and sparing with their groupings because Low-Time pilots relied on their more rehearsed declarative knowledge and High-Time pilots relied on their vast amounts of practical experience.

Appendix C - Preparing Card Sort Data for Analysis

This appendix describes the procedures for transforming the physical groups of cards that pilots created into a quantitative measure of similarity. Two procedures were used: Traditional data coding and Jaccard similarity scoring.

Traditional Data Coding

Figure C.1 shows an example of what one pilot's groupings looked like after the 15-card card sort was complete. Items in pink represent the labels created and applied to the groups by the pilot. This pilot created three groups (VFR, TAF, and IFR) with IFR having two subgroups (Icing Related and Thunderstorm-related). Cloud Proximity was identified as an outlier, and the "Ceiling" card was duplicated so that it could be represented in both the VFR and IFR categories.

An excel spreadsheet was created with all 105 pairs of concepts listed in a matrix form. The 15 concepts were listed across the top of the matrix and along the side, so that the diagonal shows each concept crossed with each other concept. Data were assumed to be symmetric, so only the bottom portion of the matrix was completed. Each of the 105 pairs was assigned a "1" if the two concepts occurred in the same group together or a "0" if they did not. For example, the Sky Conditions – Visibility pair was assigned a "1" and the Wind Direction - Precipitation Type pair was assigned a "0". Any pair that involved Cloud Proximity was assigned a "0" (because it was designated an outlier) as was any pair that involved TAF (because it placed in a group unto itself). Item pairs within the IFR category were assigned a "1" even if they occurred in different subgroups. For example, the Dewpoint-Icing pair was assigned a "1" because they both occurred under "Icing Related." The Dewpoint-Precipitation Type pair was also assigned a "1" because they both occurred under "IFR" even though they were in different subgroups.

Thus, the subgroups created by pilots were “flattened” to simply the coding and analysis. Data were then formatted into similarity matrices.

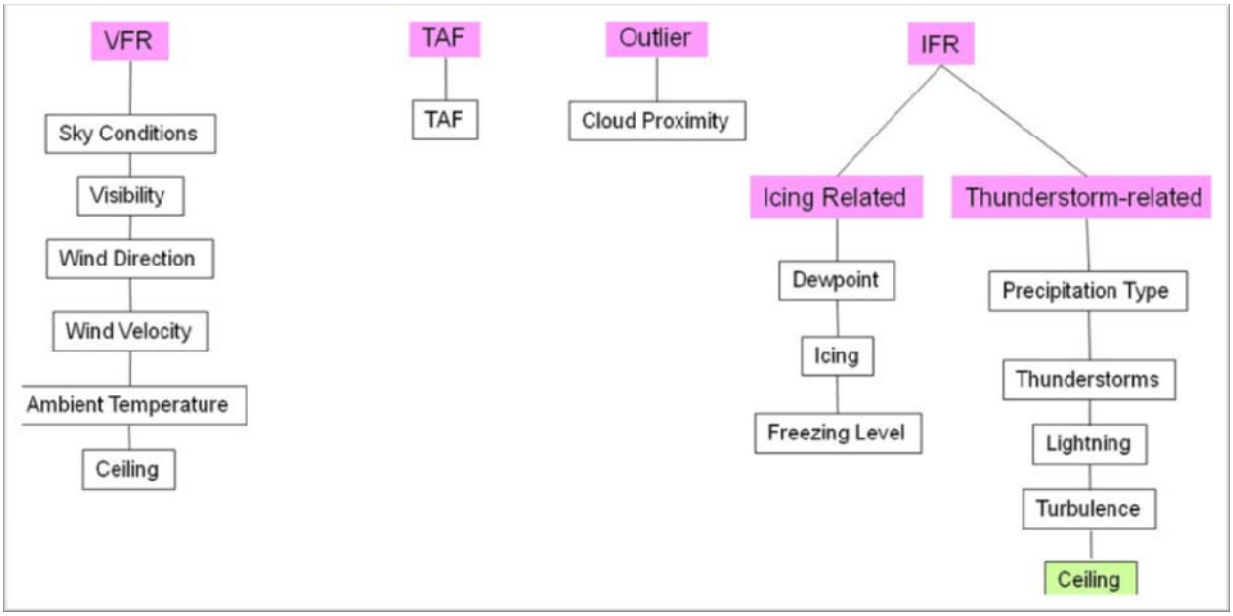


Figure C.1. Example shows an example of what one participant’s groupings looked like after the 15-card card sort was complete. Items in pink represent the labels created and applied to the groups by the participant. This participant created three groups (VFR, TAF, and IFR) with IFR having two subgroups (Icing Related and Thunderstorm-related). Cloud Proximity was identified as an outlier, and the “Ceiling” card was duplicated so that it could be represented in both the VFR and IFR categories.

The major drawback to the traditional scoring procedure is that it completely disregards the fact that participants created subgroups. In other words, it gives the relationship between Dewpoint and Precipitation Type (occurring in different subgroups) the same weight as the relationship between Dewpoint and Icing (occurring in the same subgroup) when clearly the participant saw and represented a difference.

Jaccard Scoring

The Jaccard similarity coefficient measures (Jaccard, 1912) similarity between sample sets. It represents the ratio of the number of categories two items have in common (their intersection) to the total number of categories containing the items (Union) (Capra, 2005).

Jaccard scoring allows for the accommodation of hierarchical groups and the creation of duplicate cards for a concept.

The formula for calculating the Jaccard score:

$$J = \frac{\text{Intersection}}{\text{Union}} = \frac{a}{a+b+c}$$

a = categories that contain both item #1 and item #2
b = categories with item #1 and not item #2
c = categories with item #2 and not item #1

Consider Dewpoint and Icing in the sorted items of Figure C.1 above. This pilot created six different categories (i.e., IFR is considered one category as are each of its subgroups – Icing Related and Thunderstorm –Related). To calculate the similarity between Dewpoint and Icing, the values of a, b, and c need to be defined as follows:

- Number of categories that contain both Dewpoint and Icing (a) = 2 (IFR and Icing Related)
- Number of categories that contain Dewpoint but not icing (b) = 0
- Number of categories that contain Icing but not Dewpoint (c) = 0

Using the formula above, Jaccard score for the Dewpoint – Icing pair = 2 / (2+0+0) = 2/2 = 1.

To calculate the similarity between Dewpoint and Precipitation Type, the values of a, b, and c are defined as follows:

- Number of categories that contain both Dewpoint and Precipitation Type (a) = 1 (IFR)
- Number of categories that contain Dewpoint but not Precipitation (b) = 1 (Icing Related)

- Number of categories that contain Precipitation type but not Dewpoint (c) = 1 (Thunderstorm-Related)

The Jaccard score for the Dewpoint - Precipitation Type pair = 1/(1+1+1) = 1/3 = .333.

To calculate the similarity between Dewpoint and Visibility, the values of a, b, and c are defined as follows:

- Number of categories that contain both Dewpoint and Visibility (a) = 0
- Number of categories that contain Dewpoint but not Visibility (b) = 2 (IFR, Icing Related)
- Number of categories that contain Visibility but not Dewpoint (c) = 1 (VFR)

The Jaccard score for the Dewpoint - Visibility pair = 0/(0+2+1) = 0/3 = 0.

Thus, Jaccard scores are not only able to distinguish between pairs of concepts that reside in the same group (e.g., Dewpoint-Icing = 1) from those that do not reside in the same group (e.g., Dewpoint – Visibility = 0), but it is also able to distinguish between pairs that reside in the same subgroup from those that reside in the same parent group but different subgroups. Pairs that reside in the same group (e.g., Dewpoint – Icing) are given a score of 1. Pairs that reside within the same parent group but different subgroups (e.g., Dewpoint – Precipitation Type) will have values greater than 0 but less than 1, indicating that they are not quite as similar as pairs that reside within the same group and/or subgroup. Jaccard scores were calculated for all 105 pairs for each participant who completed the 15-card Card Sort. Data were then formatted into similarity matrices.

Illustrating how Jaccard Similarity Scores are Influenced by Number of Parent Groups

Creating fewer parent groups means making less distinction in similarity of the items. Creating more groups means effort is being taken to create distinctive relationships between concepts. For example, Figure C.2 shows a Card Sort that resulted in one of the lowest Jaccard similarity scores (.19) because it has five Parent groups and no subgroups. Figure C.3 shows a

Card Sort that resulted in one of the highest Jaccard similarity scores (.48) because it has fewer Parent groups (less distinction between concepts) but a couple of subgroups. Figure C.4 shows a Card Sort that resulted on one of the highest Jaccard similarity ratings because it only consisted of two Parent groups and no subgroups.

