2010

Investigating The Reliability And Validity Of Knowledge Structure Evaluations: The Influence Of Rater Error And Rater Limitation

Michelle Harper-Sciarini
University of Central Florida

INVESTIGATING THE RELIABILITY AND VALIDITY OF KNOWLEDGE STRUCTURE EVALUATIONS:

THE INFLUENCE OF RATER ERROR AND RATER LIMITATIONS

by

MICHELLE EVONNE HARPER-SCIARINI
B.S., University of Central Florida, 2001
M.A., University of Central Florida, 2006

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the Department of Psychology
in the College of Sciences
at the University of Central Florida
Orlando, Florida

Spring Term 2010

Major Professor: Florian Jentsch

ABSTRACT

The likelihood of conducting safe operations increases when operators have effectively integrated their knowledge of the operation into meaningful relationships, referred to as knowledge structures (KSs). Unlike knowing isolated facts about an operation, well integrated KSs reflect a deeper understanding. It is, however, only the isolated facts that are often evaluated in training environments. To know whether an operator has formed well integrated KSs, KS evaluation methods must be employed. Many of these methods, however, require subjective, human-rated evaluations. These ratings are often prone to the negative influence of a rater's limitations such as rater biases and cognitive limitations; therefore, the extent to which KS evaluations are beneficial is dependent on the degree to which the rater's limitations can be mitigated. The main objective of this study was to identify factors that will mitigate rater limitations and test their influence on the reliability and validity of KS evaluations. These factors were identified through the delineation of a framework that represents how a rater's limitations will influence the cognitive processes that occur during the evaluation process. From this framework, one factor (i.e., operation knowledge), and three mitigation techniques (i.e., frame-of-reference training, reducing the complexity of the KSs, and providing referent material) were identified. Ninety-two participants rated the accuracy of eight KSs over a period of two days. Results indicated that reliability was higher after training. Furthermore, several interactions indicated that the benefits of domain knowledge, referent material, and reduced complexity existed within subsets of the participants. For example, reduced complexity only increased reliability among evaluators with less knowledge of the operation. Also, referent material increased reliability only for those who scored less complex KSs. Both the practical and theoretical implications of these results are provided.

To Olive Lily,

Thank you for being there from the beginning to the end.

Love, Mommy

ACKNOWLEDGMENTS

First, I would like to express my sincere gratitude to Florian Jentsch for being a patient and diligent advisor who shaped and guided me through graduate school; and furthermore, for his hard work at maintaining a productive research environment in which this dissertation was completed. I would also like to thank my committee members, Shawn Burke, Clint Bowers, and Eduardo Salas for their patience, support, and guidance throughout the dissertation process; Monica Ritman, whose assistance with the data collection process was essential to the completion of this dissertation; Davin Pavlas, whose knowledge of both web design and research design, was essential for developing the materials for the study; Mike Curtis, Bill Evans, and Reagan Hoeft, for the many productive brain storming sessions; Steve Fiore, who dug me out of the “weeds”; my Mom and Dad for their endless emotional support, my mother-in-law for being there to take care of me and my daughter through-out the most stressful part of this journey; Courtney Dorn, a very special friend who has been patient with my absenteeism, and has provided many words of encouragement; and most importantly, Lee Sciarini—a wonderful husband, father, and friend. I could fill every page of this dissertation with the things he has done to make the completion of my Ph.D. possible. In this limited space, I would like to say thank you for inspiring me to be the erudite person I am today. You facilitated the growth of my confidence and intelligence. I cannot find the words that express my gratitude to you for being there for me as I attained this goal. I look forward to the next chapter of our lives.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE: INTRODUCTION

Operations in high-risk environments such as commercial flight or combat often require operators or teams of operators to make crucial, quick, and effective decisions under intense stress. The likelihood of safely conducting these operations increases when operators have effectively integrated their knowledge of the operation into meaningful relationships that define the domain, procedures, and systems associated with the operation. Although it has been established that integrated knowledge of an operation is essential for safe operations, many operation-based training environments determine what an operator knows or has learned using methods that only evaluate superficial knowledge (e.g., memorization) (Day, Arthur, & Gettman, 2001). For example, many commercial airline pilot training environments emphasize procedural knowledge evaluations, such as evaluating whether pilots have memorized and can execute the hundreds of procedures required for flight including set up, equipment check, navigation, and control procedures, as opposed to evaluating integrated knowledge which would reveal a pilot's deeper understanding of flight procedures, such as understanding the relationships between the procedures and the aircraft's and/or automation's behaviors (see Dismukes, Berman, & Loukopoulos, 2007).

Methods referred to as knowledge structure (KS) evaluation methods, have been used to evaluate integrated knowledge. In fact, research on the elicitation and evaluation of KSs is prevalent within elementary education domains such as primary and secondary education of science and mathematics (J. D. Novak, 1995; Ruiz-Primo, Shavelson, Li, & Schultz, 2001; Stayanov & Kirschner, 2004; Yin, Vanides, Ruiz-Primo, Ayala, & Shavelson, 2005). In addition, researchers have investigated the application of KS elicitation and evaluation to skill-dependent tasks (Day et al., 2001), and operations and system dependent tasks such as flight (Curtis, Harper-Sciarini, Jentsch, Schuster, & Swanson, 2007; Mathieu, Heffner, Goodwin, Salas, & Cannon-Bowers, 2000; Smith & Boehm-Davis, 2008; Smith, Boehm-Davis, & Fadden, 2008) and electronic circuitry (Harper, Hoeft, Jentsch, & Boehm-Davis, 2005) . These studies have been successful in showing the benefits of both (a) having accurately interrelated knowledge connecting operation-specific information, and (b) using KS evaluations in operation-based training environments.

Although there is evidence supporting the use of KS evaluation methods for gaining a better understanding of what an operator knows about an operation, researchers have shown how unreliable and invalid these methods can be when not properly implemented (McClure, Sonak, & Suen, 1999; Ruiz-Primo, Schultz, Li, & Shavelson, 2001). In fact, the novelty of these methods, alone, will likely require steps to be taken to ensure their implementation is reliable and valid.

Purpose of the Study and Overview of Paper

The research presented here sought to investigate factors that may influence KS evaluations, with the specific goal of identifying methods that may improve their reliability and validity. How this was achieved is discussed following Chapter Two, where I review the history and application of knowledge integration, knowledge structures, and knowledge structure evaluation.

In Chapter Three, I first discuss the similarities and differences between KS evaluations and other subjective evaluations methods (i.e., job performance evaluations). Second, I explain how the reliability and validity of KS evaluation methods may be sensitive to the same or similar biases and limitations of the evaluator, or rater. Finally, I describe the framework that guided this research effort, which assumes that a rater's biases and limitations will influence the cognitive processes that occur during the rating process. This framework was then used to identify methods that may mitigate the negative influence a rater's bias or limitation may have on the reliability and validity of KS evaluations (see Figure 1).

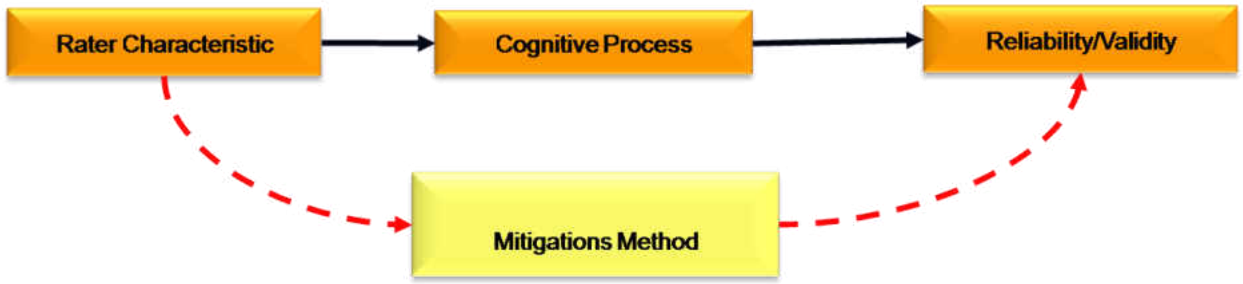


Figure 1. Framework used for investigation.

A description of the study designed to test the hypothesized affects of the mitigation methods on reliability and validity is presented in Chapter Four, the results from the study are presented in Chapter Five; and the conclusions in Chapter Six. By conducting this investigation, a better understanding of the KS evaluation process was gained, in addition to guidelines that practitioners should follow when implementing KS evaluations.

CHAPTER TWO: BACKGROUND AND LITERATURE REVIEW

Knowledge Integration

Unlike simply knowing isolated facts about an operation, having well-integrated knowledge reflects a deeper understanding of an operation. For example, one may learn the superficial facts (or surface features) of driving, such as that turning the wheels on a car is done by using a hand–operated steering wheel which is positioned in front of the driver. In contrast, however, one may also learn the deeper, conceptual features of driving, such as that the outcome of turning a steering wheel (ratio of how far you turn the steering wheel to how far the wheels turn) is often a function of the gear ratio, or the type of gearset (e.g., rack and pinion). Knowledge of surface features gives the driver general facts about steering, such as the location of the steering wheel. In contrast, knowledge of conceptual features provides an understanding of how steering is affected under given conditions (i.e., one type of gearset may require less force from the driver to turn the tires than another type of gearset).

Varying terms have been used to describe knowledge of conceptual features, and specifically, the meaningful relationships one may integrate within memory. The terms include “mental models”, “conceptual knowledge”, “schemas”, “cognitive structures”, and “structural knowledge” (Day et al., 2001; Jonassen, Beissner, & Yacci, 1993; Kraiger, Ford, & Salas, 1993; Tennyson & Cocchiarella, 1986). Although these terms have slightly varying definitions, they are, for the most part, based on the assumption that conceptual knowledge is gathered and stored in memory in the form of relational networks (see Anderson & Bower, 1973; Collins & Quillian, 1969; Deese, 1961; Johnson & O'Reilly, 1964; Shavelson, 1972). These relational networks are commonly understood as the mechanisms by which people can interact with their environment (see Rouse & Morris, 1986).

Knowledge Integration and Operator Knowledge

Given that the relational networks that one has stored in memory are used to interact with their environment, it can then be logically concluded that the relational networks that an operator has stored in memory are used to interact with the systems that are being operated. Specifically, the operator uses the relational networks to help with understanding situations that may occur during an operation, and to make predictions about future states of the operation. Furthermore, the stored relationships, if well integrated, facilitate effective and efficient memory retrieval which, in turn, facilitates quicker comprehension, better inferences, and more accurate predictions of future states of an operation (Collins & Gentner, 1987; Day et al., 2001; Rouse & Morris, 1986). Finally, and of most importance, is the quick retrieval of relevant information that well-integrated knowledge facilitates when situations or events outside of normal operations occur (e.g., mechanical failure, human error, or environmental changes) (Cannon-Bowers, Salas, & Converse, 1993). In sum, well-integrated knowledge facilitates cognitive actions that can lead to safe operations.

Representing Integrated Knowledge

Knowledge elicitation methods can be used to represent the relationships an operator has integrated and stored in memory. As mentioned in the introduction, the outcome that is elicited has been referred to as a knowledge structure (KS). KSs delineate hypothetical structures of information related to an operation. The accuracy of an operator's KS is evaluated to determine what the operator understands/misunderstands about the operation. Given that integrated knowledge may influence both how well an operator understands the functions of an operation, and an operator's success at predicting future operational requirements (Kraiger et al., 1993; Rouse & Morris, 1986), the accuracy of an operator's KS, to some degree, represents an operator's ability to conduct safe operations. For example, pilots may exhibit safe flight maneuvers when they display more accurate KSs of flight dynamics, flight procedures, and aircraft components (Curtis et al., 2007).

Knowledge Structure Elicitation Techniques

Developing techniques for eliciting KSs has been a major focus of knowledge elicitation research (e.g., Boehm-Davis, 1989; Cooke, 1999; Day et al., 2001; Hoffman, Shadbolt, Burton, & Klein, 1995; Klein, Calderwood, & MacGregor, 1989). These techniques have been referred to as conceptual knowledge elicitation techniques (Cooke, 1994). Their outcome often resembles a network of interrelated terms that define the domain or operation that is being represented (see Figure 2).

In comparison to techniques for the elicitation of other types of knowledge, such as verbal reports, interviews, and process tracing, conceptual knowledge elicitation methods require very little intervention, intuitions, and/or judgments by the administrator (Cooke, 1994). Furthermore, conceptual knowledge elicitation techniques reduce the need for the subjective interpretation of the large amounts of data that is often collected from observations, interviews, or other typical classroom techniques, such as writing compositions or essays. Essentially, conceptual knowledge elicitation techniques attempt to provide a condensed, objective representation of a learner's knowledge that is free of administrator bias.

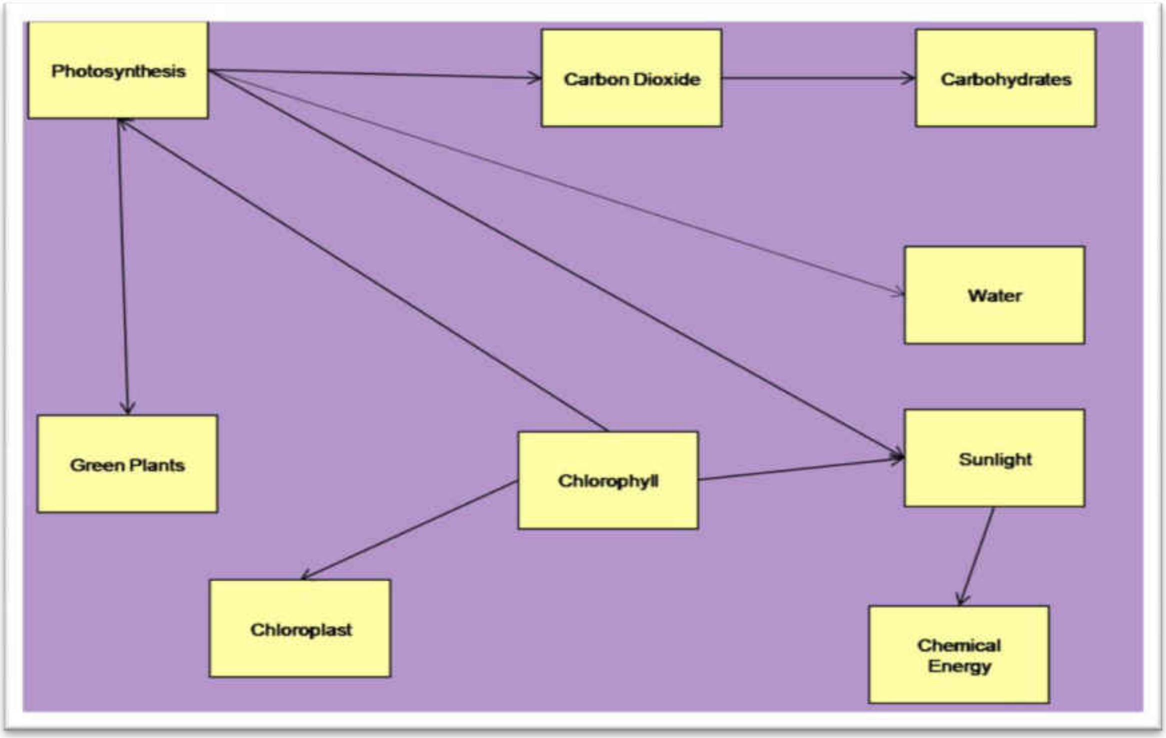


Figure 2. Depicts an example of an outcome from a conceptual knowledge elicitation technique.

> Note: The knowledge structure here represents the person's understanding of the relationships between concepts related to photosynthesis in plants.

Various KS elicitation techniques have been developed. The outcomes from these techniques only vary slightly in appearance; however, how the outcome is obtained may vary drastically, to the point of where the outcomes represent different information about a learner's knowledge. In fact researchers have suggested that the different elicitation techniques depict different aspects of a KS, and thus multiple elicitation techniques should be used in combination (Shavelson, Ruiz-Primo, & Wiley, 2005) (see also Cheatham & Lane, 2002; Evans, Jentsch, Hitt, & Bowers, 2001; Hoffman et al., 1995).

Two examples of commonly used elicitation methods are to obtain pairwise relatedness ratings and concept mapping. Pairwise ratings requires the respondent to judge the strength of the relationship within concepts presented as pairs (Cooke & McDonald, 1987; Kraiger et al., 1993; Shavelson, 1972). More specifically, users may be presented with 55 pairs of concepts formed from 10 concepts, and asked to rate their similarity on a scale from “1” (not related) to “7” (highly related). Once all pairs of concepts are rated, the ratings can be transformed into a network using a scaling algorithm, such as the Pathfinder algorithm (Davis, Curtis, & Tschetter, 2003; Anna L. Rowe, Cooke, Hall, & Halgren, 1996; Schvaneveldt, 1990).

More specifically, scaling algorithms, such as Pathfinder transform a proximity matrix into a network of concepts where the links indicate the semantic distance between the concepts, or how closely the concepts are related (Jonassen et al., 1993). When elicited from an operator, these networks can convey (a) how integrated the operator’s knowledge of an operation is, (b) the operator’s understanding of the hierarchical nature of the operation, and/or (c) the operator’s perceived strength/existence of relationships between concepts from other domains (i.e., cross-links) (see Novak & Gowin, 1984).

In contrast to the use of pairwise relatedness ratings, concept mapping requires the operator to directly generate relationships between domain concepts (Jonassen et al., 1993; J. D. Novak & Gowen, 1984; Ruiz-Primo & Shavelson, 1996). There are various methods for administering concept mapping. McClure, Sonak, and Suen (1993), through empirical research, for example, identified an administration method that is less time-consuming than other methods of elicitation, yet still provides an adequate representation of learners’ knowledge structure. For this elicitation method, operators are given a list of concepts (usually between 10 and 20) that are essential for defining the operation(s) under evaluation. Operators then create relationships by drawing links or arrows between the concepts they perceive as being related. In most cases, the user is asked to create labels between the connected concepts that describe why or how the concepts are related (refer to Figure 2). When two concepts are linked and the relationship is labeled, a proposition, or a meaningful statement about an object or event, is formed (J. D. Novak & Canas, 2006). Essentially, the concept mapping method produces a concept map (CM) made up of propositions that define an operation (see Figure 3).