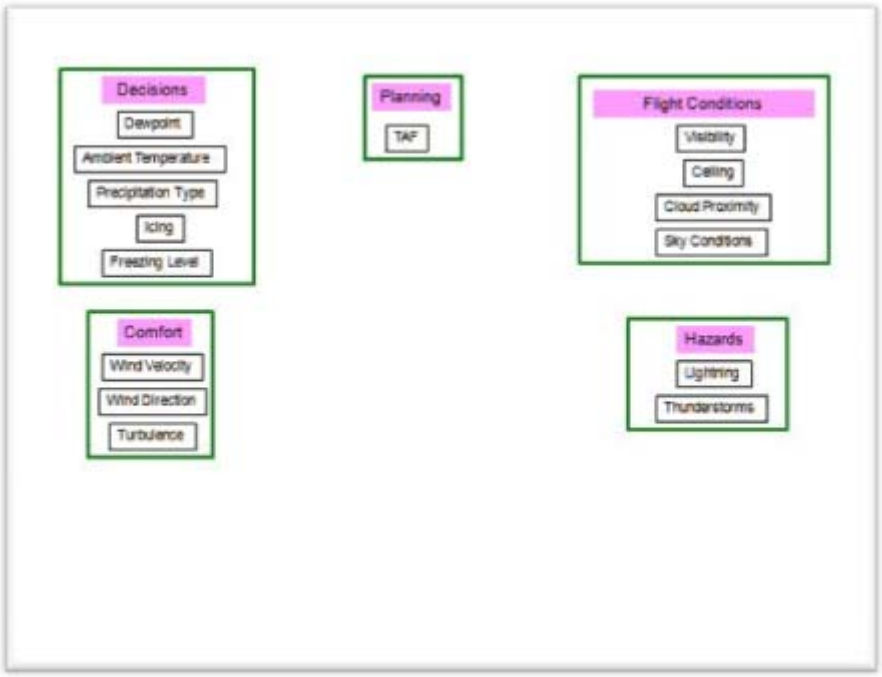


Figure C.2. Example of a Card Sort that resulted in one of the lowest Jaccard Similarity Scores in the distribution (.19). Pink shading indicates Group labels applied by the pilot.

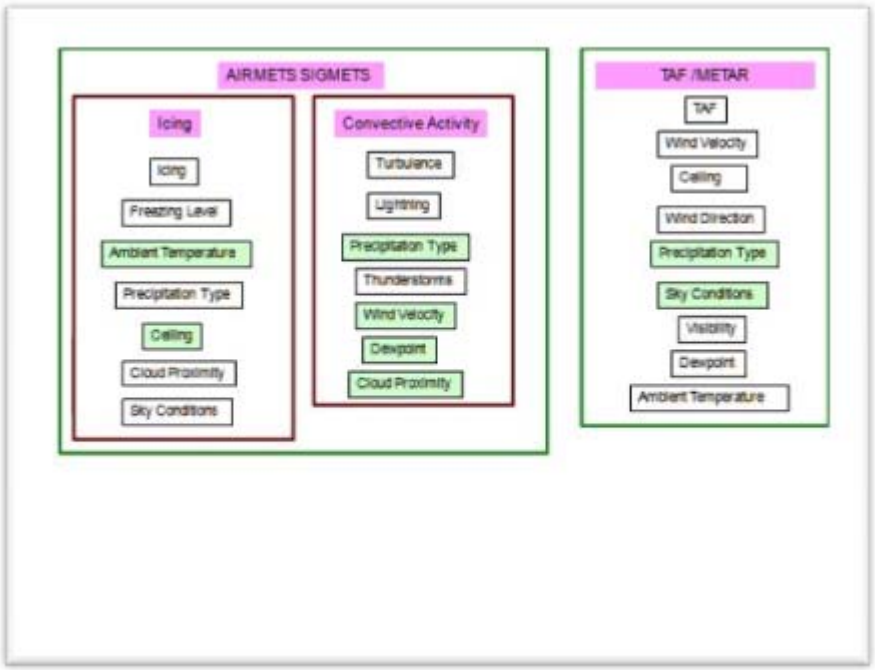


Figure C.3. Example of a Card Sort that resulted in one of the highest Jaccard Similarity Scores (.44). The Green shading indicates duplicate cards and the Pink shading indicates Group labels applied by the pilot).

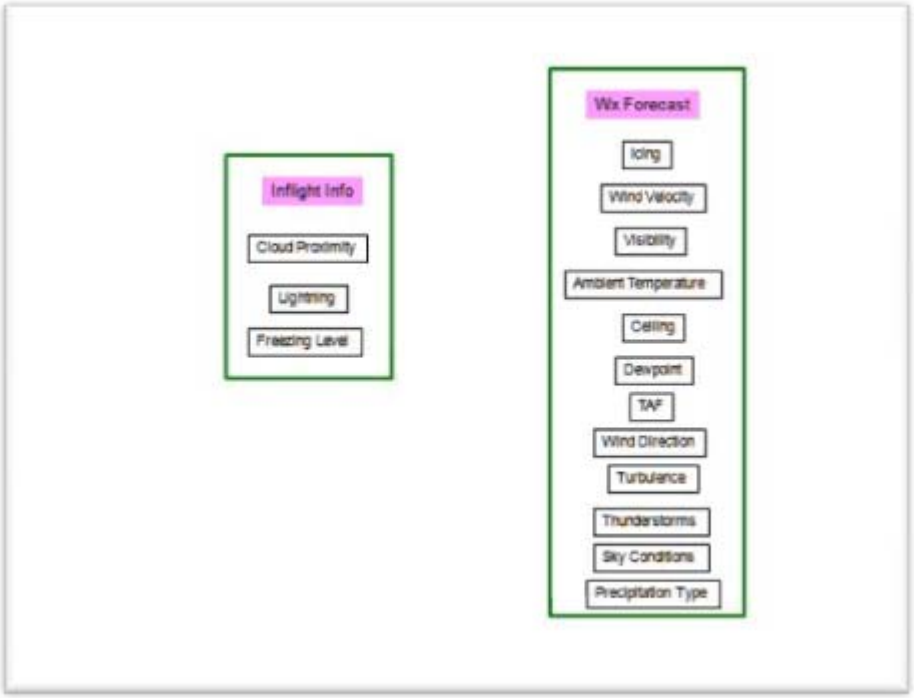


Figure C.4. Example of a Card Sort that resulted in the highest Jaccard Similarity Score (.66). Pink shading indicates Group labels applied by the pilot.

Appendix D - MDS Data Characteristics and Analysis Decisions

General MDS Procedure

In simple terms, MDS is a tool that allows for the relationships between objects to be presented spatially. Data are gathered through any number of different methods in an effort to quantify the nature of the relationship between each pair of objects. The resultant datasets are often called *proximities*. MDS attempts to represent or map proximity data as distances among points in an n-dimensional configuration, also referred to in this study as a *conceptual structure*. The mapping is specified by a representation function, f: $p_{ij} \rightarrow d_{ij}(X)$ which specifies how proximities (p_{ij}) should be related to distances (d_{ij}) , otherwise known as *disparities,* in the conceptual structure. The goal of MDS is to define the configuration (within a given dimensionality) whose disparities satisfy *f* as closely as possible (Borg & Groenen, 1997).

MDS starts by calculating some set of coordinates for the stimulus, called the starting configuration. Disparities are calculated from these coordinates and compared with the proximity data. Depending on how large the differences are, MDS will move the coordinates around and then recompute the disparities. The process continues to iterate until the disparities fit the proximity data as closely as possible. Multiple dimensions may be used to order to increase the fit of the solution to the data. In this context, the term “dimension” refers to a characteristic that serves to define a point in a space (Schiffman et al., 1981). Generally, higher dimension solutions will provide a better fit for the data. However, higher dimension solutions are not always practical and are extremely difficult to interpret. Since MDS is most often used to describe or provide insight into the latent structure of the data, it is important that the solution be interpretable on at least some of the dimensions. Typically, multiple MDS solutions are computed using a range of dimensions and then the optimal solution is chosen by the

experimenter based on a number of different factors, including *goodness of fit*, *interpretability*, and *ease of use*, all discussed in more detail below.

The specific computations used by MDS are different depending on the scaling procedure and the type of model used, both of which are discussed in further detail in upcoming sections. Many MDS algorithms are based on Euclidean distance to calculate optimal distances between objects in n-dimensional space. Euclidean distance is derived from the Pythagorean Theorem and is defined as the hypotenuse linking two points in a hypothetical right triangle. Several computer programs have been developed over the past several years to improve the efficiency and ease with which MDS can be employed. Two of the most commonly used computer programs are INDSCAL (Carroll & Chang, 1970) and ALSCAL (Takane, Young, & de Leeuw, 1977). ALSCAL is available in major statistical systems (SAS and SPSS).

The following section provides a review of basic MDS data characteristics, criteria for determining the optimal solution, and the process by which decisions were made regarding how MDS was applied to the data in the current study.

Data Characteristics

Type of Proximity Data

Typically, MDS analysis begins with data representing the relationship (i.e., proximity) between objects. This relationship is often elicited by having people directly judge the "psychological distance" or closeness of objects. *Similarity* and/or *dissimilarity* ratings are most frequently used to elicit psychological distance (i.e., proximity). The attributes used to judge the psychological distance between objects are usually not specified other than, perhaps, to indicate a specific type of similarity to judge (e.g., "political similarity" in a study about countries). Other methods to elicit proximities include *stimulus confusability* (i.e., confusability is measured by the percentage of "same" responses to a pair of physically different stimuli) and *card sorting*

(i.e., proximity reflects the number of times each item occurs with another item in the same category) (Kruskal & Wish, 1978).

The use of dissimilarity data is encouraged over similarity data for the simple reason that its relationship to distance is direct and positive – that is, objects that are highly dissimilar are assumed to have a larger perceived psychological distance between them. Thus, large proximity values will be represented by large distances in the conceptual structure. Similarity data can be transformed into dissimilarity data by subtracting the original data values from a constant that is higher than all collected values (Kruskal & Wish, 1978). For the current study, Relationship Judgment and Card Sort data (using Jaccard similarity scores) existed in raw data form as similarity data and, thus, had to be recalculated into dissimilarity ratings. Prime Recognition Task data were originally in dissimilarity form (large response times implied larger dissimilarity).

Data Matrix Shapes

MDS can handle proximity data in the form of two different shapes. *Square* data matrices occur when the same objects are represented as rows and columns of the matrix, indicating that all objects are compared to each other. The term “square” applies because there are identical numbers of rows and columns in the matrix. *Rectangular* matrices usually occur when the objects represented in the rows are different from the objects represented in the columns (e.g., rows represent nations, columns represent different types of measurement variables about those nations) (Giguere, 2006). Data from all three KETs were fitted into square matrices, as each weather concept was paired with each other weather concept to form the matrix.

Data within a square matrix is said to be *symmetric* if the order of presentation has no effect on the judgment (i.e., the judged similarity/dissimilarity between *a* and *b* is the same as the judged similarity/dissimilarity between *b* and *a*). In this case, it is typical to only include the

bottom half of the matrix in the analysis since the values above and below the diagonal are identical. If presentation order does affect the value of the judged similarity, the data is said to be *asymmetric* (Giguere, 2006). Data from all three KETs were treated as symmetric.

Number of Ways

A single data matrix of rows and columns is said to be *two-way* data because the data has exactly two ways – rows and columns. When multiple matrices comprise a data set, the data is said to be *multi-way*. For example, if participants are asked to judge the similarity of several pairs of mobile phones, then the data has *three ways* – participants and mobile phones (which comprise the rows and columns of each matrix). A study conducted before and after participants were allowed to use them would constitute four-way data, with the ways consisting of occasion (before and after use), participants, and the two ways of mobile phones. For the current study, each KET resulted in a dataset comprised of three-way data with one way corresponding to individuals (pilots) and the other two ways corresponding to the weather information concepts.

Number of Recommended Judgments per Stimulus Pair

Previous research provides the following recommendation for the number of judgments (*J*) per pair of stimuli used in any MDS analysis:

$$J = \frac{40D}{(I\text{-}1)}$$

where *D* represents the maximum number of dimensions anticipated and *I* represents the number of items used in the study (e.g., Giguere, 2006).

For the current study, $I = 15$ weather-related concepts. To ensure interpretability and ease of use of the MDS results, the maximum number of dimensions $D = 3$ (see below for more explanation about identifying the appropriate dimensionality). In each of the three KETs, each participant made one judgment for each of the 105 stimulus pairs. Therefore, the recommended

number of judgments *J* is analogous to the recommended number of participants for each analysis. Thus, according to the formula above, recommended number of participants for each MDS analysis is (40*3)/(15-1) = 9 participants. Fortunately, the total number of participants who participated in each of the KETs exceeded this number. However, when MDS was applied to individual pilot experience groups, the number of participants failed to meet this recommendation in Relationship Judgment for all pilot experience groups and in the Prime Recognition Task for High and Mid-time pilots (see Table D.1).

Table D.1. *Number of pilots who completed each KET.*

Pilot Experience Group #	Relationship Judgment	Card Sort (15 card)	Prime Recognition Task
Low-Time	6	12	9
Mid-Time	5	12	8
High-Time	8	14	7
Total	19	38	24

Missing Data

Missing data can be caused by several different events. In some cases, it can occur as a result of an experimental design decision, as in when incomplete designs are used to gather data (i.e., judgment pairs are randomly assigned to participants, with no participant providing judgments on all stimulus pairs). These designs are typically employed when not enough time is available for the number of judgments it would take to maintain a full design. Missing data may also occur inadvertently, such as through experimenter or participant error (e.g., inadvertently skipping trials). Most MDS computer programs are designed to analyze data sets that have missing values, with the exception of INDSCAL. For programs that can handle missing data, there are very few restrictions placed on the amount or pattern of missing data (Schiffman et al., 1981).

For the current study, datasets from two of the three KETS had some missing data. During Relationship Judgments, two participants (both Mid-time pilots) inadvertently chose an incorrect response (i.e., other than 1-9) as their judgment on a few occasions. These incorrect responses were removed from the data matrices, resulting in missing data (4.8% from one matrix, 1.0% from the other matrix). All but one of the Prime Recognition Task matrices had at least some data points missing. Recall that data matrices only included response times from trials in which the participant correctly indicated that the target was part of the memory set. Thus, if a participant was incorrect on a trial (i.e., answered "no" when the target was actually in the memory set), the response time for that corresponding target-prime pair was excluded from the analysis. The amount of data missing from pilots' matrices ranged from 1.0% to 10.5% with an average of 4.0% missing from matrices of Low-time pilots, 4.4% missing from matrices of Mid-time pilots, and 4.5% missing from matrices of High-time pilots.

Determining the Optimal Solution

Determining the optimal solution for an MDS output is as much an "art" is as it is a science. Of course, *goodness of fit* of the solution to the data is one important consideration. However, because MDS is typically used as a descriptive technique for representing and understanding data, determining the "true" or "statistically correct" dimensionality of a dataset is useless if the true dimensionality is too high to interpret. Therefore, in most cases the ultimate goal is to identify the *appropriate* dimensionality, rather than the *true* dimensionality of the data set (Kruskal & Wish, 1978). Several factors must be considered when deciding the appropriate dimensionality: *Measure of Fit*, *Interpretability*, *Ease of Use and Number of Stimuli*.

Measures of Fit

Scatter Diagrams

Scatter diagrams (i.e., Shepard diagrams) provide a visual assessment of the correspondence between d_{ij} (the calculated disparities) and p_{ij} (the original proximity data). Typically, the horizontal axis displays the original proximities and the vertical axis displays the disparities. Thus, each point corresponds to one pair (i, j) and has coordinates (p_{ij}, d_{ij}). If the proximities are dissimilarity data, a best fitting MDS solution should have small dissimilarities corresponding to small distances and large dissimilarities corresponding to large distances, resulting in the points in the diagram forming a rising pattern increasing from lower left to upper right on the graph. If the proximities are similarity data, the best fitting MDS solution should form a falling pattern. The amount of "scatter" is a visual indication of fit of the solution to the original data. Scatter diagrams can also be used to ensure that data does not look abnormal in any way (i.e., a degenerate solution). For example, if the function f connecting the proximities and the disparities is assumed to be linear, scatter diagrams having the wrong "shape" (i.e., not sloping upward for dissimilarity data or downward for similarity data) are an indication that a local minimum solution may have been found (Kruskal & Wish, 1978).