In addition to the concept mapping method being a more direct elicitation method than pairwise relatedness ratings, the outcome from concept mapping provides more information about the relationships than that from pairwise relatedness ratings. In particular, concept mapping is typically administered in a way that encourages the learner during the elicitation process to describe *why* two concepts are related. Descriptive information about the relationship between two concepts can, however, also be obtained from pairwise relatedness rating technique *after* the structure of the knowledge is obtained (Cooke, 1994); however, concept mapping facilitates a more fluid elicitation of knowledge right from the start.

Depicting the contextual information associated with an operator's knowledge may be invaluable not only for evaluating its accuracy, but also for diagnosing any misconceptions the operator may have about an operation (see J. Novak, Gowen, & Johansen, 1983). The propositions that are formed when creating a concept map depict not only connections between concepts, but also what the operator knows about the relationships. Indeed, only knowing that an operator has made connections between concepts may be misleading when the operator has a weak or incorrect understanding of the relationship. Concept mapping, thus, provides additional information about an operator’s knowledge that can be useful for diagnosing misconceptions or misunderstandings

Given the importance of diagnostic evaluations, this study focused on the reliability and validly of KSs elicited using concept maps. As a side note, however, I have not intended here to argue that concept mapping is frequently, or always, better than pairwise relatedness rating methods or other conceptual knowledge elicitation methods. In fact, I advocate instead that more than one method should be used to ensure a thorough evaluation of an operator’s knowledge structure. Furthermore, the pairwise relatedness rating technique has been extensively investigated in operation-based domains (Dorsey, Campbell, Foster, & Miles, 1999; Anna L. Rowe & Cooke, 1995), yet very few studies have investigated the utility of methods (i.e., concept mapping) that elicit and evaluate the contextual information within a KS. For the remainder of this paper, the term “KS evaluation” will refer to the evaluation of the information elicited within a contextual KS elicitation method such as concept mapping.

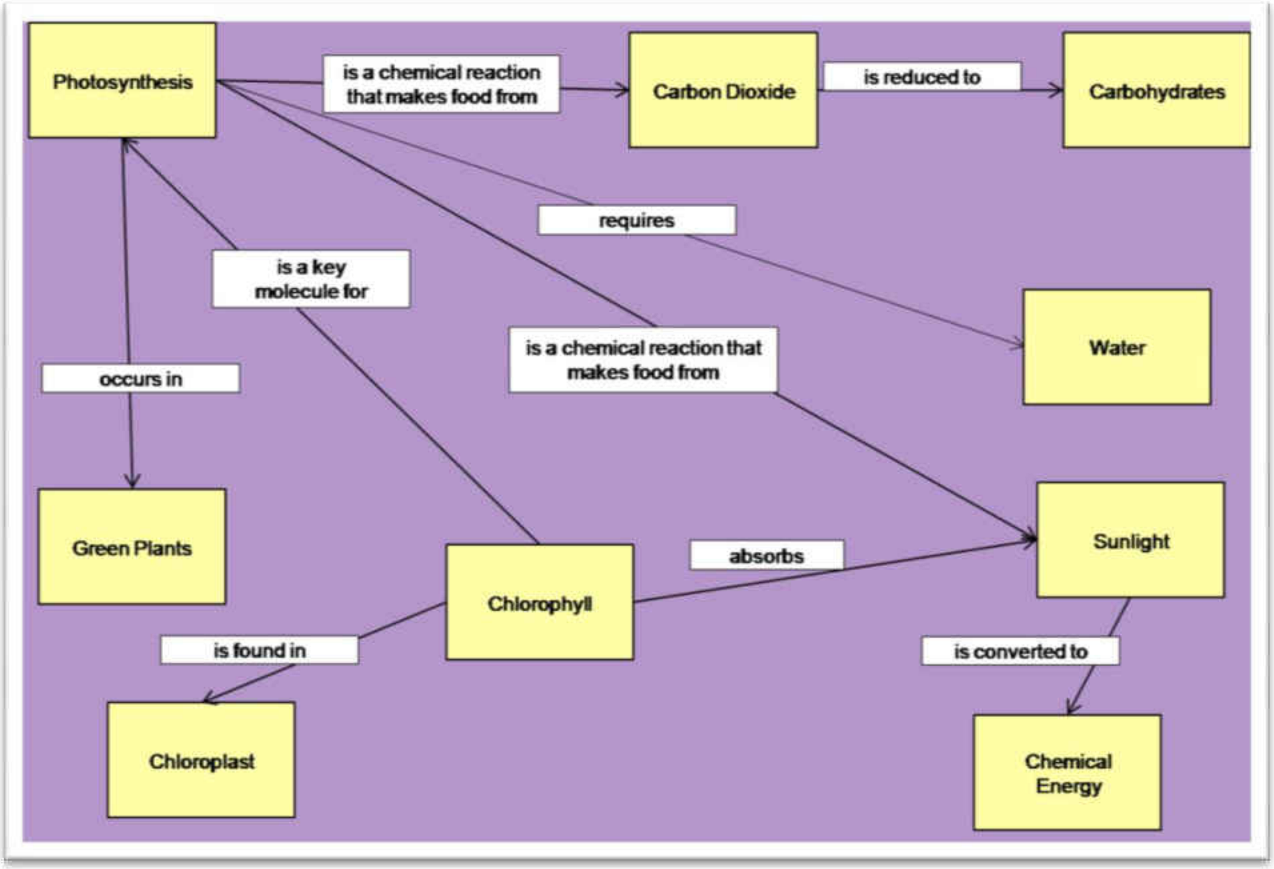


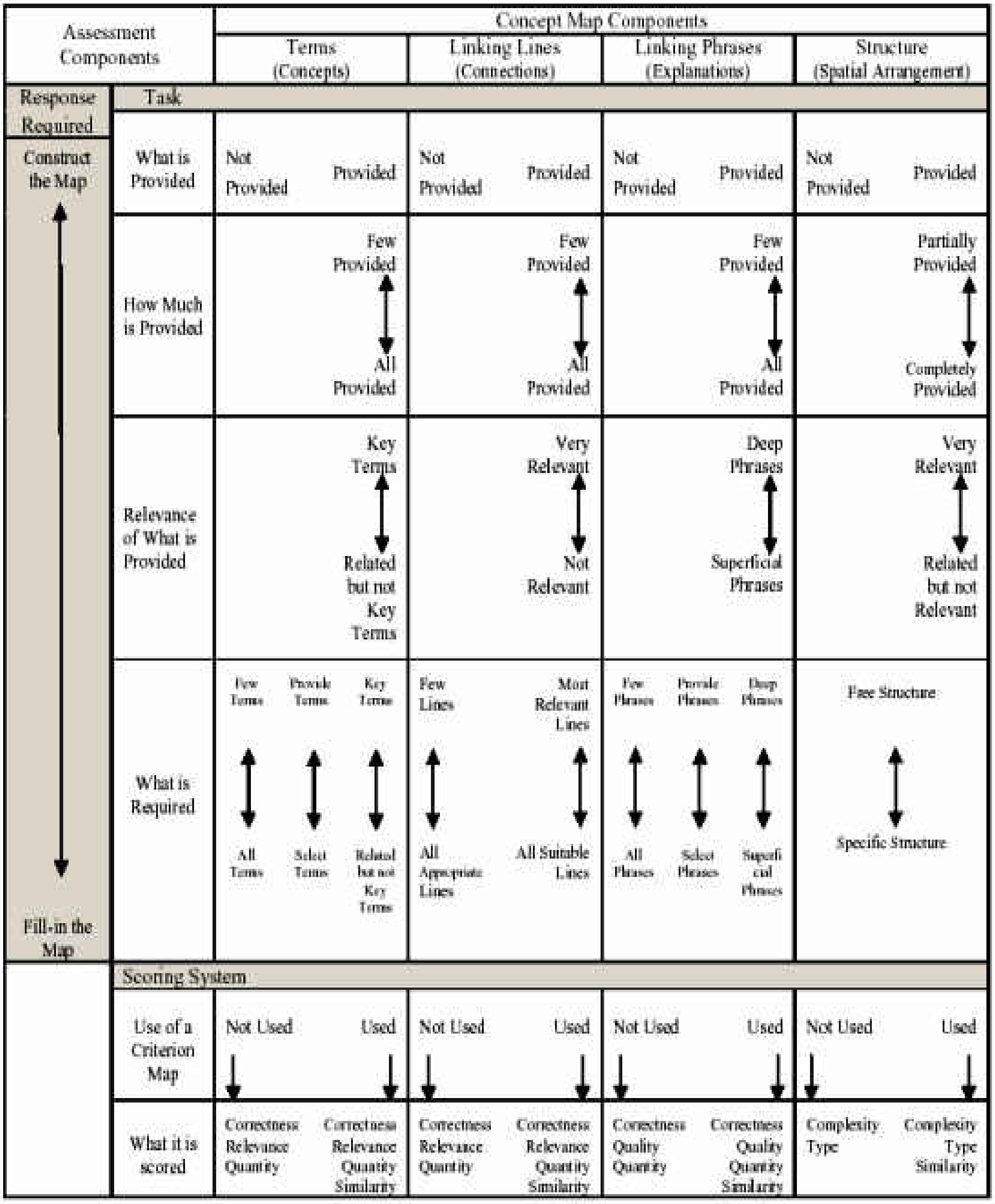
Figure 3. Example of KS with propositions.

> Note: This knowledge structure contains propositions, in comparison to the knowledge structure in Figure 2 that contains only links.

Evaluating Knowledge Structures

Once elicited, KSs are evaluated using conceptual knowledge evaluation methods. How KSs are evaluated is dependent on the method used for the elicitation, and the components within the elicited outcome. For example, Ruiz-Primo (2004) suggested that elicitation methods can be characterized along a continuum from low to high directedness (Ruiz-Primo, 2004; Ruiz-Primo, Schultz et al., 2001; Ruiz-Primo, Shavelson, & Schultz, 1997). Therefore, when choosing an evaluation method, both the directness and the components (contextual and/or structural) elicited must be considered. Ruiz-Primo (2004) used the following example to explain:

If the examinee is to provide the terms, the assessor may decide to score them as correct or incorrect without considering the relevance of the terms. If the amount of terms was not posed as a constraint, the assessor may score the quantity of terms provided (Ruiz-Primo, 2004, p. 2). In light of her findings, Ruiz-Primo (2004) developed a framework for choosing KS scoring methods based on what method was used to elicit the KS (see Figure 4).



Assessment Components		Concept Map Components			
		Terms (Concepts)	Linking Lines (Connections)	Linking Phrases (Explanations)	Structure (Spatial Arrangement)
Response Required	Task				
Construct the Map ↕ Fill-in the Map	What is Provided	Not Provided Provided	Not Provided Provided	Not Provided Provided	Not Provided Provided
	How Much is Provided	Few Provided ↕ All Provided	Few Provided ↕ All Provided	Few Provided ↕ All Provided	Partially Provided ↕ Completely Provided
	Relevance of What is Provided	Key Terms ↕ Related but not Key Terms	Very Relevant ↕ Not Relevant	Deep Phrases ↕ Superficial Phrases	Very Relevant ↕ Related but not Relevant
	What is Required	Few Terms ↕ All Terms Provide Terms ↕ Select Terms Key Terms ↕ Related but not Key Terms	Few Lines ↕ All Appropriate Lines Most Relevant Lines ↕ All Suitable Lines	Few Phrases ↕ All Phrases Provide Phrases ↕ Select Phrases Deep Phrases ↕ Superficial Phrases	Free Structure ↕ Specific Structure
	Scoring System				
	Use of a Criterion Map	Not Used ↓ Used ↓	Not Used ↓ Used ↓	Not Used ↓ Used ↓	Not Used ↓ Used ↓
	What it is scored	Correctness Relevance Quantity Correctness Relevance Quantity Similarity	Correctness Relevance Quantity Correctness Relevance Quantity Similarity	Correctness Quality Quantity Correctness Quality Quantity Similarity	Complexity Type Complexity Type Similarity

Figure 4. Ruiz-Primo's (2004) KS directness framework.

There are multiple methods for evaluating KSs (McClure et al., 1999; Ruiz-Primo & Shavelson, 1996). Ruiz-Primo and Shavelson (1996) categorized these methods by the strategies used to obtain a score/rating that represents the quality of the KS. These strategies included: (a) scoring/rating the components of the KS, (b) comparing the KS to an expert or referent outcome, and (c) using both strategies a and b. To simplify the explanation of how these strategies have been applied, they were collapsed into two groups: (a) evaluating the components of a KS and (b) using a referent map to evaluate the components within the KS. The first strategy is discussed in the following section. The latter strategy is the same as the former, only the rater uses referent materials to assist with implementing the evaluation strategies. It was proposed here that using referent materials during a KS evaluation would mitigate particular rater limitations. This is discussed further in Chapter Three, along with how referent materials have been created and used.

Evaluating the Components of a KS

The components within a KS may include concept, links, and labels. Figure 2 above depicts a KS outcome that is made up of *concepts* and *links*. As discussed above, the concepts are important terms that define a domain and the lines indicate a relationship exists between the concepts. Furthermore, KSs may contain *labels* between each linked concept which describes how the concepts are related (refer to Figure 3). The combination of concepts, links, and labels form statements that define an operation, also referred to as a *proposition.*

The components within a KS can be characterized as either structural or contextual. In this framework, structural characteristics are the linked concepts, whereas contextual characteristics are the label. Like different knowledge elicitation methods (Hoffman et al., 1995), the different components within a KS elicit different types of knowledge. Therefore, the type of evaluation, in terms of its detail or depth, is dependent upon the components available to evaluate. For example, KSs that contain only linked concepts will provide a depiction of the *structural* characteristics within a KS, as opposed to evaluating the propositions which provide a depiction of the *contextual* information within the KS (Yin et al., 2005).

Structural KS Evaluations. Evaluating the structure of a KS will indicate whether an operator can correctly identify relationships that are important for defining an operation (Johnson-Laird, 1980). Examining the linked concepts by, for example, counting the number of correctly linked concepts within a KS indicates how many relationships an operator can correctly identify. Essentially this examination method reveals the *density* of an operator's KS which can distinguish an expert operator, who has a denser network of correctly linked concepts, from a novice who has a less dense network of correctly linked concepts (see Bedard & Chi, 1992).

Structural evaluations may also refer to evaluating features of a KS such as hierarchies. Hierarchical KSs are characterized by super-ordinate concepts at the top and crosslinks between hierarchies. Scores may be assigned based on how many levels of hierarchies are present in the structure, or how many relevant crosslinks there are (J. D. Novak, 1995). Not all KSs, however, have a hierarchical structure. KSs that reflect a hierarchical structure are often developed based on a Learning theory (i.e., Ausubel's theory) which posits that new information is related to and subsumable under, more general, existing information (see J. D. Novak & Gowen, 1984). As a result, the cognitive structure of this information should be elicited (and learned) in a hierarchical manner.

Researchers in support of non-hierarchical structures suggest that KSs are more like semantic networks. They argue that not all domains are suitable for hierarchical structures (Ruiz-Primo & Shavelson, 1996). Furthermore, eliciting a hierarchical structure from an operator may require that the operator has learned the information in a hierarchical manner. Whether KSs containing operation relevant information should be hierarchical in nature is an empirical question that is, however, beyond the scope of this study.

Contextual KS Evaluations. Evaluating contextual information typically entails examining the quality of propositions. For example, each proposition within a KS could be given a rating in accordance with a protocol that considers the correctness of the proposition (see Figure 5) (McClure et al., 1999). Evaluating the accuracy of the labels within a KS is considered a more detailed KS evaluation method. Contextual evaluation can indicate whether the operator correctly understands the relationships between the linked concepts.

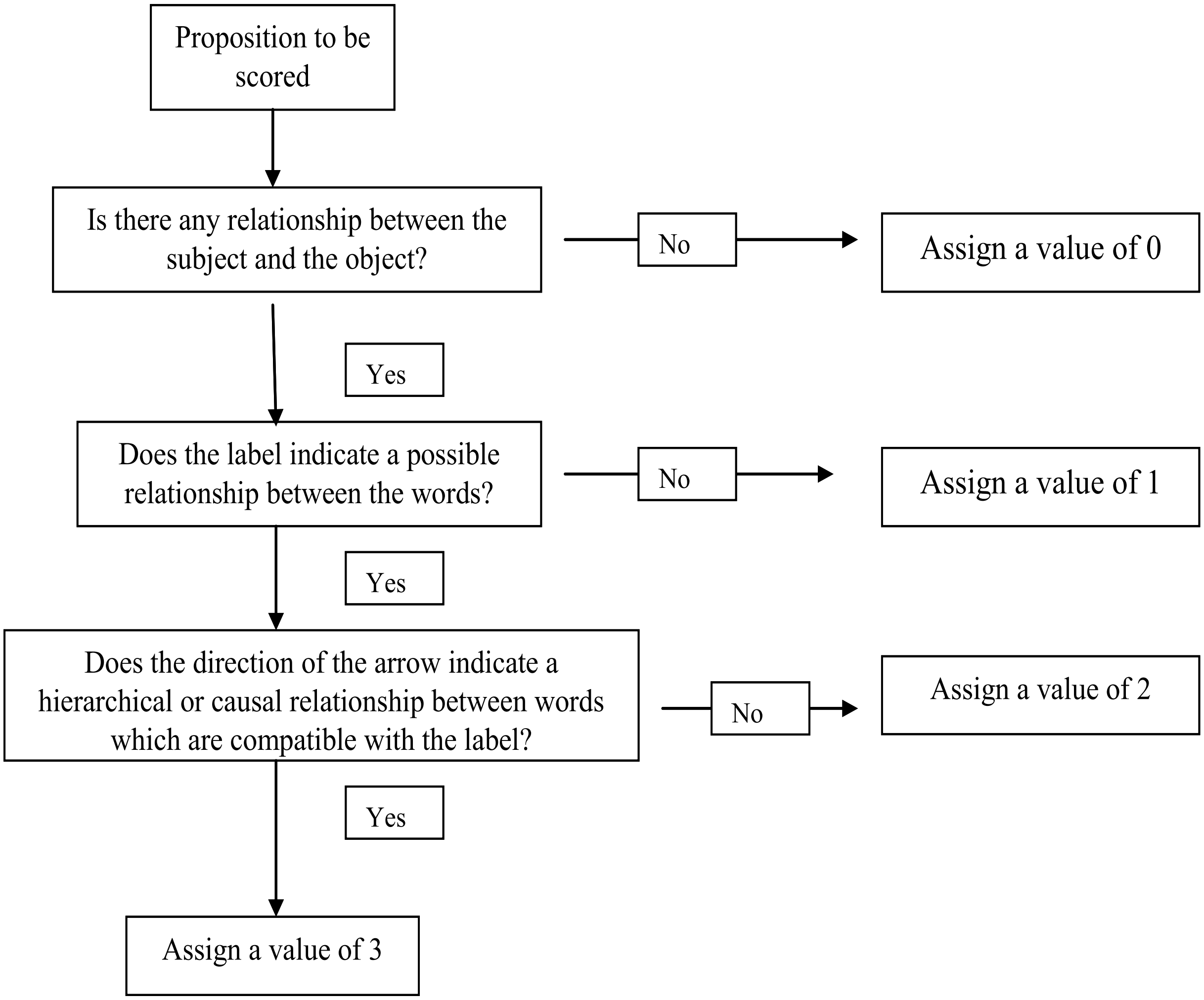

Proposition to be scored
Is there any relationship between the subject and the object?
No
Assign a value of 0
Yes
Does the label indicate a possible relationship between the words?
No
Assign a value of 1
Yes
Does the direction of the arrow indicate a hierarchical or causal relationship between words which are compatible with the label?
No
Assign a value of 2
Yes
Assign a value of 3

Figure 5. McClure and Bell's (1999) protocol for contextual KS evaluations.