Stress

Stress is one of the most widely used goodness-of-fit measures for MDS. Stress is defined as the square root of a normalized "residual sum of squares" and provides an indication of how well the configuration represents the data. Stress is sometimes referred to as a "badness-of-fit" measure given that larger values of Stress indicate a worse fit to the data (Kruskal & Wish, 1978). Computer programs such as ALSCAL typically start the process of finding the best-fitting MDS representation with some initial configuration of coordinates and improve this configuration by moving points in small iterative steps to approximate the ideal

model relation $f(S_{ij})= \delta_{ij}(X)$. To determine the "badness-of-fit" between the MDS representation and the original data, SPSS ALSCAL uses a loss function called S-STRESS defined as:

$$SS1 = \left[\frac{\sum_{(i,j)} (\delta_{ij}^2 - d_{ij}^2)^2}{\sum_{(i,j)} (d_{ij}^2)^2} \right]^{1/2}$$

where δ_{ij}^2 is the squared disparity between items *i* and *j*, d_{ij}^2 is the related squared distance between items *i* and *j* in terms of the proximity data, *I* is the number of rows and *J* is the number of columns in the matrix. S-STRESS is derived from Kruskal's (1964) Stress formula 1 (SS1) measure. When using Replicated MDS or Weighted MDS, the S-STRESS given in the iteration history is calculated differently to take into account the individual differences in the multiple data matrices contributing to the representation:

$$SS1 = \left[\frac{1}{m} \sum_{(i,j)} (SS1_k)^2 \right]^{1/2}$$

where $SS1_k$ is the corresponding S-STRESS measure calculated for participant *k*, and *m* is the number of data matrices (participants) contributing to the analysis (Giguere, 2006).

Typically, computer programs compute an S-STRESS measure after each program iteration to indicate how far off the model is from the original proximity matrix, with lower values indicating less stress or a better model). SPSS ALSCAL keeps trying to improve the model by adjusting the coordinates until the S-STRESS does not improve very much with the next iteration.

In addition, SPSS ALSCAL also gives Kruskal's Stress formula 1 measures for each of the matrices used in the individual differences scaling model and for the overall model. Kruskal's Stress formula 1 is similar in concept to the S-STRESS measure except that it uses a

different equation that makes comparisons between different analyses with different programs easier (George & Mallery, 2009).

Kruskal & Wish (1978) have proposed some guidelines for interpreting Stress-1 (see Table D.2).The interpretation of stress values may be dependent upon the number of objects being examined (*I*) and the dimensionality (*m*) of the configuration. As long as the number of objects is large compared to the number of dimensions (general rule of thumb: $I > 4m$), the interpretation of stress is not sensitive to *I* or *m*. However as the number of objects approaches the number of dimensions, the interpretation of stress values is affected. For example, a stress value of .02 may generally be considered good fit when $I \geq m$. However, for $I = 7$ objects in $m=3$ dimensions (i.e., $I < 4m$), a stress value of .02 or less would occur for contentless random data about 50% of the time (see Kruskal & Wish, 1978 for more detail). Many factors can affect the value of stress. In general, stress is *higher* when 1) using metric MDS, 2) using a higher number of stimulus pairs or data matrices, and 3) there is a high level of error in the data. Stress is *lower* when 1) dimensionality of the MDS representation is higher, 2) there is missing data, and 3) when using nonmetric MDS (Giguere, 2006).

Table D.2. *Guidelines for interpreting the level of MDS model fit from Kruskal's Stress formula 1 measure (i.e., Stress-1) (Kruskal & Wish, 1978; Giguere, 2006).*

Stress-1 value	Interpretation of Fit
>.20	Poor
≥ .10 but ≤ .20	Fair
≥ .05 but ≤.10	Good
≥ .025 but ≤ .05	Excellent
.00	Perfect

Stress should always decrease as dimensionality decreases and this is most often visualized by a *Scree Plot*. Scree plots visually depict the Stress value as a function of the dimensionality of the configuration and can be used to determine if adding an extra dimension significantly decreases the badness-of-fit. One way to do this is to look for an "elbow" in the screen plot, indicating that additional dimensions do not result in significant decreases in the

badness-of-fit. However, elbows are often very difficult to identify or non-existent in the data so other criteria for identifying dimensionality must be used.

In addition to Stress measures, SPSS ALSCAL also provides an $\boldsymbol{R^2}$ value for each matrix in an individual differences scaling model and for the overall model. R^2, also referred to as the squared correlations, is an indicator of the *proportion of variance of the disparities accounted for by the MDS model*, thus higher numbers of R^2 are better (George & Mallery, 2009; Schiffman et al., 1981). Some argue that R^2 provides a better indicator of how well the model fits the data than Stress because R^2 is simpler to interpret. Schiffman et al. (1981) advocate the use of R^2 and provide examples of studies where R^2 does provide a better indicator of the appropriate dimension than Stress (see Chapter 9 of their book for more detail).

Interpretability

Many authors (e.g., Kruskal & Wish, 1978; Davidson, 1983) suggest that the ability to interpret a configuration should be a central consideration for choosing dimensionality, especially when a range of reasonable dimensionalities has been suggested by goodness-of-fit measures. In other words, it may not be necessary to add dimensions that do not contribute to the interpretation and understanding of the underlying dataset, just for the purposes of reducing the stress value. Similarly, one should consider removing dimensions when it helps the interpretation even though the stress value may increase (Kruskal & Wish, 1978, Giguere, 2006). However, caution should be used in trusting any interpretation when the configuration fits the data too poorly (Kruskal & Wish, 1978).

Ease of Use

Generally, interpretation is easier on configurations of fewer dimensions. According to Kruskal & Wish (1978, p. 58), “when an MDS configuration is desired primarily as the foundation on which to display clustering results, then a two-dimensional configuration is far more useful than one involving three or more dimensions.” Kruskal and Wish also note that configurations

based on a higher number of dimensions may only be useful when supplementary techniques are employed to find understandable and interesting structures. Of course, by constraining the interpretation to fewer dimensions, there is always the risk that important aspects of the structure are missed because they are not represented in the small number of dimensions. However, configurations based on lower dimensions are typically easier to interpret and easier to explain to a general audience. Thus, interpretations will only be discussed for MDS solutions in two-dimensions and occasionally three-dimensions when the meaning of third dimension provides additional clarity to the understanding of the conceptual structure.

Number of Stimuli

The number of stimuli has a direct effect on the number of dimensions that can be reliably observed and interpreted in the MDS output. However, recommendations for the proportion of stimuli to the number of dimensions vary across researchers. Kruskal & Wish (1978) recommend at least 9 stimuli in order to identify two-dimensional solutions, 13 stimuli for 3 dimensions, and 17 stimuli for 4 dimensions. Spence and Domoney (1974) recommend at least 11 stimuli for two-dimensional solutions and 17 stimuli for 3 dimensional solutions. Schiffman et al. (1981) recommend 12 stimuli for two-dimensional solutions and 18 stimuli for three-dimensional solutions. However, Schiffman et al. (1981) also note that recommendations can be weakened if there are many matrices that contribute to the analyses (e.g., > 10 matrices), although they admit they have no empirical data on which to base this recommendation. Given the use of 15 stimuli in the current study, most guidelines would suggest that no higher than a three-dimensional solution should be interpreted.

Decisions Made Regarding the Application of MDS to KET Data

Several decisions had to be made to ensure the appropriate MDS analysis was applied to the KET data. First, the decision was made to apply *nonmetric MDS*, even though the measurement level of the KET data could be considered metric (interval and ratio). Second, the

decision was made to treat all data from each KET as *matrix conditional*, rather than unconditional. Third, the weighted MDS model was identified as the appropriate model to answer the research questions. Fourth, the decision was made to employ procedures to untie any ties that were in the original proximities matrices. The following sections describe each issue and the rationale behind the decisions that were made.

MDS Scaling Procedure

The terms “metric” and “nonmetric” are often used to mean two different things within the context of MDS. First, the terms may be used to describe the *measurement level* of the data. Used in this context, nonmetric refers to data measured at a nominal (objects sorted in groups) or ordinal level (objects ranked in order of magnitude) and metric refers to data measured at an interval (the magnitude of the difference between objects is shown by a scale) or ratio level (position along a scale represents absolute magnitude of the attribute, with the scale having a zero point) (Schiffman et al., 1981).

Metric and nonmetric can also refer to *MDS scaling procedures*. In metric MDS, scaled distances preserve the original proximity data in a linear fashion (i.e., by applying linear transformations to the data). In nonmetric MDS, scaled distances only preserve the rank order of the original proximity data. That is, monotonic transformations applied to the original data maintain the rank order of the proximities and allow the performance of arithmetic operations on those rank orders (Schiffman et al., 1981). Thus, nonmetric solutions can be found for proximity data whether that data is measured at a metric (i.e., interval or ratio) or nonmetric (i.e., ordinal) level. Used in this context, “nonmetric” indicates that a nonlinear monotone transformation has been applied to the original data.

In most cases, the use of metric scaling vs. nonmetric scaling has very little effect on the resultant conceptual structure but the goodness of fit (i.e., stress) is typically lower when using nonmetric scaling (Kruskal & Wish, 1978). According to Schiffman et al. (1981), “in general,

nonmetric scaling, where only rank order relationships are maintained, provides spaces with better fit in low dimensionality than metric solutions" (p. 6). Also, previous research in knowledge elicitation where MDS has been employed suggests using nonmetric scaling to analyze similarity judgments (e.g., Jonassen et al., 1993). For these reasons, the current study employed nonmetric MDS.[17]

Matrix Conditionality

Conditionality refers to any relationships that may exist among observations within sets of observation categories (Young & Hamer, 1987). Data is said to be matrix conditional if individual differences between participants are hypothesized. For example, when participants judge the similarity of all pairs of a set of stimuli, it may be inappropriate to compare one participant's response to another because those participants may not be using the response scale in identical ways. That is, a response of "7" on a similarity scale cannot be assumed to represent more similarity than another participant's response of "6" and one participant's response of a "6" may not be assumed to indicate the same magnitude of similarity as another participant's response of "6". Thus, the term "matrix conditional" is used because the meaning of the data is conditional upon which participant is responding (Young & Hamer, 1987). Conversely, unconditional data matrices are able to be meaningfully compared with each other. Objective measures such as response times and test scores most likely can be compared across participants in a meaningful way (Giguere, 2006; Young & Hamer, 1987).

Generally, MDS is performed when datasets are considered matrix conditional. In terms of the current study, clearly the *Relationship Judgment* data must be considered matrix conditional, as it fits the exact description of matrix conditionality above. While response time data (from the *Prime Recognition Task*) can generally be considered unconditional according to

[17] Separate analyses were conducted using metric scaling with the results confirming previous research. Conceptual structures were relatively unchanged between metric and nonmetric scaling, but the nonmetric scaling solutions provided a better fit to the data.

Young and Hamer (1987), it was considered matrix conditional for the purposes of the current study. Some participants were just naturally faster or slower in making most of their judgments in the Prime Recognition Task and therefore, both their range and magnitude of the response times varied widely. The following hypothetical example illustrates the reasoning behind this decision. Consider a case where a specific prime-target pair elicited a response time of 656ms for both Pilot 1 and Pilot 2. However, if Pilot 1's response times ranged from 300-700ms while Pilot 2's response times ranged from 600 ms – 1300 ms, that response time of 656ms has different implications for the relative similarity of that pair compared to other pairs depending on whether it is part of the dataset for Pilot 1 or 2. Thus, Prime Recognition Task data were considered to be matrix conditional for the analysis. *Card sort* data was also treated as matrix conditional for the analysis because the value of each data point can be considered to be conditional upon how many groups each participants felt it necessary to create.

Choosing the Appropriate MDS Model

There are three main models by which MDS is applied to a dataset. Generally, these models differ in terms of 1) number of matrices contributing to the analyses and 2) accommodation of individual differences.

Classical MDS (CMDS)

Classical MDS is the simplest of the MDS models. It uses only one matrix of raw or averaged data (i.e., *one-way* data). The algorithm produces a hypothetical stimulus space where Euclidean distances are arranged such that they match the original data as much as possible. The original proximity data are transformed into disparities using a linear (for metric MDS) or monotonic (for non-metric MDS) function. Thus, if a study contains multiple participants, the participants' proximity matrices must averaged, with the averaged matrix used as the dataset for the MDS analysis.

Replicated MDS (RMDS)

Replicated MDS uses several data matrices (e.g., *three-way data*) and allows for the accommodation of individual differences in the response biases between the participants (i.e., different ways participants may have used the response scales), depending on the conditionality of the data. It does so by permitting the use of different response transformation functions, one for each participant (proximity matrix). However, only one stimulus space is produced (e.g., Schiffman et al., 1981).

Weighted MDS (WMDS)

Weighted MDS also uses three-way data, with the third way typically corresponding to participant differences. However, true WMDS need not be reserved to just individual differences. The third way could include occasions or experimental conditions other than individuals. For the current study, however, the third way does correspond to pilots.

WMDS can be used when it is assumed that participants differ in the degree to which a given dimension (e.g., characteristic, attribute, feature, etc.) affects their similarity judgments. The model assumes there is a common stimulus space that reflects the dimensions used by the entire group of participants. The model calculates a vector of weights, with a weight value for each dimension that reflects how important that dimension was to the participant's judgment. A large weight value indicates that a dimension was highly salient to the participant when making the judgment. A zero weight value indicates that a dimension was not used by the participant when making the judgment (Young & Hamer, 1987). It is the introduction of weights that sets WMDS apart from RMDS. While RMDS allows individual differences in response bias through the data transformations, the individual differences are not represented in the Euclidean model. In WMDS, the weight values that specify differences between participant matrices are included in the Euclidean model (e.g., Schiffman et al., 1981). Thus, WMDS uses several data matrices (one per participant) and provides not only a representation of the item configuration in common

stimulus space but also a *participant space* that indicates the differential weighting given by each participant to each of the dimensions depicted in the common stimulus space (Giguere, 2006). While Classical MDS and Replicated MDS require matrix conditional data, Weighted MDS can use several matrices of either matrix-conditional or unconditional data. Table D.3 specifies the model equations for the three MDS models.

Table D.3. *Model Equations for each of the MDS models (Giguere, 2006).*

Classical MDS (CMDS)	**Replicated MDS (RMDS)**	**Weighted MDS (WMDS)**
$T(P) = D^2 + SSE$ P = original proximities matrix T(P) = disparity matrix stemming from transformation T (linear or monotonic depending on metric or nonmetric MDS) D^2 = squared Euclidean distances fit by ALSCAL SSE = sum of squared errors between the distances and disparities	$T_k(P_k) = D^2 + SSE_k$ P_k = original proximities matrix for participant *k* $T_k(S_k)$ = individual disparities matrix for participant *k* stemming from unique transformation T_k (linear or monotonic depending on metric or nonmetric MDS) D^2 = squared Euclidean distances fit by ALSCAL for the common stimulus space SSE_k = sum of squared errors between distances and disparities for participant *k*	$T_k(P_k) = D^2_k + SSE_k$ P_k = original proximities matrix for participant *k* $T_k(P_k)$ = individual disparities matrix for participant k stemming from unique transformation T_k (linear or monotonic depending on metric or nonmetric MDS) D^2_k = squared Euclidean distances fit by ALSCAL for participant *k* SSE_k = sum of squared errors between distances and disparities for participant *k*

Of primary importance to the current study is the examination of individual differences in knowledge structures as represented by the MDS output (i.e., conceptual structures). Therefore, all analyses employed the WMDS procedure. It should be noted that the terms "Weighted MDS" or WMDS and "individual differences scaling" (INSCAL) are often used interchangeably in the literature. However, Schiffman et al (1981) identify a distinction between the two terms that will be adopted and used in the current study. INDSCAL will be used to refer to the computer program (originally developed by Carroll & Chang, 1970) while Weighted MDS or WMDS will refer to the analysis. This distinction is important because WMDS is not reserved

just for individual differences. As previously noted, the third way in the data can also correspond to occasions or experimental conditions other than individual differences.

ALSCAL and INDSCAL are of the more common two computer programs that perform WMDS, but do so in slightly different ways. INDSCAL was the first to perform WMDS and provides only a metric solution by optimizing the fit of scalar products to a transformation of the data. ALSCAL can provide both nonmetric and metric solutions to WMDS and does so by optimizing the fit of squared distances in the data (Young, 1985). Thus, ALSCAL was chosen to conduct the MDS analysis because it is 1) well known, 2) can accommodate datasets with missing data, 3) provides the flexibility to construct both metric and nonmetric MDS solutions, and 4) was easily accessible (through SPSS 13).

The Measurement Process and its Implications for Managing Ties in the Data

Measurement Process

The type of measurement process (discrete or continuous) used to generate the proximity data has implications for the treatment of ties in the data. Since the data for the current study are all obtained from participants, then the distinction between discrete and continuous refers to the nature of what is assumed to be occurring in the mind of the participants as they are making those judgments. If their internal scale is assumed to be continuous but they are forced to provide discrete numbers, then the data are continuous. If the internal scale really is discrete, then the data can be thought of as discrete. Ties are left tied when the process is considered discrete but they should become untied if the process is considered continuous. Most programs only allow the option for ties to become untied for the ordinal level of measurement. The choice of whether or not to untie the data is only crucial when there are a large number of ties in the data. Otherwise the solutions between analyses on tied and untied data are typically not very different (Schiffman et al., 1981).