Structural vs. Contextual Evaluation. Although knowing how many relationships an operator can correctly identify may provide some insight into his/her knowledge, evaluating the accuracy of the labels that connect the concepts will determine whether an operator correctly understands the relationship. Indeed, researchers have argued that only evaluating the structural characteristics of a person's knowledge will lead to a less accurate depiction of the accuracy or quality of the knowledge in comparison to evaluating the contextual information. As argued before, contextual information may be invaluable for not only evaluating an operator's knowledge, but also for diagnosing any misconceptions the operator may have about an operation (J. Novak et al., 1983).

In comparison to structural evaluations, however, contextual evaluations are more subjective. For example, there is only a finite number of concepts that could be correctly connected within a KS; yet, there are varying descriptions that could correctly represent the relationship between the concepts (West, Pomeroy, Park, Gerstenberger, & Sandoval, 2000). This can be better understood by considering the difference between evaluating multiple-choice tests and essays.

With a multiple-choice test, there is typically one correct or more accurate answer to choose from among multiple wrong or less accurate answers. The evaluation method, therefore, is done by calculating how many times one chooses the correct answer. Essays, in comparison, allow responders to present information from their own perspective which may reflect varying levels of correctness. As a result, raters must judge how correct a description is, sometimes over multiple evaluations. It is here where the limitations of the rater affect the rating process.

Like essays, examining the propositions within a KS requires judging and rating the correctness of conceptual information. As a result, evaluating the contextual information within a KS is susceptible to the same limitations as other subjective evaluation methods (i.e., essays and job performance evaluations). The following section discusses how these limitations have been demonstrated in studies investigating the reliability and validity of KS evaluation methods.

The Psychometrics of KS Evaluations

Ruiz-Primo and Shavelson (1996) conducted a thorough review of several studies that investigated the psychometric properties of KS evaluations. Included in their review was a study by Anderson and Huang (1989) which indicated substantial correlations between KS evaluation scores, education achievement tests and ability tests. Furthermore, Acton (1994) found that evaluation scores for KSs elicited from instructors were higher than the evaluation scores for KSs elicited from students. These studies suggested that KS evaluation methods have concurrent validity and can show known group differences.

From their review, Ruiz-Primo and Shavelson (1996) concluded that many of the studies reported high inter-rater reliability coefficients; however, they warned that these scores must be interpreted with caution as reliability was often calculated on scores produced from only evaluating the structural dimensions (e.g., density) of the KS. The reliability and validity of evaluating the contextual information within a KS, however, was seldom reported. Furthermore, in recent studies researchers have shown that the scores derived from contextual evaluations were less reliable than the scores derived from structural evaluations (see M. E. Harper, Hoeft, Evans, & Jentsch, 2004; West et al., 2000).

Overall Summary

Researchers have recognized the influence that well integrated knowledge structures has on performance (Goldsmith & Kraiger, 1997). When strong, accurate relationships have been formed between concepts that define an operation, then operators are better able to make effective decisions, and quickly problem solve when necessary. These relationships can be depicted in the form of Knowledge Structures (KSs). To effectively measure the accuracy of KSs, contextual KS evaluation methods must be employed which are essential for both evaluation and diagnosis.

KS evaluations often require human evaluation, and thus may be influenced by the characteristics a rater may bring to the evaluation process (e.g., knowledge of the operation or knowledge of the evaluation process). These characteristics may often negatively affect evaluations as they may be in the form of biases (i.e., the halo effect) or limitations (cognitive limitations). The influence of these characteristics is often reflected in the reliability and validity of the evaluation outcome, which in this study refers to the ratings an evaluator assign to represent the quality of a KS.

Very few, if any, studies have investigated how a rater's characteristics may influence the reliability and validity of KS evaluations. The most relevant research is found in the Industrial/Organization Psychology literature, which has extensively focused on improving behavioral ratings, such as those derived from job-performance evaluations. As discussed in Chapter three below, the KS evaluation process may be influenced by the same or similar rater biases or limitations. Furthermore, how these limitations were uncovered and how they can be mitigated is discussed.

CHAPTER THREE: THEORETICAL FRAMEWORK AND HYPOTHESES

Researchers have described the procedural steps for evaluating job performances as including (1) observing a performance, (2) examining the quality of the performance, and (3) rating the performance (Borman, 1978) (see Figure 6). While the KS evaluation process may include the examine and rate steps, it lacks the complexity of the observation stage. In performance evaluations, the observation process includes detecting, perceiving, and recalling or recognizing a specific behavioral event (Woehr & Huffcutt, 1994). In KS evaluations, the information is presented in a single instance; therefore, detection and perception of the information is unnecessary. Furthermore, the process does not require an evaluation based on more than one instance, only the instance presented within the KS at the time of the evaluation. Therefore, a rater's ability to detect, perceive, and recall/recognize a behavior is inconsequential to KS evaluations.

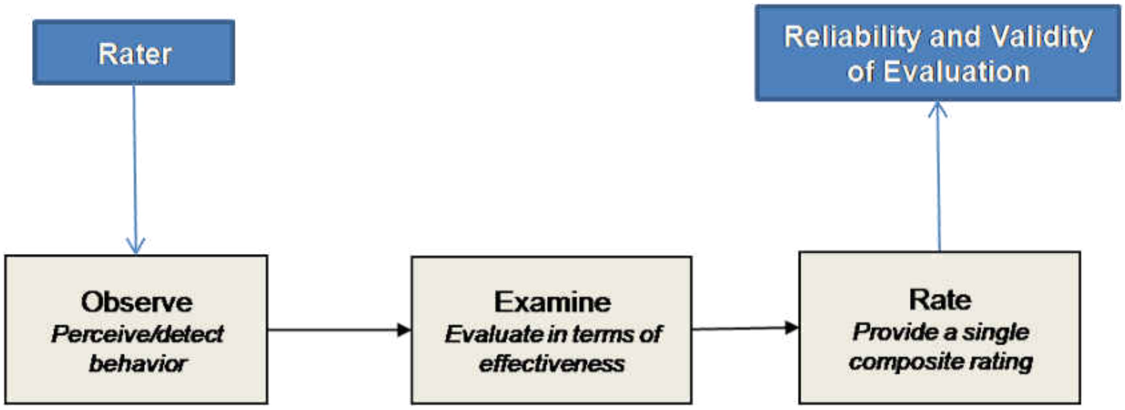

Rater
Reliability and Validity of Evaluation
Observe
Perceive/detect behavior
Examine
Evaluate in terms of effectiveness
Rate
Provide a single composite rating

Figure 6. Stages of Borman's performance judgment process.

Knowledge Structure Evaluation Procedure

In place of the three step performance judgment process described above, a two step process which includes only the Examine and Rate steps (see Figure 7) was used in order to delineate KS evaluation procedures. In this procedure, the examine step is when the performance is examined in terms of the effectiveness it represents. Similarly, the examine step in KS evaluations is where the quality of the content within the KS is examined. For job performance and KS evaluates, the result of the examine process is depicted during the Rate step. In other words, during the Rate step, a single outcome (i.e., rating) is derived to represent the outcome of the Examine step. As depicted in Figure 7, the rating reflects the reliability and validity of the KS evaluation.

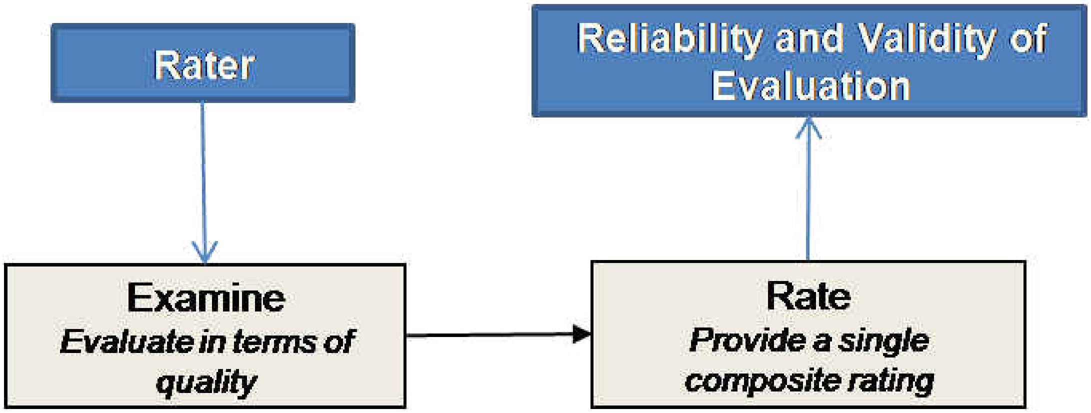

Rater
Reliability and Validity of Evaluation
Examine
Evaluate in terms of quality
Rate
Provide a single composite rating

Figure 7. KS evaluation procedures.

Guiding Frameworks

To reiterate, the goal of this study was to identify methods that improve the reliability and validity of KS evaluations. Therefore, it was necessary to identify factors that influence the KS evaluation procedures described above. Two theoretical frameworks were used to guide the identification of these factors, including Landy, James, and Farr's (1980) process model of job performance ratings and Baddeley's (1981) working memory model.

Performance Evaluation Process Model. The performance rating process model delineates the subsystems that form the overall job performance rating process. The main assumption of the model is that the rater brings certain characteristics (e.g., domain knowledge and cognitive capacities) to the rating process which inevitably influences the rating outcome. The "rating process" component within the model contains two subsystems, the cognitive process of the rater and the administrative rating process of the organization. This effort specifically focuses on the former subcomponent; however, it is recognized that organizational influences must also be acknowledged and investigated.

Baddeley's Working Memory Model. The second framework used for this effort was Baddeley's (2000) working memory model. In general, working memory models propose a system with limited capacity which temporarily stores information that is necessary to complete a task. A more current working memory model is composed of an episodic buffer that can be conceived of as an interface between the various components of working memory and LTM (Baddeley, 2000). Based on this model, it could be assumed that during the KS evaluation process the episodic buffer plays a role in retrieving operation specific information from long term memory; and furthermore, temporarily stores the information while the central executive component of working memory uses the information to form a mental model that can be used to decide the quality of the information presented in the KS (to assign a rating). Figure 8 depicts the working memory process as it is assumed to occur during KS evaluations. First, the rater makes a decision by attending to the information within the KS, once it is attended to, then the rater searches long term memory for relevant information, retrieves and stores the information in the episodic buffer while the central executor forms a mental model to use for determining what rating to assign.

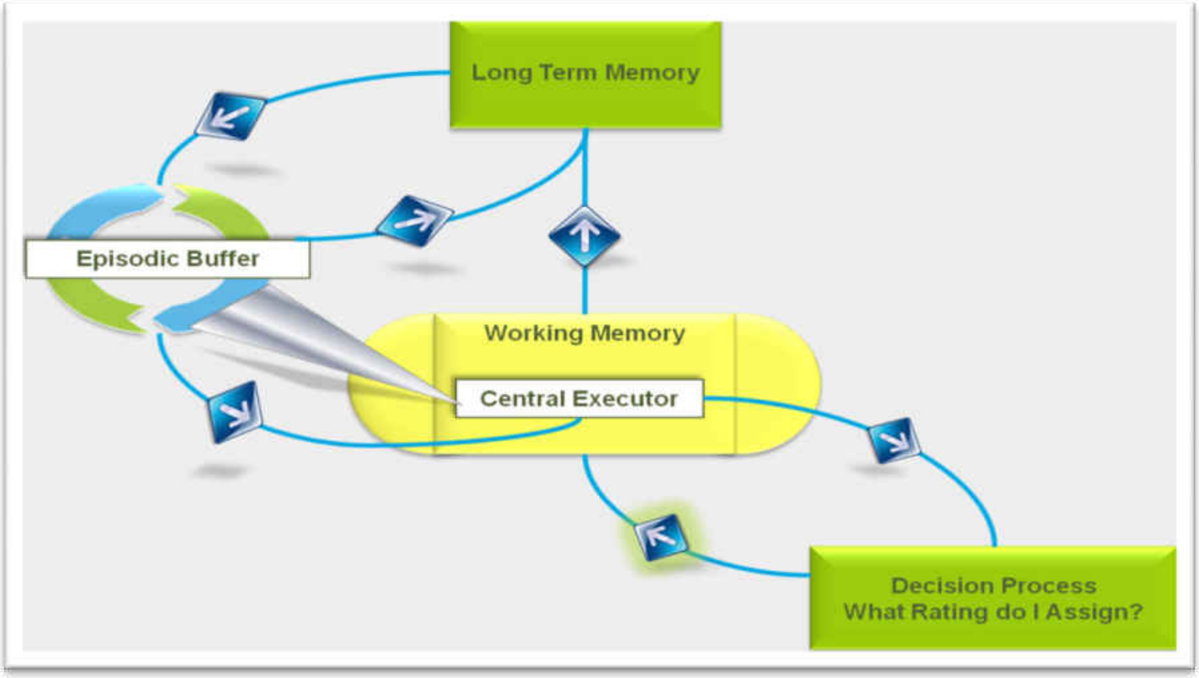


Figure 8. Model of working memory process during the KS evaluation process.

Experimental Framework

From these two models, the framework depicted in Figure 9 was delineated. Essentially, the framework represents the characteristics (i.e., biases and limitations) a rater may bring to the evaluation process, and the processes within working memory that may be affected by those characteristics. The KS evaluation framework guided the hypotheses presented in the following Chapter.

The hypotheses presented below propose how the reliability and/or validity of KS evaluations are affected by specific mitigation techniques. Reliable KS evaluations are defined here as the ability of an assessor to produce a rating that consistently represents the quality of an operator's KS; in particular, the reliability within a rater's ratings of the same information within a KS. Validity is defined as the ability of an assessor to provide ratings that converge with ratings derived by raters considered to be domain experts.

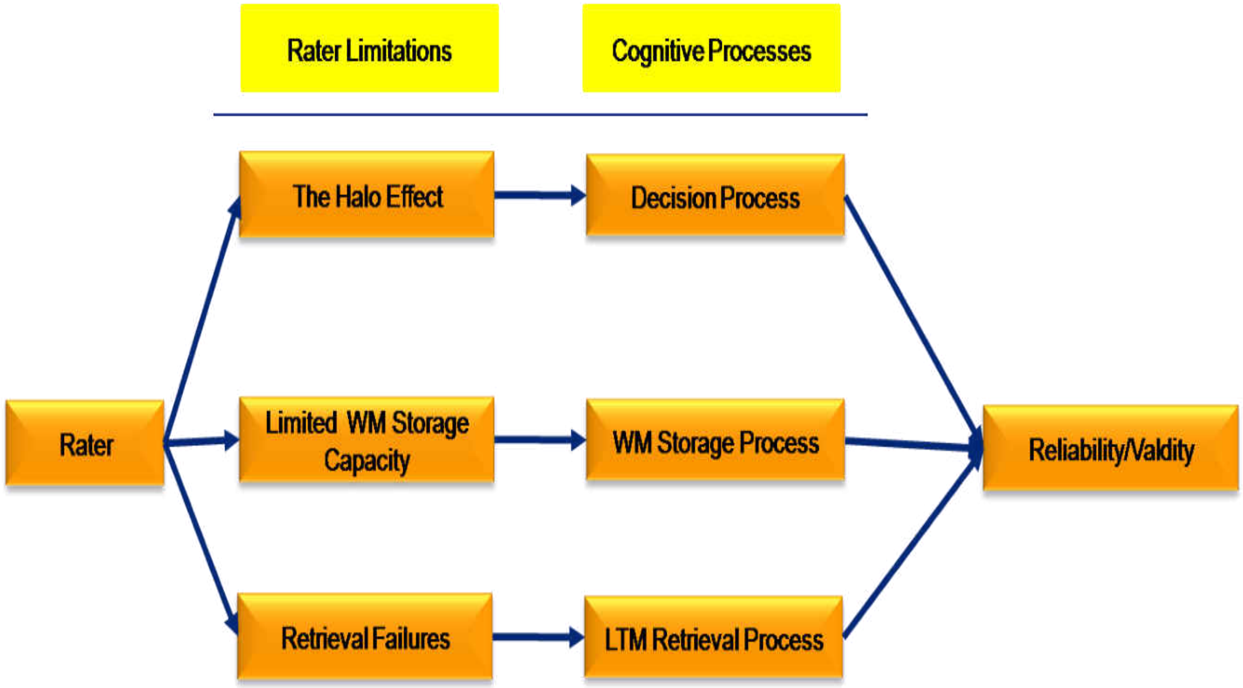


Figure 9. Framework of knowledge structure evaluation process.

The Halo Effect and the Decision Process

As described in Chapter Four, the decision process occurs when the rater uses the mental model formed in the central executor to decide what rating to assign. The decision that is made may be influenced by the contents of the mental model, and also, by the tendency of a rater to exhibit idiosyncratic behaviors that result in rater error. Rater errors are often associated with the biases of a rater (Borman, 1978; Kavanagh, MacKinney, & Wolins, 1971; Landy & Farr, 1980; Weekley & Gier, 1989). Rater error has been extensively studied in the context of performance ratings in industrial organizations (Viswesvaran, Schmidt, & Ones, 2005), and to a lesser extent in the context of conceptual ratings in education environments (Eckes, 2008; G. Engelhard, Jr., 1994). Researchers have identified various categories of rater error including severity/leniency, central tendency, restriction of range, and halo (Saal, Downey, & Lahey, 1980). All of these errors should be addressed when studying the reliability of a measurement. As discussed above, one key aspect of KS evaluations is the multiple dimensions of knowledge that are represented within the KS (e.g., structural and contextual). This characteristic makes KS evaluations particularly susceptible to the rater error, referred to as the halo effect.