The decision of how to treat ties in the data for the current study was based on both logical reasoning and empirical evidence. Logical reasoning took into account the assumptions that could be made regarding participants' internal response scales when participating in the KETs. The empirical evidence was obtained by comparing the results for MDS analyses conducted on data from each KET when ties were allowed to exist and when procedures were used to untie the data ties. The next section provides more information on the assumptions regarding pilots' internal scales and comparison of the MDS output for data in which data was left tied or untied.

Decision to Use Tied or Untied Data

An important consideration for nonmetric MDS is how to handle ties in the data, given that a monotone function is used to preserve the order of the proximities.[18] One approach, known as the *primary approach*, is to allow the model to break the ties when fitting distances (disparities) if it increases the goodness of fit. Thus, just because proximities p_{ij} and p_{kl} are equal ($p_{ij}= p_{kl}$) does not necessarily mean that their scaled disparities d_{ij} and d_{kn} will be equal as well. The secondary approach retains all ties in the fitting of distances such that if $p_{ij}= p_{kl}$ then also $d_{ij} = d_{kl}$. Many MDS program use the primary approach as the default for ordinal data, although ALSCAL does not (Borg & Groenen, 1997).

The type of measurement process used to generate or collect the data (i.e., discrete vs. continuous) has implications for how ties should be treated in the data. Since the data for the current study are all obtained from pilots, then the distinction between discrete and continuous refers to the nature of what is assumed to be occurring in the mind of the pilots as they are making those judgments. If their internal scale is assumed to be continuous but they are forced to provide discrete numbers, then the data are continuous and any ties in the data should be

[18] Most programs only allow the option for ties to become untied for the ordinal level of measurement. For other levels of measurement, the measurement process is implied to be discrete (Schiffman et al., 1981).

untied when distances are fitted (i.e., the primary approach). If the internal scale really is discrete, then the data can be thought of as discrete and ties in the data should be maintained (i.e., the secondary approach). (Borg & Groenen, 1997; Schiffman et al., 1981).

Borg & Groenen (1997) provide some examples of different data collection methods that can result in ties in proximity data. Variants on a Card Sort task and Relationship Judgment task were highlighted as two major examples:

- Card Sort: Imagine participants are asked split a stack of cards into two piles, one containing more similar pairs, the other more dissimilar pairs. Then the participant repeats the sort for each pile multiple times until he or she determines that it is not possible to further discriminate between pairs of objects in any pile. If at the conclusion of the sort each pile has only one card left, then the result is a complete similarity order of pairs of objects, with no ties occurring. However, the more likely result is that some piles have multiple cards, with smaller piles for pairs objects that are extremely similar and larger piles for pairs of objects that have intermediate similarity. However, it would be inappropriate to assume that objects within the same group hold equal similarity. Instead, group membership merely means that the pairs in some piles do not appear to be sufficiently different to warrant further meaningful sorting. In this case, Borg & Groenen (1997) advise breaking ties in the data for analysis.
- Relationship Judgment: Say that participants are asked to rate the degree of similarity between each pair of 12 objects on a 9 point scale, with end points labeled “very different” and “very similar.” In this situation, ties in the data are expected, because there are 66 pairs of objects and would thus require a rating scale of at least 66 categories in order to be able to assign a different proximity value to each pair. In this case, the 9 point scale acts as a high-level approximation of true

similarity, with the data representing intervals on a continuum of similarity. Borg & Groenen (1997) also advise breaking ties in the data because the ties occur as a result of the data collection method.

The choice of whether or not to untie the data is only crucial when there are a large number of ties in the dataset, otherwise the solutions between analyses on tied and untied data are typically not very different. However, extreme caution should be used when interpreting solutions from datasets that contain an abundance of ties and those ties are at the extreme end of the data range. Through the use of scattergrams and plots of transformations, Schiffman et al. (1981) showed how transformations of continuous (i.e., untied) data resulted in many of the original tied "no difference" observations being transformed into large disparities. In other words, untying data from their study provided transformed disparities that were either basically zero or approximately uniform but large numbers. While their suggestion is still to untie the data, they caution against over-interpreting any solutions that come from similar types of extreme data.

Implications for the Current Study

Arguments can be made for data in all three KETs arising from an internal continuous scale. Prime Recognition Task data are comprised of response times, and response time is considered a continuous variable. Also, as Borg & Groenen's (1997) previous examples illustrate, one can argue that the participant uses an internal continuous scale of similarity on which to base their Relationship Judgments and Card Sort behaviors as well. Not surprisingly, both the Relationship Judgment and Card Sort data contain a considerable number of ties because of their relatively constrained response scales.

Although several sources suggest the primary approach (i.e., untying data) is most appropriate in most cases of nonmetric MDS, the cautionary example provided by Schiffman et al. (1981) suggested the need to conduct MDS analyses using both the primary (i.e., untied

data) and secondary (i.e., leave data tied) approaches to ultimately understand the best approach for each of the KET data sets. Thus, individual nonmetric WMDS analyses were conducted on datasets from each of the three KETs with SPSS ALSCAL using both the primary and the secondary approach to data ties. Two-dimensional solutions for both participant spaces and stimulus spaces were compared across each approach to examine the effect that untying the data has on the MDS solutions.

Determining the Optimal Solution

Table D.4 shows the values of stress and R^2 for 2-6 dimensional solutions when using tied and untied data for each KET dataset. As Tables D.4a and D.4b indicate, there is a vast improvement in the amount of variance explained by the solution when MDS is performed on untied data in Relationship Judgment and especially in Card Sort. There is also some substantial improvement in the Stress measures. The differences in R^2 and Stress between untied and tied Prime Recognition Task data are negligible, which is to be expected because there was a very low instance of ties in the original response time data. Therefore, because of the nature of the assumed internal scales for the responses in the KETS and the improvement seen in the goodness of fit measures, results will be interpreted from the MDS analysis using untied data only for each KET.

Table D.4. *Values of Stress and R^2 for each of the 2-6 dimensional solutions for untied data are shown for a) Relationship Judgment, b) Card Sort, and c) Prime Recognition task data.*

a) Relationship Judgment (RJ)

	Data Left Tied		Data Untied	
Dimension	*Stress*	*Ave R^2*	*Stress*	*Ave R^2*
6	.14	.50	.10	.80
5	.16	.51	.12	.79
4	.20	.48	.14	.76
3	.24	.48	.17	.74
2	.30	.45	.22	.70

b) Card Sort (CS)

Dimension	Data Left Tied		Data Untied	
	Stress	*Ave R^2*	*Stress*	*Ave R^2*
6	.18	.33	.08	.91
5	.21	.35	.09	.89
4	.26	.31	.11	.87
3	.29	.30	.14	.88
2	.38	.29	.18	.85

c) Prime Recognition Task (PRT)

Dimension	Data Left Tied		Data Untied	
	Stress	*Ave R^2*	*Stress*	*Ave R^2*
6	.13	.27	.13	.27
5	.16	.23	.16	.23
4	.20	.22	.20	.21
3	.26	.21	.26	.21
2	.37	.15	.36	.16

Stimulus Spaces[19]

The stimulus space resulting from the Relationship Judgment task stayed fairly consistent whether the primary approach or secondary approach was used to handle ties in the data (Figure D.1). Items that clustered together when ties were allowed tended also to cluster together when ties were removed from the data. Untying the data resulted in greater dissimilarity between Icing and Ambient Temperature and between Ambient Temperature and Freezing level, although the respresentation still maintained an indication of overall similarity between those three items. Untying the data also resulted in slightly greater similarity between TAF, Sky Condition, Visibility, Ceiling and, to a lesser extent, Dewpoint.

Not surprisingly, untying the data had more of an effect on Card Sort stimulus spaces (Figure D.2). When ties were allowed to exist, items tended to cluster closer together in more well defined groups near the extremes of the two dimensions. When data were untied, these

[19] Note: In this section, stimulus spaces will only be discussed insofar as the relative placement of items will be compared and contrasted across the stimulus spaces that result from the primary and secondary approaches. Interpretation of the underlying structure occurs in other sections of the main document.

clusters still seemed somewhat apparent, but location of items in the representation seemed to be distributed more uniformly across the dimensions.

The relative locations of the items within the Prime Recognition Task stimulus space stayed fairly consistent regardless of the primary or secondary approach to ties in the data (Figure D.3). However, note that many of the items shift from one side of a dimension to another as the approach to ties in data changes. For example, Precipitation Type is positioned on the right side of Dimension 1 when ties were allowed, but is positioned on the left side of Dimension 1 when ties were not allowed. The same phenomenon occurs for Cloud Proximity, Lightning, and others. Note that the same repositioning occurs along Dimension 2 (e.g., Wind Velocity is positioned at the lower end of Dimension 2 when ties were allowed and at the upper end of Dimension 2 when ties were not allowed). However, the relative position of each item stays very consistent in both stimulus spaces. The reason for this repositioning is uncertain.

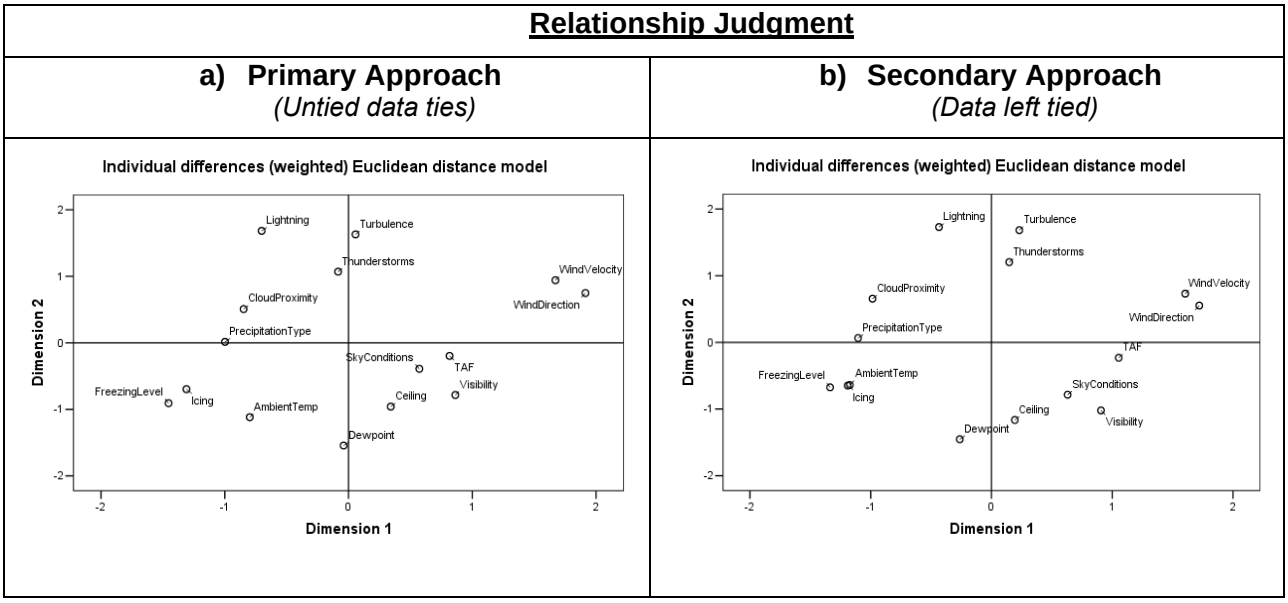


Figure D.1. Comparison of stimulus space for Relationship Judgment data resulting from WMDS analysis when ties in the data were handled using a) the primary approach and b) the secondary approach.

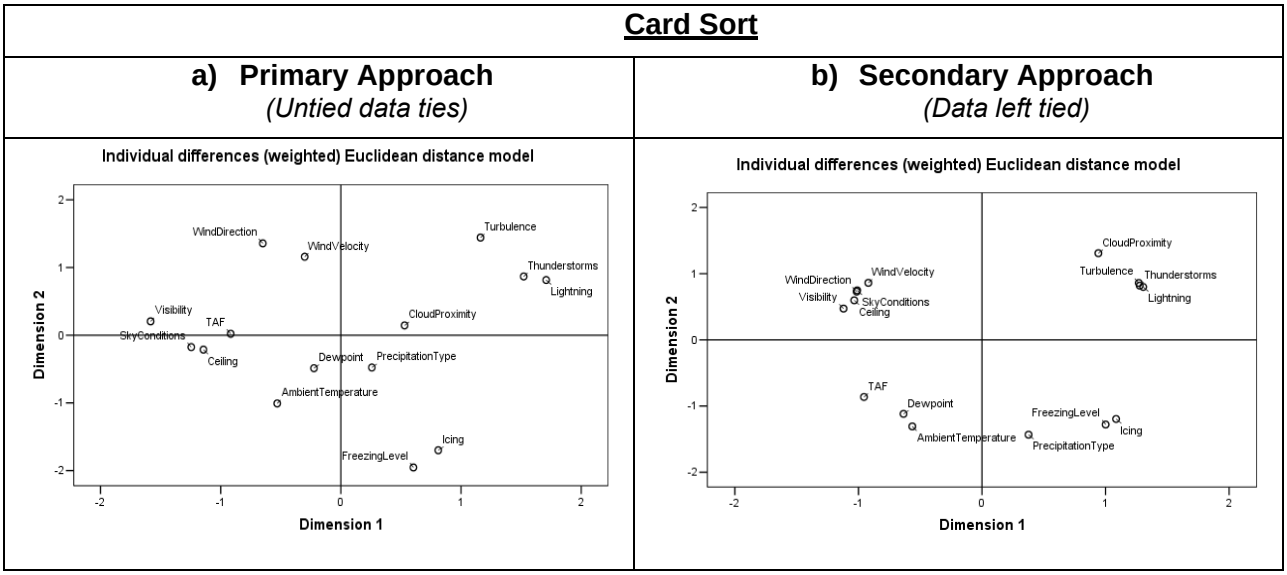


Figure D.2. Comparison of stimulus space for Card Sort data resulting fromWMDS analysis when ties in the data were handled using a) the primary approach and b) the secondary approach.

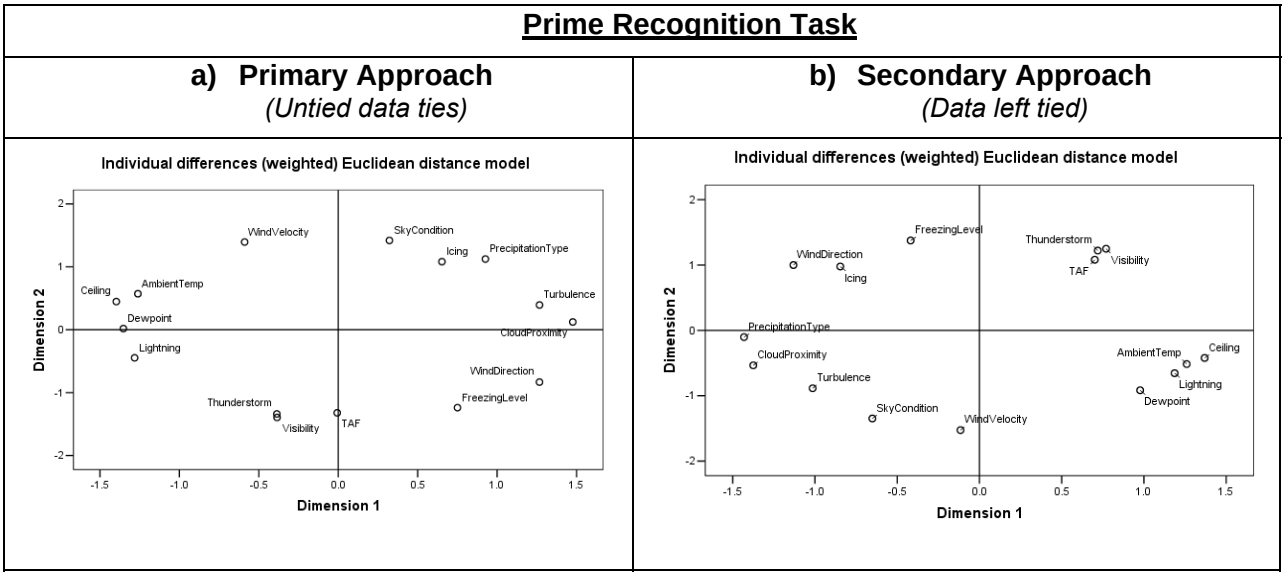


Figure D.3. Comparison of stimulus space for Prime Recognition Task data resulting from WMDS analysis when ties in the data were handled using a) the primary approach and b) the secondary approach.