Halo and KS Evaluations

The halo effect, is an error commonly addressed throughout past research on both performance ratings and conceptual ratings (Carter, Haythorn, Meirowitz, & Lanzetta, 1951; Cooper, 1981; Dennis, 2007; Thorndike, 1920). This effect has differing manifestations depending on the context of an evaluation. For behavioral ratings, halo often refers to the tendency of a rater to evaluate an individual's performance on the merit of that individual, rather than on the actual performance being evaluated (Thorndike, 1920). In educational settings, where more conceptual information such as essays or compositions is evaluated, halo refers to the tendency of the assessor to apply a singular approach to the evaluation, when a multidimensional approach is more appropriate; for example, raters may tend to provide general ratings based on a subset of dimensions within a an essay (e.g., context, structure, or syntax) rather than spreading their evaluations out evenly across all relevant dimensions (Eckes, 2008; G. Engelhard, 1994).

KSs, like performance behaviors and essays, often represent multiple dimensions of one's integrated knowledge. As discussed previously, KSs may contain both structural characteristics (e.g., density) and contextual characteristics (e.g., accuracy). These characteristics can be broken down even further; for example, contextual information represents both the relevancy and accuracy of one's KS; and, structural may represent the density and hierarchical nature of one's KS.

To decide what rating to assign, a rater may conduct a general evaluation of a KS by considering only the relevancy of the content or only the structure of the content; in this case, he/she is demonstrating the halo effect. Failure to evaluate the multiple dimensions within a KS may lead to an over-/under-estimated representation of an operator's knowledge. For example, if during the evaluation process a rater examines how many concepts are accurately linked, the rating will then represent only the structural dimensions of the operator's knowledge, which can be misleading if the operator does not accurately understand the relationship between the concepts that are linked.

Demonstrating the halo effect within KS evaluations suggests that a rater does not have an accurate conceptualization of how to effectively evaluate KSs. In this case, the rater may not only provide misleading ratings, but may decide what rating to assign based on different rating criteria across different evaluations. Given this, the reliability of KS ratings may be dependent on whether the decision process is affected by the halo effect (see Figure 10).

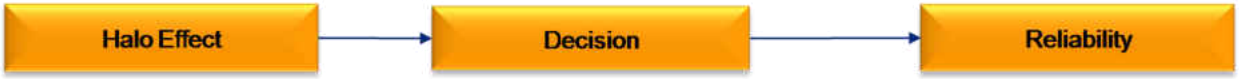


Figure 10. Depicts the reliability of KS ratings as being dependent on whether the decision process is influenced by the halo effect.

Mitigating Halo

Within the behavioral evaluation literature researchers have successfully reduced the halo effect using training methods referred to as rater error training (Bernardin, Bernardin, & Walter, 1977; Woehr & Huffcutt, 1994). Originally, rater error training was used to mitigate not only halo but also the errors mentioned above (e.g., severity, leniency, and central tendency). The goal of the training was to familiarize raters with the concept of rater error. This was achieved by identifying different types of rater errors and describing why these errors may occur (Bernardin, 1978; Borman, 1978; Woehr & Huffcutt, 1994). This training paradigm was based on the assumption that the accuracy of performance ratings could be established by training raters on how to evenly distribute their ratings across a rating scale.

More recently, researchers have suggested that training which focuses on appropriately distributing ratings across a scale is less effective than theory-based rater training (Lievens, 2001; Schleicher, Schleicher, Day, Mayes, & Riggio, 2002). In particular, researchers have demonstrated the effectiveness of a training referred to as Frame-of-Reference (FOR) training. FOR training has been shown to reduce rater error within the context of (a) behavior-based performance measures such as instructor performance (Uggerslev & Sulsky, 2008) and (b) performance-based ratings for both workers' job competencies (Lievens & Sanchez, 2007) and management competencies (Schleicher, Schleicher, & Day, 1998).

Frame-of- Reference Training

FOR training emphasizes a theory of performance in terms of the dimensions that define the performance (Lievens, 2001; Schleicher et al., 2002; Uggerslev & Sulsky, 2008). For example, the theory of performance for evaluating an instructor's effectiveness may be defined by dimensions such as presentation skills, lecture organization, and lecture content. Essentially, FOR training defines dimensions that are important for the evaluation, and also provides examples of effective behaviors related to the dimensions (Lievens, 2001). This approach encourages the assessor to evaluate dimensions that effectively represent performance.

In behavioral performance ratings, FOR training provides raters with a conceptualization of what dimensions within a behavior should be rated, which in turn, helps the rater adequately rate the dimensions within the observed performance (Sulsky & Day, 1994). For KS evaluations, FOR training would provide raters with a conceptualization of what dimensions should be evaluated and provide examples of what ratings would be assigned to varying levels of quality. For example, raters should be provided with propositions that accurately explain phenomena relevant to the operation; or, may inaccurately explain relevant phenomena.

Studies have shown that FOR training is more effective at increasing the accuracy of behavioral performance ratings, as compared to traditional rater error training (Woehr & Huffcutt, 1994). Additionally, FOR training provided a deeper level of processing which resulted in more retention of the training material, as compared to traditional training methods (Athey & McIntyre, 1987; Sulsky & Day, 1994); In sum, FOR training enforces a conceptualization of the rating process that is resistant to decay, over time.

Given the prior success of FOR training, I proposed that the application of this training paradigm to KS rater training would lead to the mitigation of the halo effect within KS evaluations. More specifically, KS FOR training would provide a deeply encoded conceptualization which allowed the rater to decide what rating to assign using the same rating method across multiple evaluations (see Figure 11). Therefore,

H1: after FOR training, raters would produce more consistent ratings than before training

H2: after FOR training, the reliability of the ratings would remain significantly higher than before training, a day following the training.

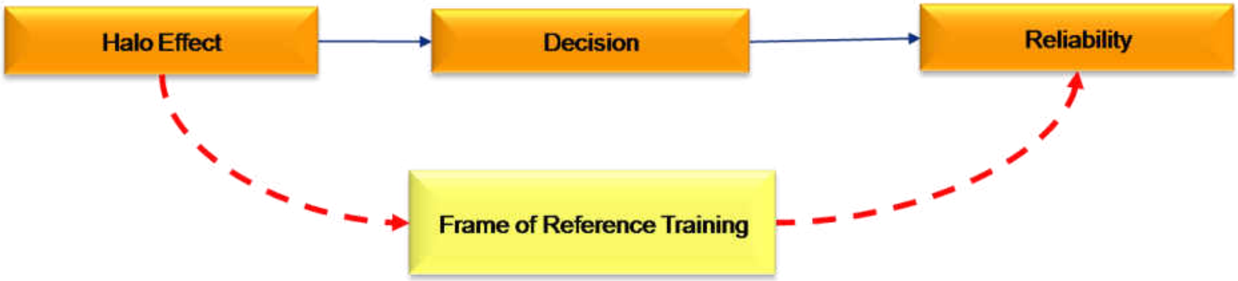


Figure 11. Depicts the mitigating effect of FOR training on the reliability of KS ratings.

Cognitive Demands on the Retrieval and Storage Process

Very few studies have investigated how cognitive demand may influence the reliability of KS evaluations (Plummer, 2008). For example, how the complexity of a KS affects the storage process that occurs during the overall evaluation process. Researchers have, however, suggested that when narrowing down reliable and valid KS techniques (both elicitation and evaluation techniques) one must consider the cognitive demands required for the task (Ruiz-Primo et al., 1997). For example, McClure et al. (1999) suggested that the reliability of scores, assigned to concept maps by raters, is related to the cognitive complexity of the evaluation process used to derive those scores.

McClure et al. (1999) referred to cognitive complexity as the demand the evaluation process places on the rater's cognitive processes (e.g., the storage and retrieval process). They argued that as the cognitive demands of the process increases, the reliability of the ratings decreases. Other than a limited description of the potential demand on working memory, there has been no explanation of how the cognitive demands of a KS evaluation influence reliability and validity. Here, an explanation was derived by delineating the cognitive processes that may occur during the evaluation process. More specifically, the processes involved with the temporary storage of information in working memory and the retrieval of information from long term memory. As mentioned previously, this explanation was used to help better understand the KS evaluation process, and to determine methods for mitigating the cognitive demands of the evaluation.

Storage Process

Within Baddeley's (2000) working memory model described in Chapter Three, the storage process refers to the temporary storage of information within the episodic buffer that occurs when the central executor forms a mental model. Given the limited amount of information a rater is capable of storing at one time, the amount of information a rater must store during an evaluation may affect the rating process. More specifically, if the capacity of the temporary store is exceeded, then information retrieved from long term memory may be "kicked out" or not reach the storage cycle. This will then affect what information is used to create the mental model in the central executor. If this occurs over time, then the rater may form mental models composed of different information about the domain across multiple evaluations. Therefore, the reliability of KS ratings may be dependent on whether the storage process is affected by the amount of information the rater must store during the evaluation process (see Figure 12).

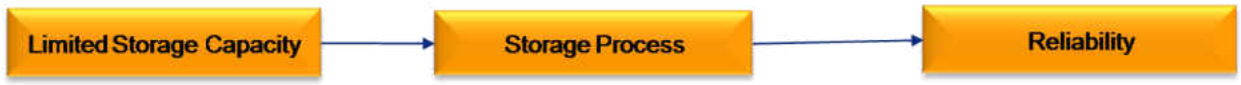


Figure 12. Depicts the reliability of KS ratings as being influenced by the effect that a rater's limited storage capacity has on the storage process.

Mitigating the Effects of a Limited Storage System

Reducing the complexity of a KS may reduce the amount of information the rater must hold in memory during the KS evaluation process. More specifically, limiting the number of concepts that are represented within the KS may alleviate the demands placed on the storage process. As a result, raters can maintain the information within the storage system that has been retrieved from long term memory, while continuing to add new information for the creation of the mental model in the central executor.

Reducing the number of concepts, however, must not interfere with the accurate representation of an operator's knowledge or the rater's ability to identify any misconceptions about the operation. This presents a catch-22. Specifically, limiting the number of concepts used to create a KS may limit the development of a KS that accurately depicts knowledge (Novak, 2006); having too many concepts however, may limit the rater's ability to accurately assess the KS.

To my knowledge, there was limited to no research on exactly how many concepts should be used to effectively measure an operator's KS. Researchers have speculated that 15 to 20 concepts would suffice (J. D. Novak & Canas, 2006). Yin (2005) suggested that KS elicitation and evaluations are more manageable when they are limited to between 8 and 12 concepts. After reviewing 15 studies that used concept map methods to elicit or evaluate KS, I found that the number of concepts ranged from 9 to 36. Six of these studies used between 10 to 15 concepts, three ranged from 15 to 20, five used above 20, and one used below 10.

An, obvious, lack of consensus on how many concepts should be used to depict a KS exists. The decision should, more than likely, be based on characteristics of the domain of interest. The point here, however, is that if raters have less complex KSs to rate, then the demands placed on a rater's storage process may be mitigated or eliminating, thus, resulting in more consistent ratings (see Figure 13). More specifically,

H3: raters who had less complex KSs to evaluate would produce more consistent ratings than raters who had more complex KS to evaluate.

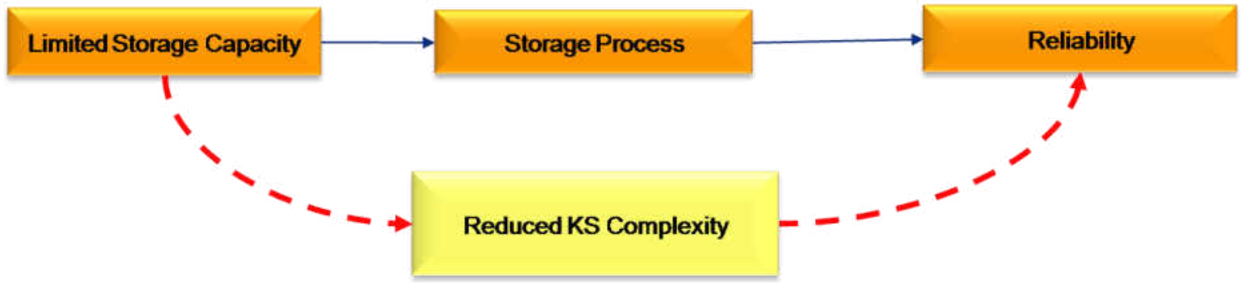


Figure 13. Depicts the mitigating effect of reduced complexity on the reliability of KS ratings.

Long Term Memory Retrieval Process

The retrieval process that occurs during the knowledge structure (KS) evaluation process may be affected by retrieval failures. A retrieval failure is defined by the degree to which a rater can access information in long-term memory (Tulving & Pearlstone, 1966). One type of retrieval failure may result from failing to recall information that has been encoded in long term memory. This failure may be influenced by whether cues are available to trigger recall (Tulving & Thomson, 1973). Another type of retrieval failure may result from the information not being available to retrieve, and can be influenced by the degree to which the information was initially encoded in memory (Fisher & Craik, 1977). In either case, retrieval failure may lead to the rater recalling different or irrelevant information across multiple evaluations. Therefore, both the reliability and validity of KS evaluations may be dependent on whether retrieval failures influence the retrieval process (see Figure 14).

Retrieval Failures
Retrieval Process
Reliability

Figure 14. Depicts how the reliability of KS ratings is dependent on whether retrieval failures affect the retrieval process.

Mitigating Retrieval Failures with Referent Material

As mentioned above, retrieval may be influenced by cueing availability. Therefore, providing raters with material that contains information about a domain during the evaluation process may reduce retrieval failures. Essentially, the information would serve as a trigger for recalling information stored in long-term memory.

Referent material for KS evaluations has been represented in many forms. For example, McClure et al. (1999) developed referent material, referred to it as a "master map", by creating a concept map depicting propositions of that were considered ideal for defining the domain. This is a common type of referent material that is often developed using an aggregation of several experts' KSs.

Referent material of this type has been used to evaluate KSs by calculating the overlap or correlation between the contents of the referent KS and the contents of the operator's KS (Acton, Johnson, & Goldsmith, 1994). For example, researchers have calculated the overlap between the number of links within a KS and the number of links within the referent KS (M. Harper, Evans, Hoeft, & Jentsch, 2004; M. E. Harper, Schuster, Hoeft, & Jentsch, 2008). This method is effective for assessing the structure of a KS however for more contextual evaluations the referent material must assist a rater with the evaluation process. This is achieved when the referent material acts as a cue; thereby, triggering the recall of information that may otherwise be unattainable from long term memory. Having the referent material available for each assessment should therefore allow raters to consistently access information to use for rating KSs across multiple evaluations (see Figure 15). More specifically,

H4: raters who use referent material during a KS evaluation will produce more consistent ratings than those who do not use referent material.

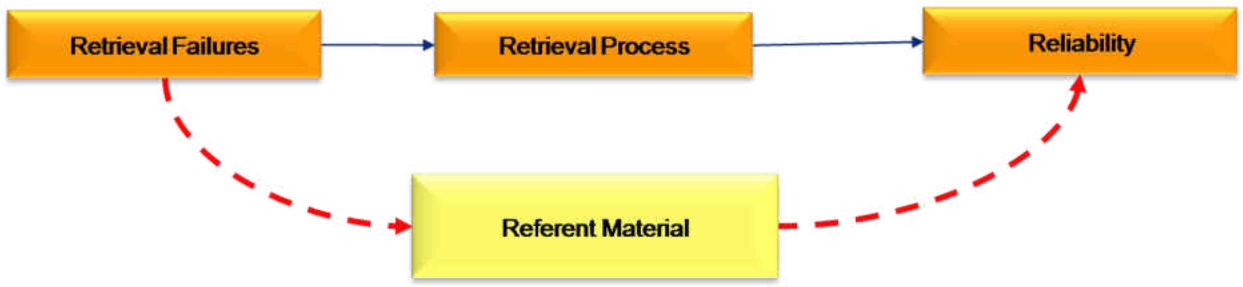


Figure 15. Depicts the mitigating effect of reduced complexity on the reliability of KS ratings.

Mitigating Retrieval Failures with Domain Knowledge

During the KS evaluation process, the information a rater retrieves from long term memory is based on their knowledge and understanding of the operation. Most theories on memory retrieval are based on activation models where concepts in a semantic network are activated by a source. While examining the contents of a KS, the semantic networks related to the operation or specific aspects of an operation should activate. According to Anderson (1983), activation of the concepts related to the source is a function of the strength between those concepts. In other words, more activation will occur between concepts that have stronger and closer relationships to the source. Therefore, a rater's ability to retrieve accurate information about an operation should be related to the amount of accurate and relevant information that can be activated. As a result, during the KS evaluation process, limited to no activation will result in the rater failing to retrieve information that may be used in the mental model development process. This failure has implication for both the reliability and validity of the KS ratings. Therefore, both the reliability and validity of KS evaluations may be dependent on the degree to which domain knowledge can be activated.

As one gains knowledge and experience with an operation, the interconnections between concepts associated with the operation become stronger (Glaser & Bassok, 1989). Minimal to no knowledge or experience with an operation, will result in limited to no connections between concepts. As a result, someone with more knowledge of an operation should be able to successfully retrieve relevant information from their memory, while one with little to no operation knowledge will have minimal to no retrieval. In the former case, the rater has a more accurate conceptualization of the operation that could lead to both more accurate (H5) and more consistent (H6) ratings; particularly, when compared to raters in the latter case who may use different, irrelevant information across multiple evaluations (see Figure 16). More specifically,

H5: raters with more knowledge of an operation would produce more consistent ratings than raters with little to no knowledge of the operations.

H6: raters with more knowledge of an operation would produce more accurate ratings than raters with little to no knowledge of the operations.