Participant Spaces

Figures D.4-D.6 show the participant spaces for tied and untied data in each KET. Each point in the graph represents the location of a participant's dimension weights, with the number indicating the participant's label in the dataset. However, it should be noted that the label meaning does not stay consistent across KETs (e.g., Pilot 2 in Relationship Judgment is not the same participant as Pilot 2 in Card Sort). Pilot experience level is noted by corresponding icon: Low-Time Pilots (green circle), Mid-Time Pilots (orange triangle) and High-Time Pilots (blue diamond).

For the most part in Relationship Judgments and Card Sort, representation of pilots' weights maintains the same order of magnitude regardless of the approach to ties in the data. Untying the data appears to increase the distance between pilots compared to when data were left tied. The biggest change in participant weights as a function of data tie approach was seen in the Card Sort data (Figure D.5). In this data set, notice that pilots who originally had very low weights for both dimensions when data were left tied (e.g., Pilot #'s 32, 7, 18, etc.) actually resulted in having much higher weights for Dimension 2 compared to Dimension 1 when data were untied. The effect of untying the data did not seem to be as extreme for weights along Dimension 1, as most of the weights were merely stretched along the access. Not surprisingly, participant spaces for the Prime Recognition Task were relatively unchanged when comparing the primary and secondary approaches to data ties.

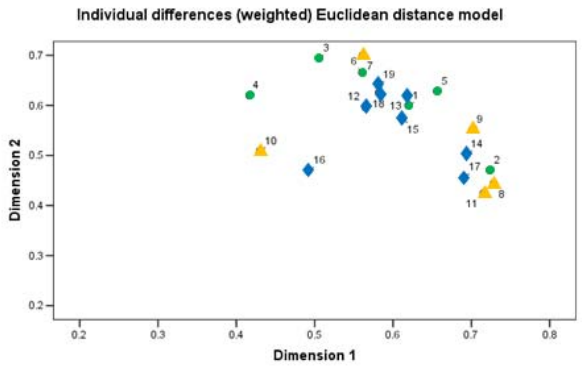
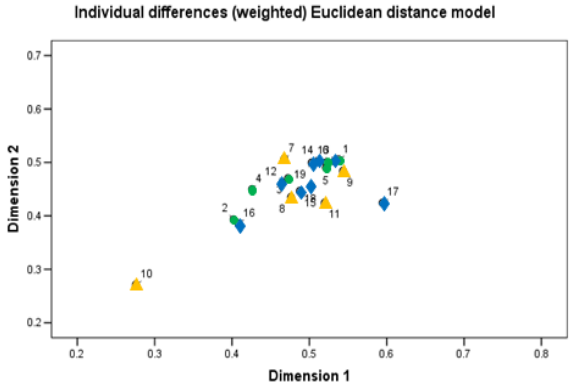
Relationship Judgment	
a) Primary Approach *(Untied data ties)*	**b) Secondary Approach** *(Data left tied)*
	

Figure D.4. Comparison of the participant space for Relationship Judgment data resulting from WMDS analysis when ties in the data were handled using a) the primary approach and b) the secondary approach.

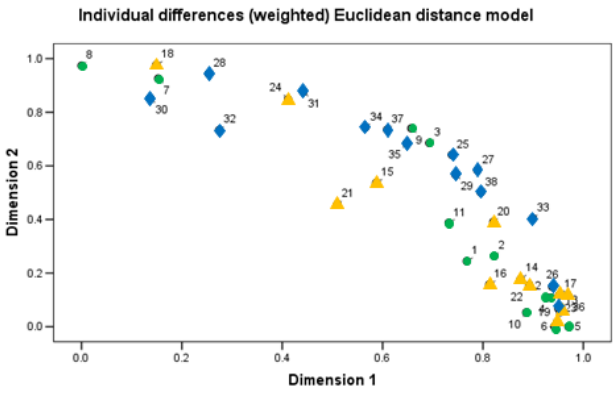
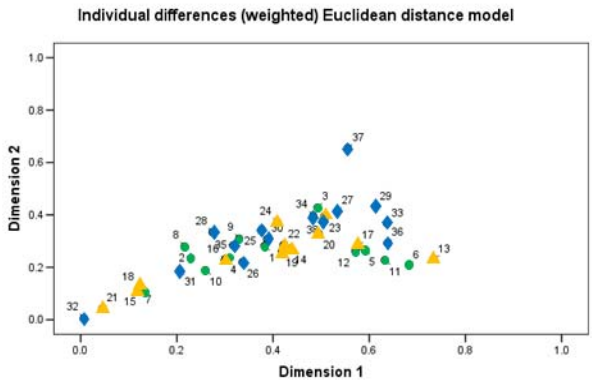
Card Sort	
a) Primary Approach *(Untied data ties)*	**b) Secondary Approach** *(Data left tied)*
	

Figure D.5. Comparison of the participant space for Card Sort data resulting from WMDS analysis when ties in the data were handled using a) the primary approach and b) the secondary approach.

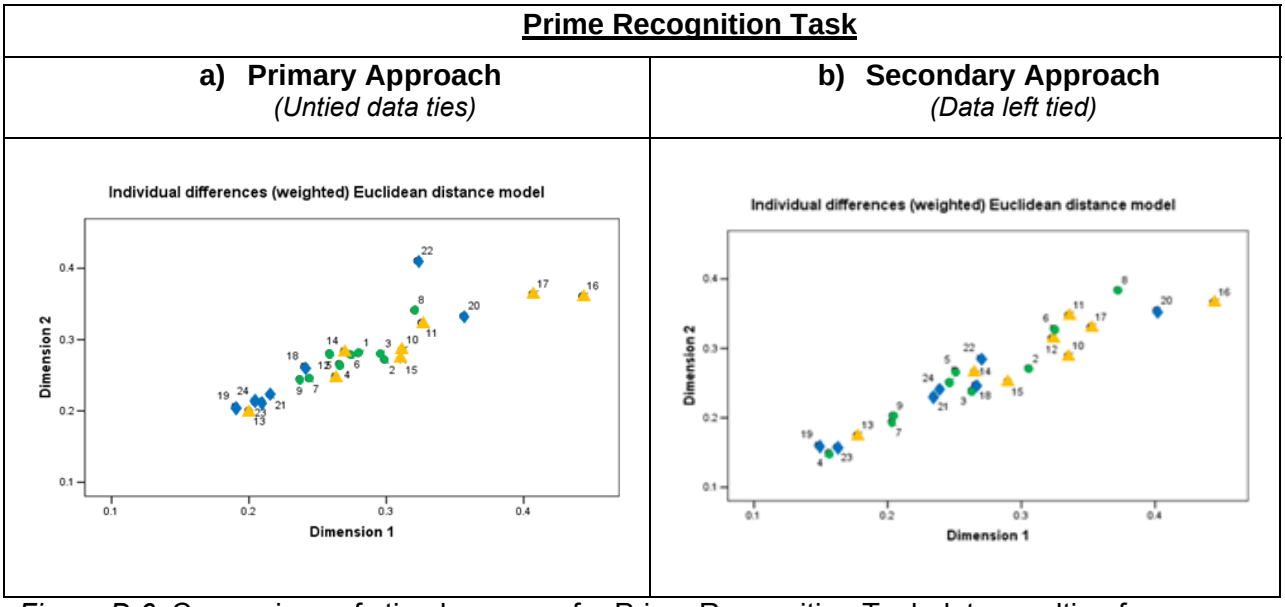


Figure D.6. Comparison of stimulus space for Prime Recognition Task data resulting from WMDS analysis when ties in the data were handled using a) the primary approach and b) the secondary approach.

Summary

Overall, there was vast improvement in the amount of variance explained by the solution when MDS was performed on untied data from Relationship Judgment and especially from Card Sort. There was also some substantial improvement in the stress measures. The differences in R^2 and stress between untied and tied Prime Recognition Task data was negligible, which was to be expected because there was a very low instance of ties in the original response time data.

Untying the data seemed to influence how tightly or loosely the items were clustered. However, caution should be taken when interpreting the distance between items in a cluster, as MDS is generally more robust to representing global structure than it is local structure (Schvaneveldt et al., 1985). Untying had the most effect on the order of magnitude on a given dimension in the Card Sort Task. However, overall the global structure seemed to maintain throughout when data were untied compared to when data were left tied.

Thus, final interpretation of the MDS analysis will occur on the untied data for the following reasons: 1) increased fit of the model to the data while maintaining logical relationships between items, 2) participants' internal scales are assumed to be continuous, which logically implies data should be continuous (untied) as well, 3) the primary approach is the one most often advised by previous researchers in the field of MDS.