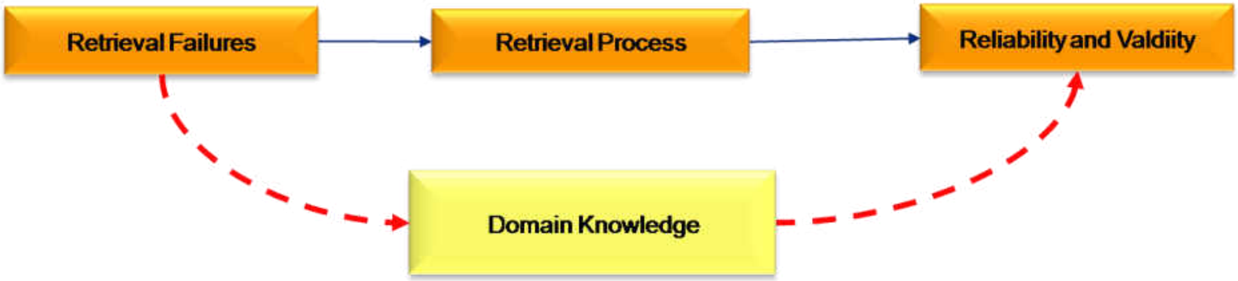


Figure 16. Depicts the mitigating effect of domain knowledge on the reliability of KS ratings.

Overall Summary

The framework described in Chapter Three led to the discovering of one factor (domain knowledge) and three techniques (i.e., FOR training, reducing the complexity of a knowledge structure, and providing referent material) proposed to mitigate or eliminate errors related to the limitations of raters. Figure 17 summarizes the hypotheses of how each mitigation method will influence reliability/validity. Chapter Four below describes the study that was conducted to investigate whether: (a) a FOR training, which focuses on how to evaluate KS, would mitigate rater errors associated with the halo effect; and (b) whether reducing the complexity of a KS; (c) providing a rater with referent materials; and (d) having more knowledge of the operation being evaluated would mitigate the demands on the cognitive processes that occur during the evaluation process.

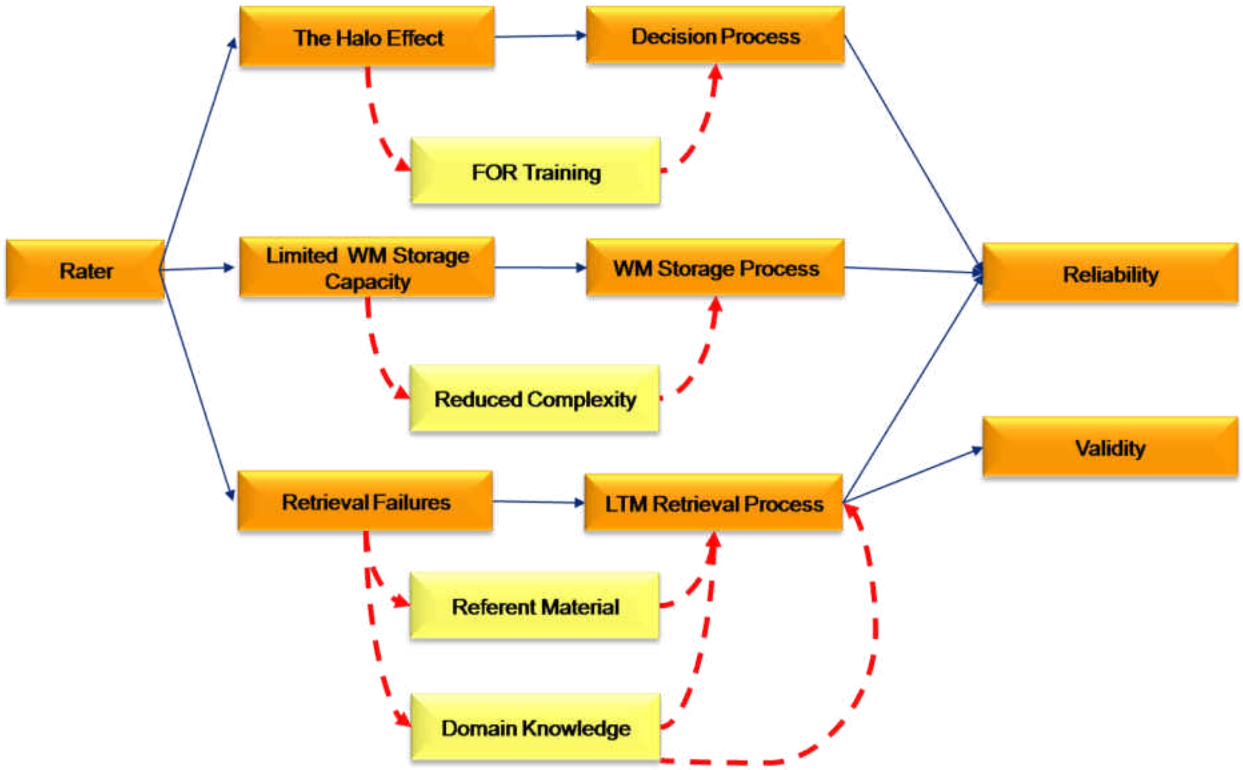


Figure 17. Summary of hypotheses.

CHAPTER FOUR: RESEARCH DESIGN AND METHODOLOGY

Design

Based on the hypotheses presented above, two between-subject factors and four repeated-measures were used. The between-subjects factors were (a) referent map at two levels (referent vs. no referent map), and (b) KS complexity at two levels (KSs with 7 concepts vs. KSs with 10 concepts). The within-subjects factors were four sets of KS evaluations. This resulted in a 2 x 2 x 4 mixed-model design.

Participants and Experimental Operation

Ninety-three volunteer participants evaluated KS outcomes that defined the process and mechanics of steering an automobile. This operation was chosen based on the assumption and prior observations that the population sampled would have a range of experience with and understanding of an automobile's operations.

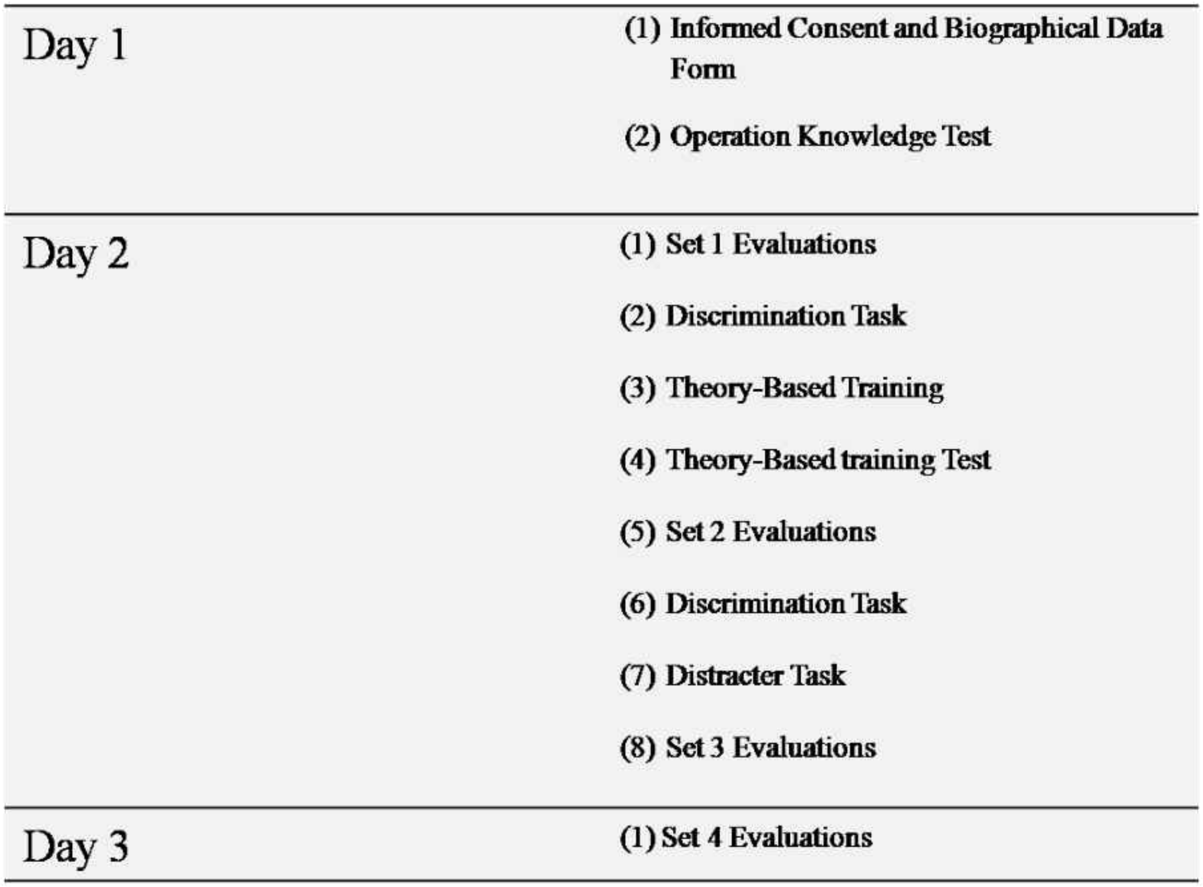
Undergraduates seeking course credit in their psychology classes constituted the sample for this study. The sample was comprised of 75 females and 18 males who ranged in age from 18 to 42 years (*M* = 20.25, *SD* = 3.58). Fifty-five of the participants had no knowledge of automobile mechanics, 22 had a basic understanding, and 16 had a moderate to intermediate understanding of automobile mechanics. No one reported having an expert understanding of automobile mechanics. Out of the 38 participants who reported they had some automobile mechanics knowledge, 20 reported having some understanding of steering mechanics. Number of years driving ranged from 0 to 12 (*M* = 5.05, *SD* = 2.31). Finally, number of days driven per week ranged from 0 to 7 (*M* = 5.05, *SD* = 2.31).

Data Collection Schedule and Activities

Data collection occurred over three consecutive days (see Table 1). During Day Two, the data was collected in a laboratory. On Day One and Day Three, the data was collected over the internet. For Day One, participants first completed biographic forms and then an operation knowledge test. On Day Two, participants completed four sets of KS evaluations, a discrimination task, and the FOR training. For Day Three, participants completed one more set of KS evaluations.

Table 1

Administration Procedures for Study across Three Days



Day 1	(1) Informed Consent and Biographical Data Form (2) Operation Knowledge Test
Day 2	(1) Set 1 Evaluations (2) Discrimination Task (3) Theory-Based Training (4) Theory-Based training Test (5) Set 2 Evaluations (6) Discrimination Task (7) Distracter Task (8) Set 3 Evaluations
Day 3	(1) Set 4 Evaluations

Materials and Procedures

As mentioned previously, the study took place over three consecutive days. Except for the training, the materials for this study were administered from a webpage. Participants were given a username and password to use throughout the study. Participants completed Day 1 at least 24 hours before Day 2's scheduled session. Furthermore, participants were denied access to the experimental webpage for 24 hours after they completed Day 2. Day Three needed to be completed within 24 hours after the 24 hour delay from Day 2. On Day 1, participants read the waiver of consent (Appendix B) and completed the biographical data form (Appendix C). All other materials are discussed in detail in the following sections.

The materials for this study were developed with the assistance of three experienced mechanics. Each mechanic had more than five years of experience working on vehicles including cars, trucks, and jeeps. Two of the experts were automotive technicians and one was a diesel technician. All three experts had received formal classroom instructions in the area of their expertise.

Operation Knowledge Test

Participants' domain knowledge was collected using an operation-specific knowledge test (Appendix D). On the first day of the study, participants answered 15 questions pertaining to the components and mechanics of steering a car. The questions were obtained from the Website www.howitworks.com and the Prentice Hall ASE Test Preparation Series, Steering and Suspension workbook (Halderman & Mitchell, 2004). The website provided questions that target novice-level to intermediate-level knowledge of automobile steering, while the workbook provided questions written by automobile industry experts and educators which targeted more of an expert level of automobile steering knowledge.

To rank the difficulty of the test questions, the expert mechanics were asked to sort the 15 questions into three groups including questions that novices should be able to answer, questions that intermediate and expert people should be able to answer, and questions that only an expert should be able to answer. All of the questions were in a multiple-choice format with three incorrect answers and one correct answer.

Knowledge Structures

The Team Performance Lab - Knowledge Assessment Testing Suite (Hoeft et al., 2003) which is based on the concept map knowledge elicitation method was used to create the KSs used for the evaluations. The KSs contained both structural (i.e., concepts, links) and contextual (i.e., labels) components.

A total of four KSs were developed which contained the same contextual information, however, the spatial location of the information varied. In order to accomplish this, an initial KS (A) was created; KS (A) was then flipped to the right to produce a mirror image for KS (B). KS (B) was then flipped upside down to produce KS (C). Finally, KS (D) was created by inverting KS (A) (Appendix E). The purpose of this technique was to create a repeated measure that would contain the same information, yet appear as if it was different. A pilot study was conducted to determine the probability of the participants recognizing that the KSs were exactly the same. The results indicated a low probability recognition rate; more specifically, when the pilot participants evaluated two KSs back-to-back, only one out of ten reported that the information contained within was the same. Furthermore, the participant who recognized the KSs as being the same had extensive knowledge of automotive mechanics, KSs, and KS evaluations.

KS Evaluation Administration

Once the KSs were developed, a program was created that allowed the participants to click on each label within the KS and assign a rating using the prompted rating scale. Ratings were chosen by clicking on a radio button. Once a label was rated, the rating appeared beside the proposition on the map. Once all the propositions on the screen were scored, participants were prompted to provide an overall rating (Appendix F shows the evaluation procedure in screen shots).

As seen in Table 1 above, one set of KSs were evaluated back to back before the training was administered, then two times after the training was administered, and then one time on Day 3. For the first set (pre-training), participants were instructed to evaluate each label within the KS and the overall KS based on the correctness of the information. For Sets 2 through 4, participants were asked to use the procedures they learned in the training to evaluate the KSs.

KS and Complexity

To determine whether the complexity of a KS would decrease the cognitive demands associated with exceeding the capacity of working memory, the number of concepts within each KS was manipulated. In particular, participants were randomly assigned to either 7-concept KSs with 9 links and labels or 10-concept KSs with 12 links and labels (see Appendix G). As mentioned previously, an ideal number for both eliciting and evaluating KSs is unknown. The average number of concepts used in past studies was somewhere between 10 and 20. The number of concept used in this study was determined by the automotive experts who decided on the minimum amount of concepts and links that could provide a basic KS of steering. The addition of three more concepts allowed for three more links and labels which maximized the difference between the 7-concept KS and the 10-concept KS while at the same time minimizing the time it took to evaluate them.

Referent Material

To investigate whether referent material would assist with reducing the cognitive demands associated with the long term memory retrieval process, a referent KS was developed and randomly assigned to participants. Those in the referent condition were given a referent KS to use for each KS evaluation. To develop the referent KS, KSs from the three expert mechanics were averaged to create one ideal KS (Goldsmith, Johnson, & Acton, 1991). The ideal KS was elicited from each expert using the concept mapping elicitation method which, as mentioned above, facilitates the elicitation of both structural and contextual dimensions of one's knowledge. Once the expert's individual KSs were created, the connections shared among them were used to create an ideal referent KS. The mechanics were then given the KS that depicted only the shared linked concepts. Each mechanic was asked to provide labels for the linked concepts. These labels were then examined for similarities, and a single label was created. The mechanics were given a KS that depicted the linked concepts and labels, and then asked to (as a group) determine whether there were any discrepancies within the KS. The final referent KS was complete when all the mechanics agreed that the KS depicted accurate and relevant information about automobile steering (Appendix H).

FOR Training

To investigate whether errors resulting from the halo effect could be mitigated using training, theory-based rater error training was designed and administered to each participant (Appendix I). The training was developed based on the FOR training paradigm (see Bernardin, Buckley, Tyler, & Weisse, 2002) discussed in Chapter Three. The training instructed the participants on the various dimensions within a KS (i.e., accuracy, relevancy, and density) and how to evaluate the KS. The domain used for the training was photosynthesis. The training described each dimension and explained how to assign a single score that best represents the accuracy of the dimensions. Power point slides were used to administer the training. Each participant had 15 minutes to review the training slides.

Training Effectiveness

A measure was developed and administered before and after training to determine whether theory-based training was effective at teaching participants how to discriminate between varying levels of quality among the dimensions within a KS (see Appendix J). For the task, participants viewed twelve pairs of KSs and determined which of the pair was of better quality or if the pair was of the same quality. Four different KSs were developed using familiar driving concepts and propositions. Familiar domain information was used to allow participants to focus on comparing the dimensions rather than focusing on the accuracy and relevancy of each proposition; therefore, reducing confounds such as experience or knowledge of the domain.

This task was administered using a power point slide format. Once the task started, a pair of KSs appeared, participants had 1.5 minutes to record, on paper, which KS was of a higher quality, or if they were of the same quality. Once their response was recorded, the participants then clicked to the next pair of KSs. If a choice was not made within 1.5 minutes, the participants were prompted to make a decision.

Each KS represented high, low, or both levels of density and accuracy; in particular, the maps contained high density and high accuracy, high density and low accuracy, low density and high accuracy, and low density and low accuracy. The KSs were paired so that six were used as distracters/manipulation checks which were spread throughout the six experimental KSs. Every participant viewed the twelve pairs of KSs in the same order. The order for the pre-training administration was different from the post-training order. Table 2 explains what discriminations were being made by participants when they correctly identified which KS represented a higher level of quality.

Table 2

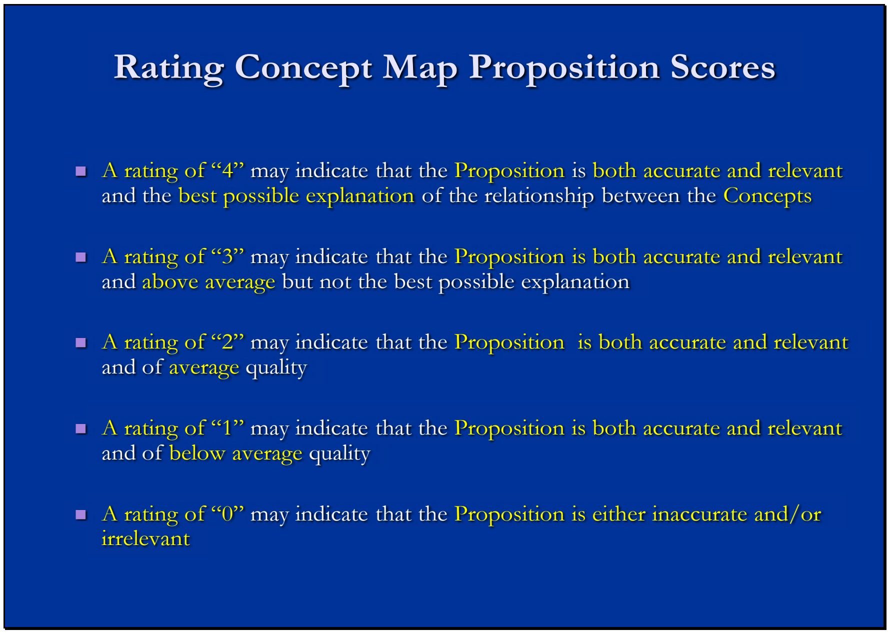
Discrimination Task

Pair	Density and Correctness	Answer	Explain
1 and 2	HD/HC vs. HD/LC	1	demonstrates discrimination between correctness with same density; high correctness is better than low correctness when same density
1 and 3	HD/HC vs. LD/HC	1	demonstrates discrimination between density with same correctness; high density is better than low density when correctness is equal
2 and 3	HD/LC vs. LD/HC	3	demonstrates discrimination between density and accuracy; low density is better than high density when all is correct in LD and non are correct in HD
2 and 4	HD/LC vs. LD/LC	same	demonstrates that density is irrelevant when nothing is correct
1 and 1	HD/HC vs. HD/HC	same	manipulation check
3 and 3	LD/HC vs. LD/HC	same	manipulation check

Dependent Variables

Rating Scale

The rating scale used to evaluate the KSs range from “0” to “4,” where “0” indicated poor quality and “4” indicated excellent quality. This scale is representative of protocols used to score KSs. For example, the McClure et al (1999) protocol depicted in Figure 5 above follows a “0” to “4” point scoring system for assessing both the contextual and structural information within a KS. Furthermore, Ruiz-Primo et al (1997) developed a proposition inventory which provided raters with examples of varying qualities of propositions. The qualities included (a) Excellent: outstanding and correct, shows a deep understanding of the relationship between two concepts; (b) Good: complete and correct; shows a good understanding of the relationship between the two concepts; (c) Poor: incomplete but correct; shows partial understanding of the relationship; (d) Don't Care: a valid relationship but doesn't show understanding; and (e) Inaccurate: incorrect proposition. Following the same schema, a similar scoring protocol was used for this study; however, each point on the scale represented the correctness of the important dimensions within a KS, particularly the accuracy and relevancy of the propositions (see Figure 18). After training, all participants had a copy of the evaluation protocol to refer to when evaluating the remaining KSs. Prior to training, however, participants were only given the end points of the scale labeled "unacceptable" and "exceptional."



Rating Concept Map Proposition Scores

- A rating of "4" may indicate that the Proposition is both accurate and relevant and the best possible explanation of the relationship between the Concepts
- A rating of "3" may indicate that the Proposition is both accurate and relevant and above average but not the best possible explanation
- A rating of "2" may indicate that the Proposition is both accurate and relevant and of average quality
- A rating of "1" may indicate that the Proposition is both accurate and relevant and of below average quality
- A rating of "0" may indicate that the Proposition is either inaccurate and/or irrelevant

Figure 18. Slide from training that describes what each point on the rating scale represents.

Reliability Scores

Reliability scores represented the percentage of times a rater gave the same rating for the same propositions within each Set of KSs (Set 1, 2, 3, and 4). More specifically, there were a total of eight KS evaluations including one set pre-training and three sets post training, the average amount of matches between each proposition within each of the four sets of evaluations was calculated (see Equation 1). For example, if a participant gave the same ratings for 7 out of twelve propositions within Set 2, then their reliability score would have been .58; meaning that 58% of the time they gave the same ratings to the same propositions within Set 2.

Each participant had one reliability score for each of the four sets of evaluations.

$$M\ (\text{reliability}) = \sum [(F_1 \ldots c + S_1 \ldots c)] / c \qquad (1)$$

Where,
M = the KS Set (1, 2, 3, or 4);
F = the rating for the proposition within the first KS; (9 or 12)
S = the rating for the proposition within the second KS
c = the total number of propositions within the KS (9 or 12)

Validity Scores

Validity scores reflected how accurate participants' ratings were as compared to true ratings, or ratings obtained from the expert mechanics. To calculate the convergence between the expert and the participant's ratings (Validity), the same formula for calculating the reliability score was used, however, each rating within a KS was compared to the true rating (see Equation 2); therefore, the score represents the percentage of times the participants' ratings matched the expert ratings. Each participant had a total of eight Validity scores.

$$(K)\ \text{Validity} = \sum [(F_1 \ldots c - E_1 \ldots c)] / c \qquad (2)$$

Where,
K = the Knowledge Structure (A, B, C, D, A2, B2, C2, D2);
F = the rating for the proposition within the KS (1 thru 9 or 12)
E = the expert rating for the proposition; and,
c = the total number of propositions within the KS (9 or 12)

CHAPTER FIVE: RESULTS

Overview

All analyses were conducted using SPSS 15 for Windows. Unless otherwise noted, an alpha level of .05 was used. Below, data cleaning efforts and manipulation checks are described followed by the results from the hypotheses testing. To reiterate, both reliability and validity were used to evaluate the psychometric properties of the ratings derived from a KS evaluation method. Analyses and the results from hypotheses testing are presented in the following order: first, the analyses and results related to FOR training and its ability to mitigate halo, second, the analyses and results related to the impact that domain knowledge had on mitigating retrieval errors, and finally, the analyses and results related to KS complexity and the role of referent material in mitigating cognitive demands on working memory. Additional analyses included analyzing the effects that different levels of the independent variables had on the dependent variables.

Data Cleaning

One-hundred and three participants completed the entire study. SPSS EXPLORE was used for evaluating the normality of the data. The data was first inspected for accuracy by looking for out of range variables, outliers, plausible means, and plausible standard deviations. Inspection of the data led to the deletion of nine cases due to participants assigning the same rating for all propositions in more than two Sets of evaluations (i.e., assigning a 4 to all the propositions in Set 1 and Set 2). Of the remaining 93 cases, there was one with an extreme reliability score (i.e., fell more than 3.5 standard deviations from the mean). The participant only had one outlying score in their data set; therefore, instead of deleting the case, the participant's outlying score was adjusted to be one unit smaller than the next most extreme case in the distribution. This allowed the participant's data to be included without having an extreme influence on the distribution. Table 3 through 6 presents the descriptive statics and correlation matrices of the reliability scores for Set 1, 2, 3, 4, and Validity scores for the initial KS evaluation (KS A), the KS evaluation immediately following training (KS B), and the first KS evaluation on the following day (KS C).

Table 3

Centrality Statistics for Reliability Scores

Reliability Scores	N	Minimum	Maximum	Mean	Standard DV
Set 1	93	0.111	1.000	0.524	0.189
Set 2	93	0.000	1.000	0.664	0.204
Set 3	93	0.222	1.000	0.746	0.180
Set 4	93	0.000	1.000	0.701	0.209

Table 4

Centrality Statistics for Validity Scores

Expert Convergence Scores	N	Minimum	Maximum	Mean	Standard DV
KS A	93	0.000	0.667	0.331	0.135
KS B	93	0.000	0.667	0.303	0.130
KS C	93	0.000	0.667	0.307	0.133

Table 5

Correlations Matrix for Reliability Scores

reliability	Set 1	Set 2	Set 3	Set 4
Set 1	1.000			
Set 2	0.168	1.000		
Set 3	0.379**	0.370**	1.000	
Set 4	0.107	0.306**	0.283**	1.000

** Significant at $p < 0.01$ (2-tailed)

Table 6

Correlation Matrix for Validity Scores

Validity	KS A	KS B	KS C
KS A	1.000		
KS B	0.198	1.000	
KS C	0.020	0.274**	1.000

** Significant at $p < .01$ (2-tailed)

Multivariate normality of the reliability scores and Validity scores among the four sets of evaluations was examined using the QQ plot function in SPSS which plots observed values against a normal distribution. As seen in Figure 19 and 20 below, the reliability scores for all Sets of KSs and the Validity scores for KS A, KS B, and KS C were closely distributed around the normal line.

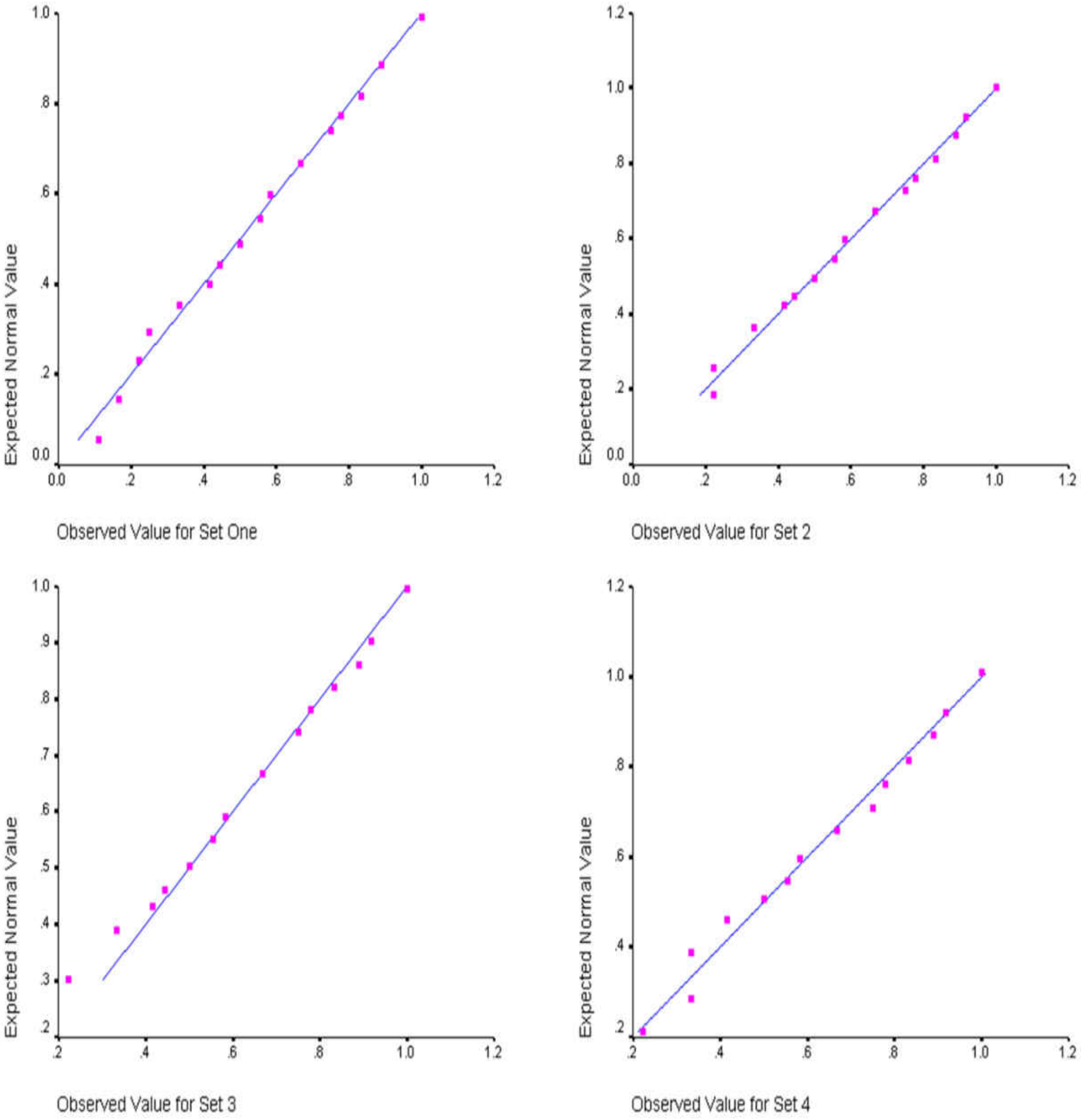


Figure 19. QQ plots showing normal distribution for the overall reliability scores.

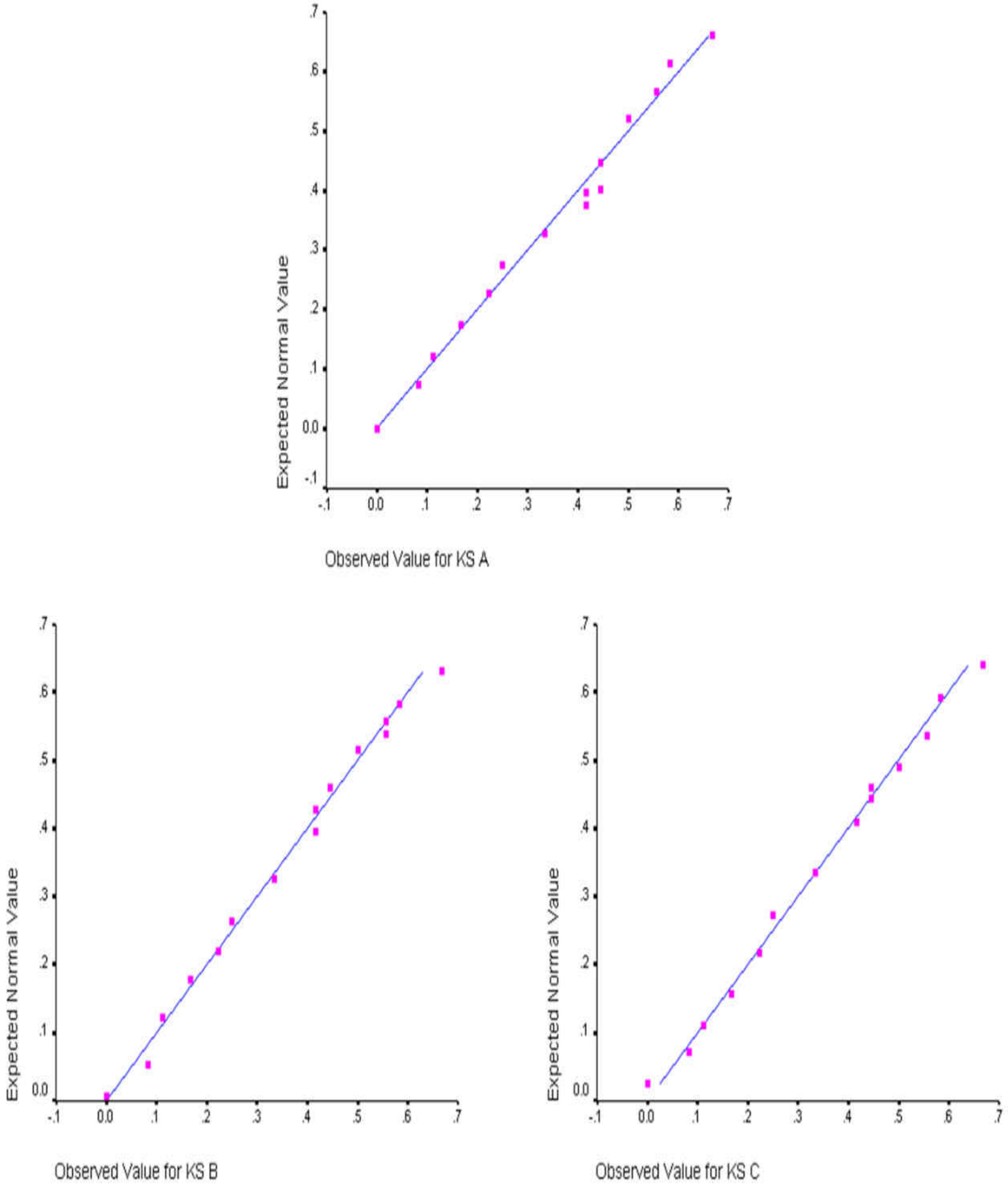


Figure 20. QQ plots showing normal distribution for the overall Validity scores.

Furthermore, QQ plots were used to assess the normality of the reliability scores and Validity scores for each of the 4 sets of evaluations within each group, where group 1 was the

10-concept / no referent group (see Figure 20 and 21), group 2 was the 7-concept / no referent group (see Figure 21 and 22), group 3 was the 10-concept / referent group (see Figure 23 and 24), and group 4 was the 7-concept / referent group (see Figure 25 and 26). All distributions were closely distributed along the expected value line with minimal to no deviation. It was concluded that the multivariate assumption of normality for the reliability and validity scores within each group was met.

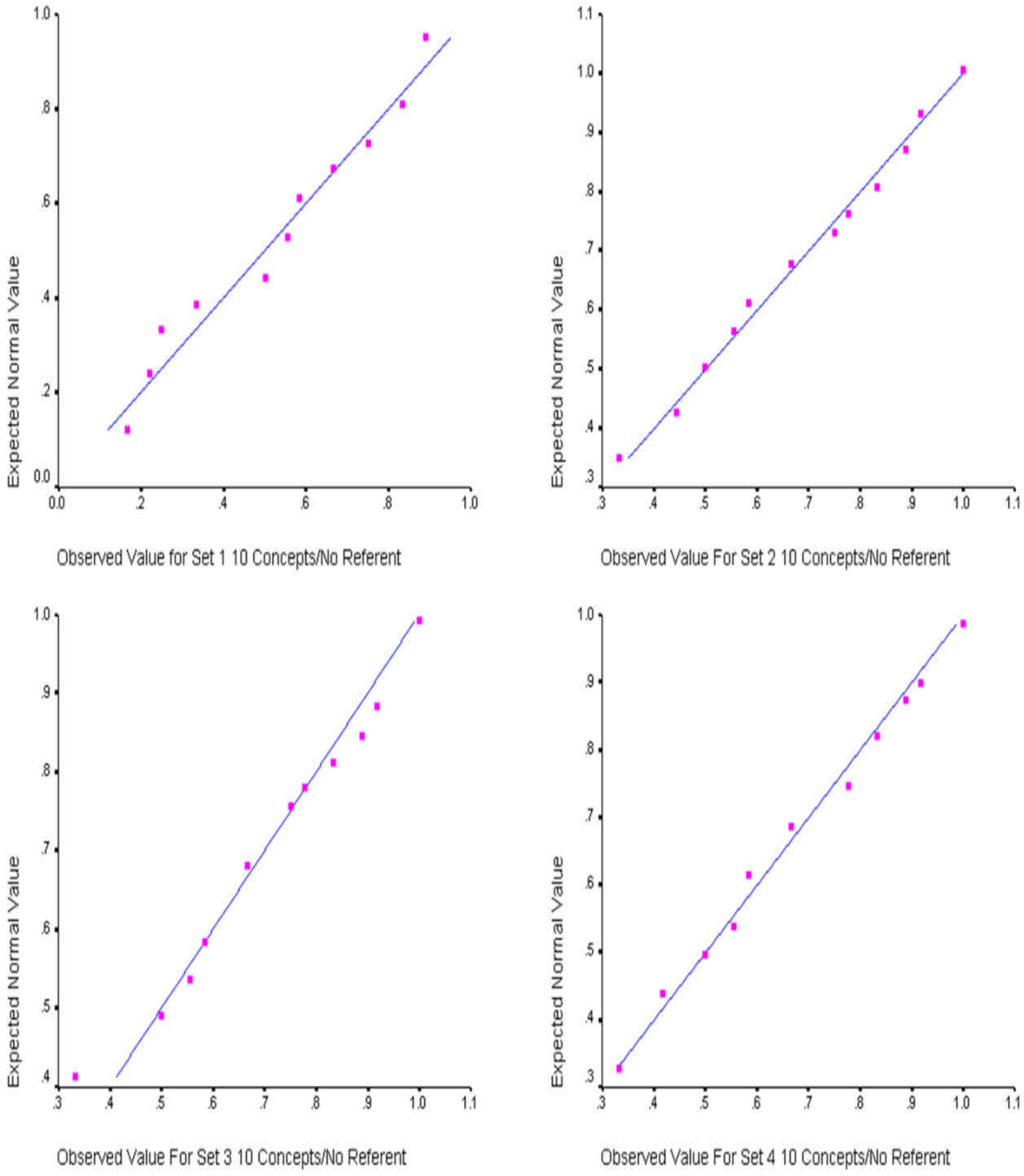


Figure 21. QQ plots showing normal distribution for reliability scores in the 10-concept / no referent group.

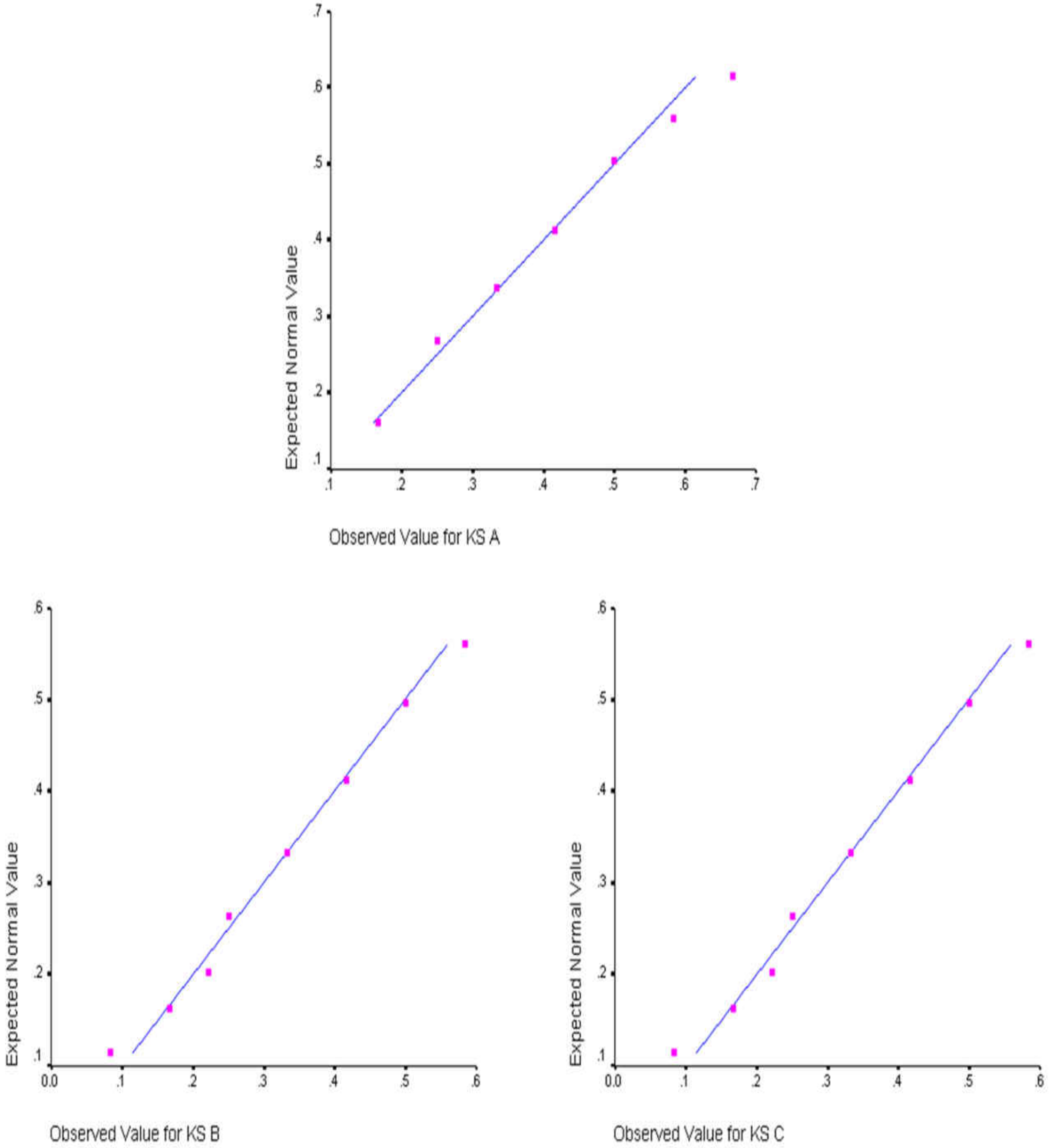


Figure 22. QQ plots showing normal distribution for Validity scores in the 10-concept/no referent group.

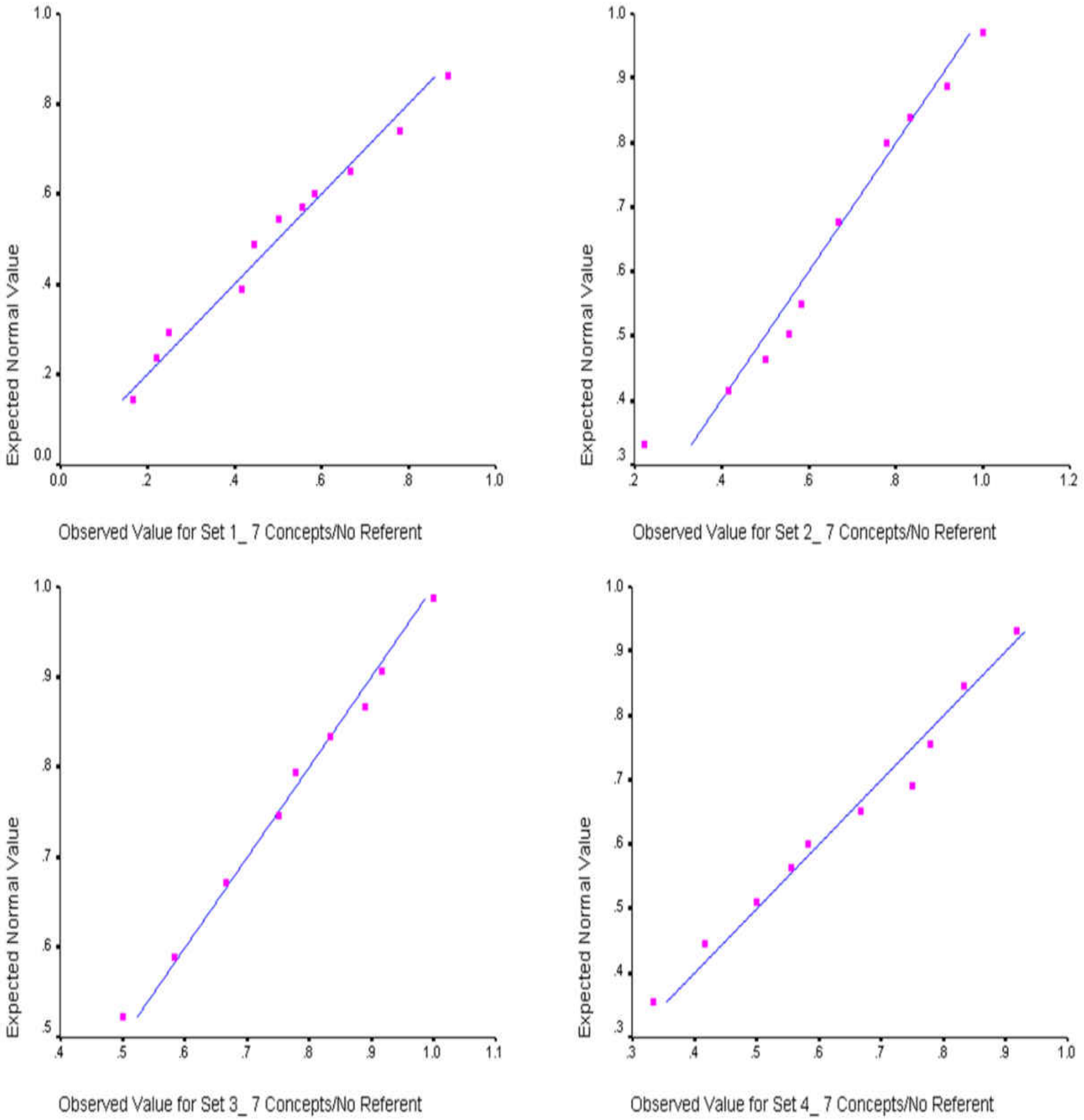


Figure 23. QQ plots showing normal distribution for reliability scores in the 7-concept / no referent group.

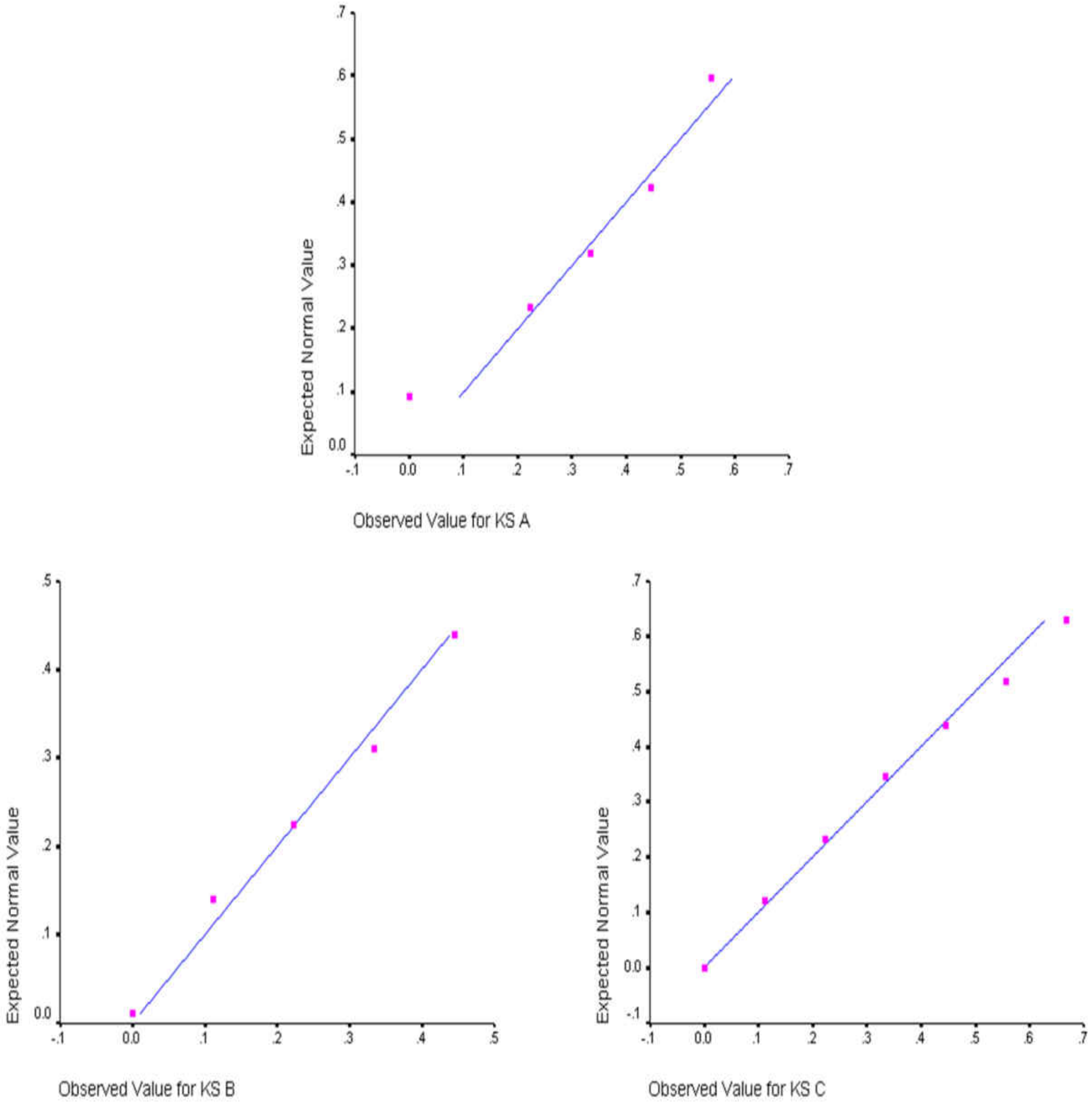


Figure 24. QQ plots showing normal distribution for Validity scores in the 7-concept/ no referent group.

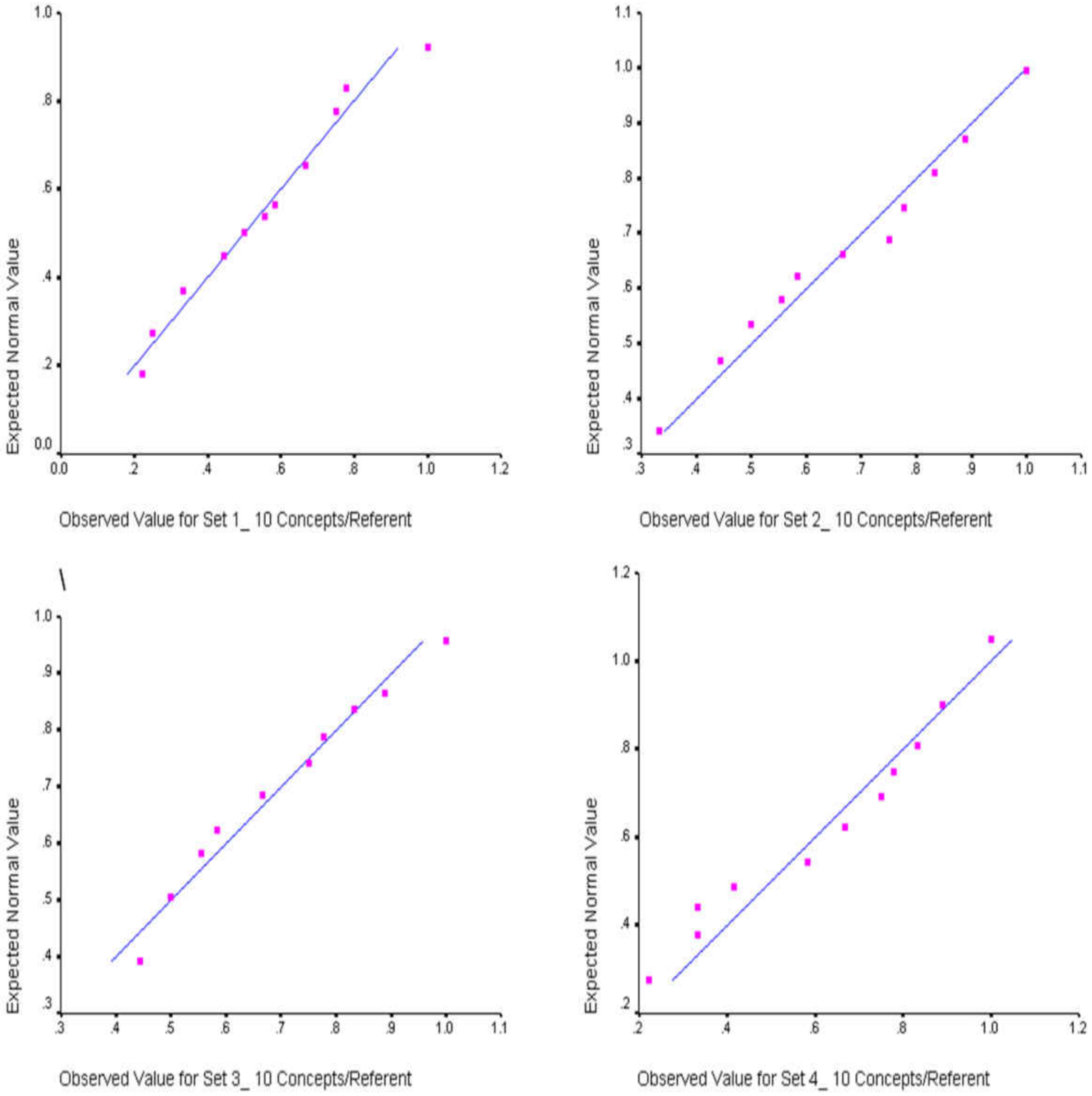


Figure 25. QQ plots showing normal distribution of reliability scores in the 10-concept / referent group.

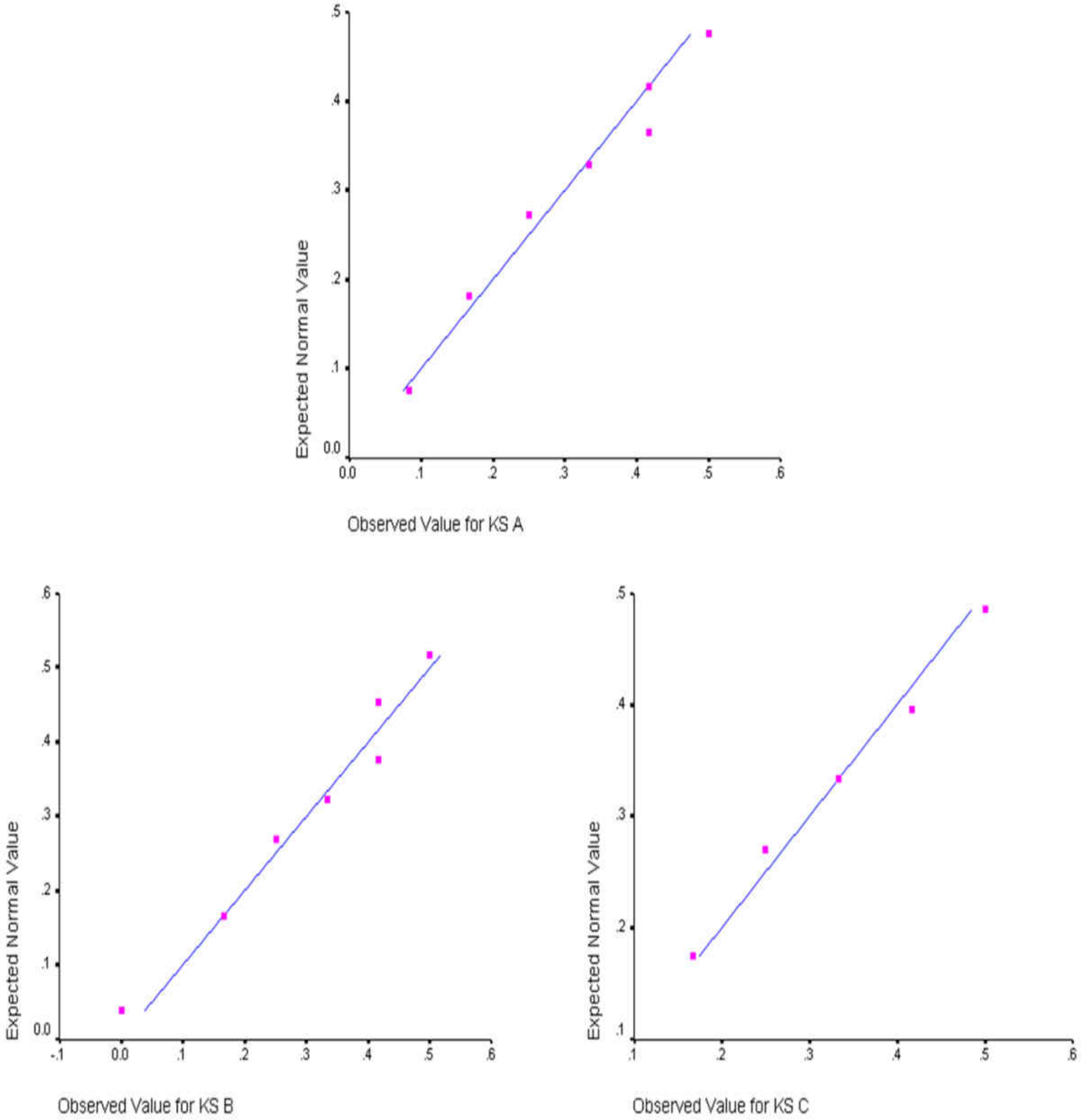


Figure 26. QQ plots showing normal distribution for Validity scores in the 10-concept / referent group.

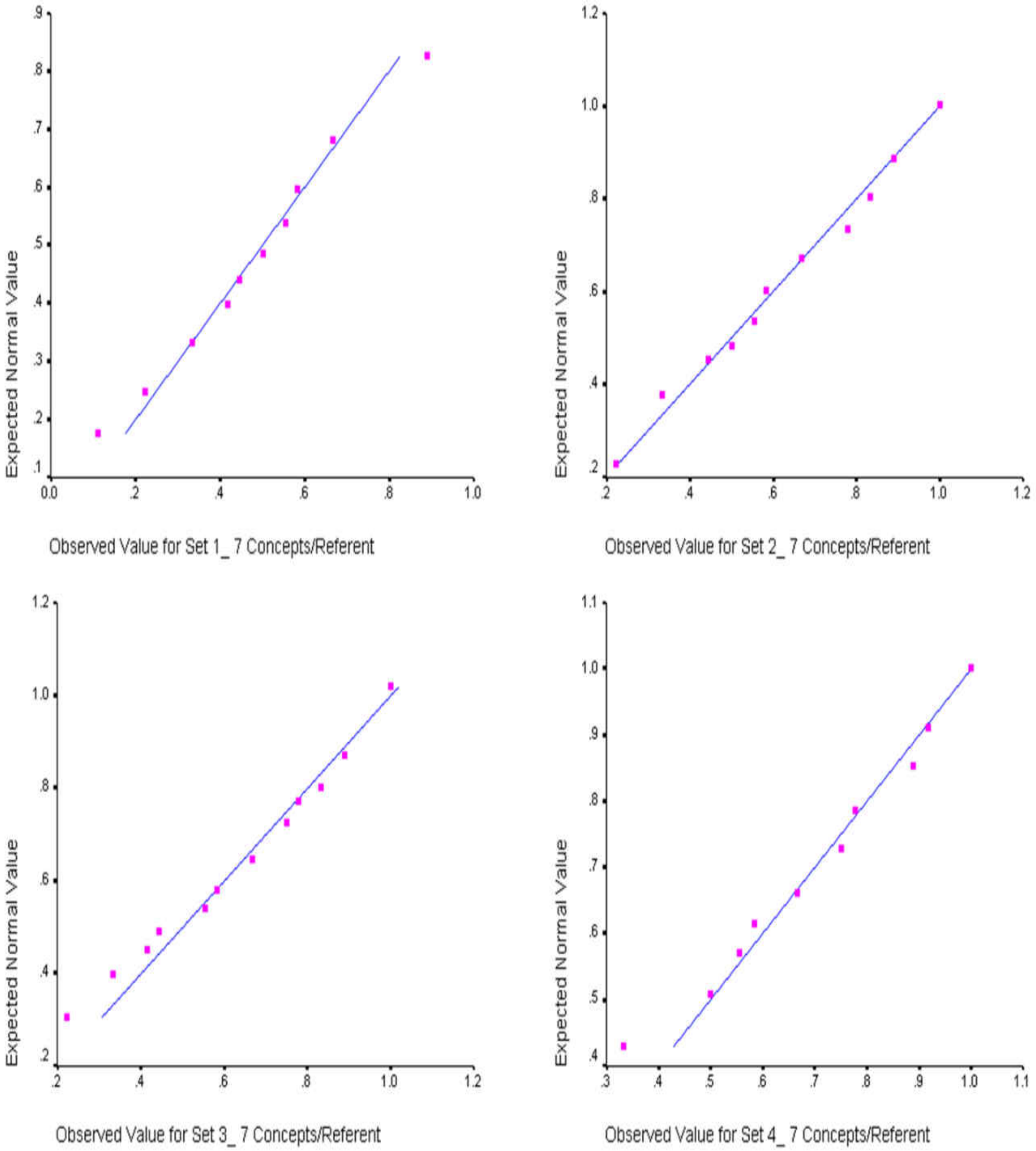


Figure 27. QQ plots showing normal distribution for reliability scores in the 7-concept / referent group

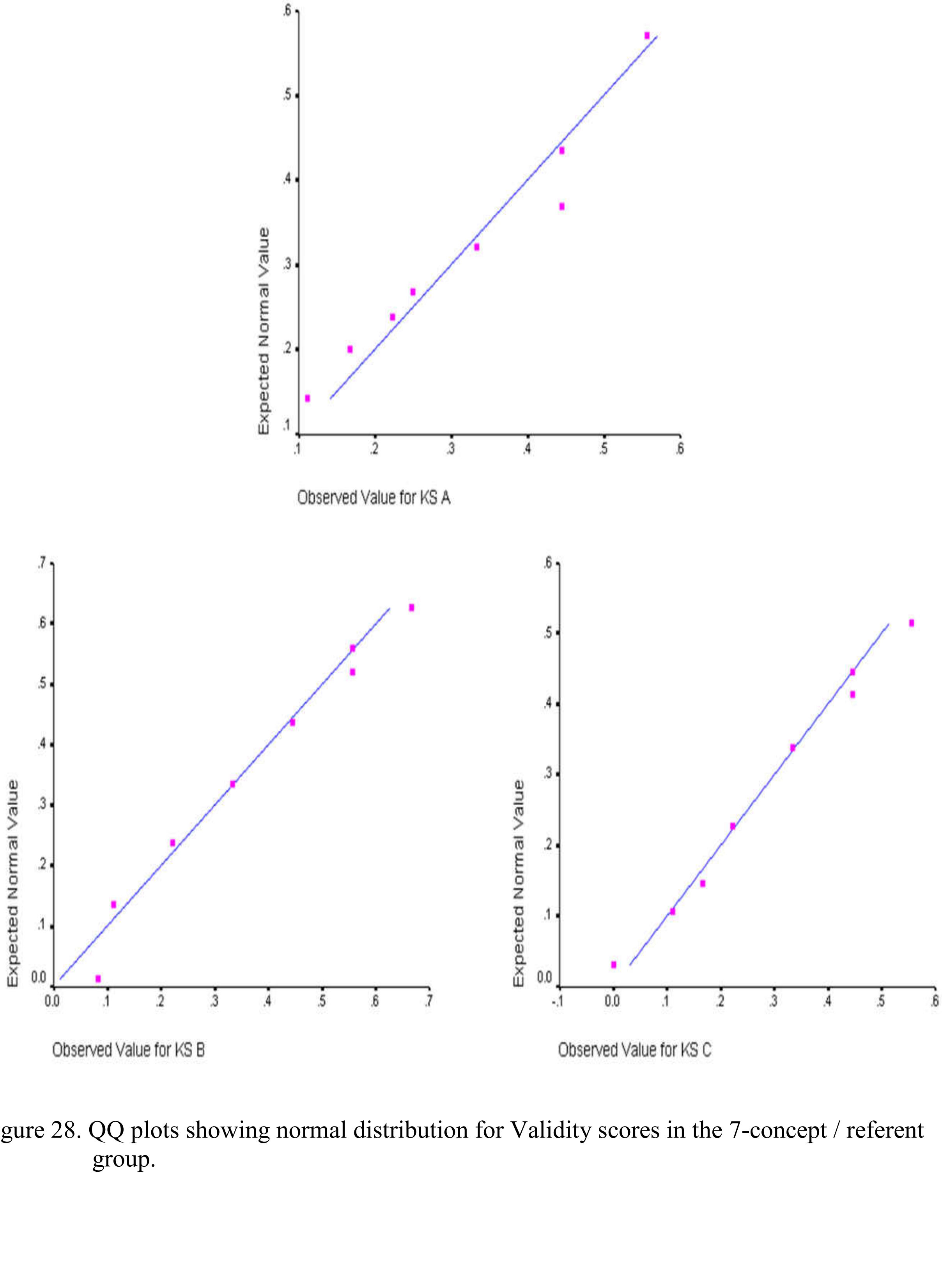


Figure 28. QQ plots showing normal distribution for Validity scores in the 7-concept / referent group.

Manipulation Check

To test whether the frame-of-reference (FOR) training was effective at teaching what dimensions with a knowledge structure (KS) should be evaluated, the participants' ability to identify KSs that correctly represented both structural and contextual characteristics was assessed. This was done by using the total of correct responses for the pre-, and then post-administration of the discrimination task described above. A paired samples t-test was used to determine if the participants' overall scores on the discrimination task were higher after receiving rater training. The test showed that participants had significantly higher scores post training (M = 9.01, SD = 1.691) than pre-training (M = 7.81, SD = 1.548); t (91) = 6.87, $p <$ .0001, η^2 = .34.

Hypothesis Testing

Analysis of Hypothesis 1 and 2

FOR training was identified as a method for reducing the halo effect within KS evaluations; specifically, by providing raters with a deeply encoded conceptualization of the evaluation process that could be consistently applied across multiple evaluations. Therefore, it was hypothesized that FOR training would increase the reliability within a participant's ratings. A multivariate repeated-measures analysis of variance with reliability scores for the pre-training set (Set 1) and the three post-training sets (Set 2, 3, and 4,) as the repeated measures was used to test this hypothesis. Under the assumption of Sphericity, a main effect of FOR training was found, F (3, 276) = 30.233, p < .0001; $\eta_{\mathrm{p}}{}^2$ = .247. Pairwise comparison using Least Squares Differences (LSD) tests indicated that Set 1 was significantly different from Sets 2, 3, and 4. Furthermore, Set 2 was significantly different from Set 3 (see Figure 29).

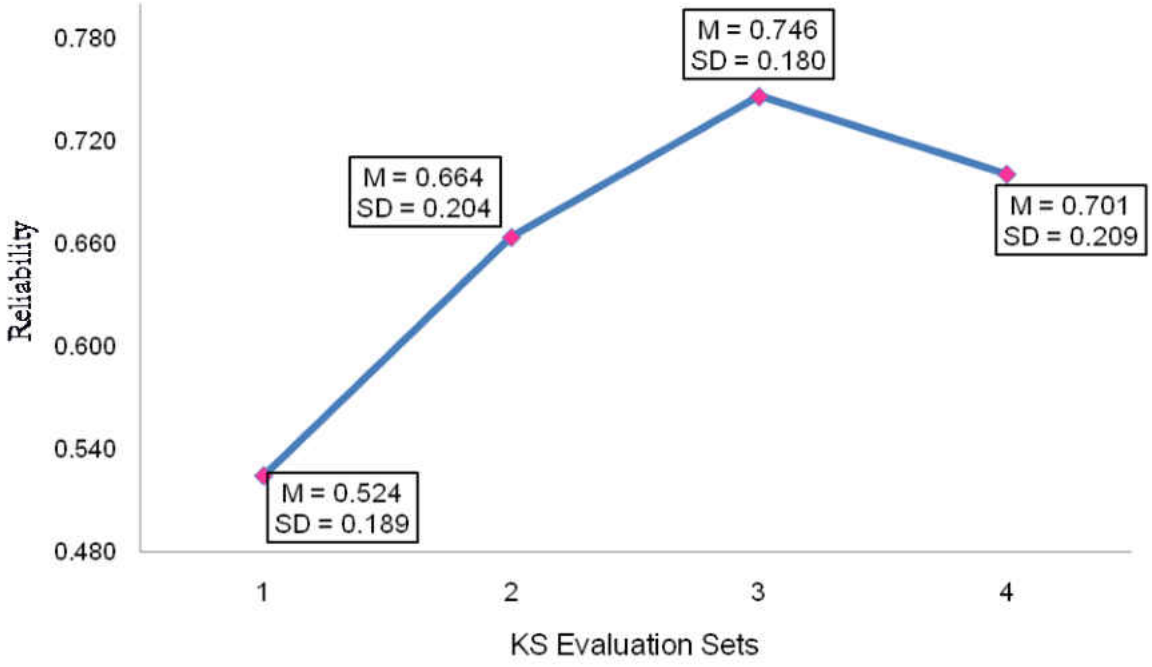


Figure 29. Means from the reliability scores across four Sets of KSs showing the upward trend on Day 1, then the leveling on Day 2.

For subsequent reliability analyses, Set 1 data was tested as a covariate and used where applicable. Furthermore, Set 4 was used for the remaining analyses. Set 4 was chosen because it was assumed that Set 4 had been the least influenced by practice effects. Practice effects for this study would have been related to memorizing the ratings. Participants evaluated the same propositions within each KS; therefore, any significant effects may have resulted from remembering the ratings they had previously assigned to each proposition, rather than as a result of the manipulated variables. The higher reliability scores for Set 3 show that there was most likely a *memorization effect* between Set 2 and 3. Without this effect, Set 2 and Set 3 would not have been significantly different. It appears, however, that the memorization effect was eliminated for Set 4 when the time period between the evaluations was the greatest. In fact, Set 4 was not significantly different from either Set 2 or Set 3 which suggests that any benefits gained from the memorization effect diminished, resulting in the participants regressing to their mean reliability scores.

Analysis of Hypothesis 3

Reducing the complexity of a KS was proposed to reduce the demand on a rater's storage process, thereby facilitating the development of a mental model that is consistently used across multiple evaluations. Therefore, hypothesis 3 stated that reducing the complexity of a KS would lead to more consistent ratings. First, the assumption of equality of slopes, which was required for using Set 1 as a covariate, was tested. The interaction between the variables was not significant, so Set 1 was used as a covariate. A one-way between-subjects analysis of covariance with complexity (10-concepts vs. 7-concepts) as the IV and reliability scores for Set 4 as the DV was conducted. Although the 7-concept group had a higher percentage of matches than the 10-concept group, the ratings were not significantly more reliable when the complexity of the KS was reduced, F (1, 90) = 0.839, p = .362 (see Figure 30)

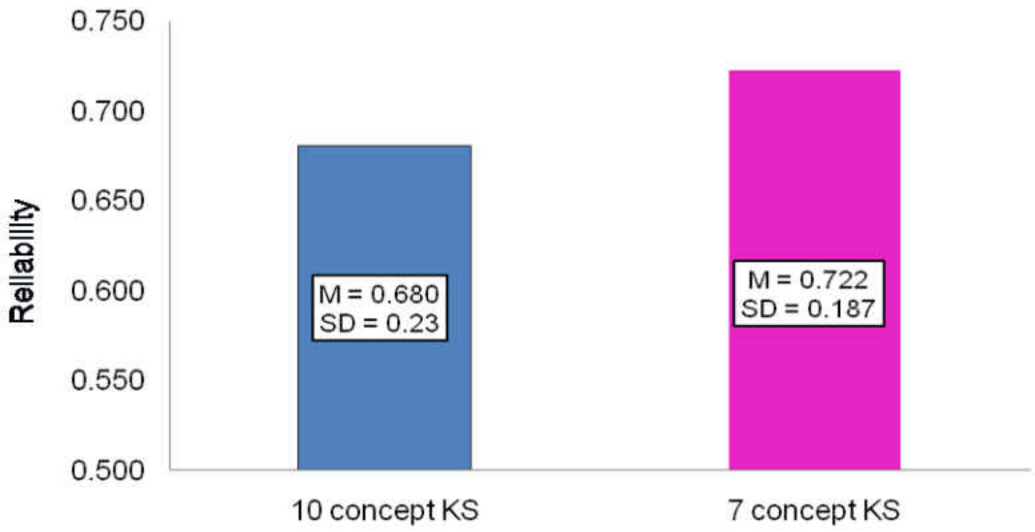


Figure 30. Mean reliability scores for the 10 concept KS group (left) and the 7 concept KS group (right).

Analysis of Hypothesis 4

Referent material was proposed to facilitate the retrieval of information stored in long term memory, thereby allowing the rater to recall and use the same information to form a mental model that is consistently applied across multiple evaluations. Hypothesis 4, therefore, stated that participants who used referent material while evaluating KSs would have more consistent ratings than those who did not use referent material. The equality of slopes assumption was met; therefore, Set 1 was used as a covariate. A one-way between-subjects analysis of covariance with Set 1 as the covariate, referent conditions (no referent vs. referent) as the IV, and the reliability scores for Set 4 as the DV was conducted. Although the referent group had more matches within their ratings, the referent model alone did not significantly increase the reliability of the ratings F (1, 90) = 2.948, p = .089 (see Figure 31).

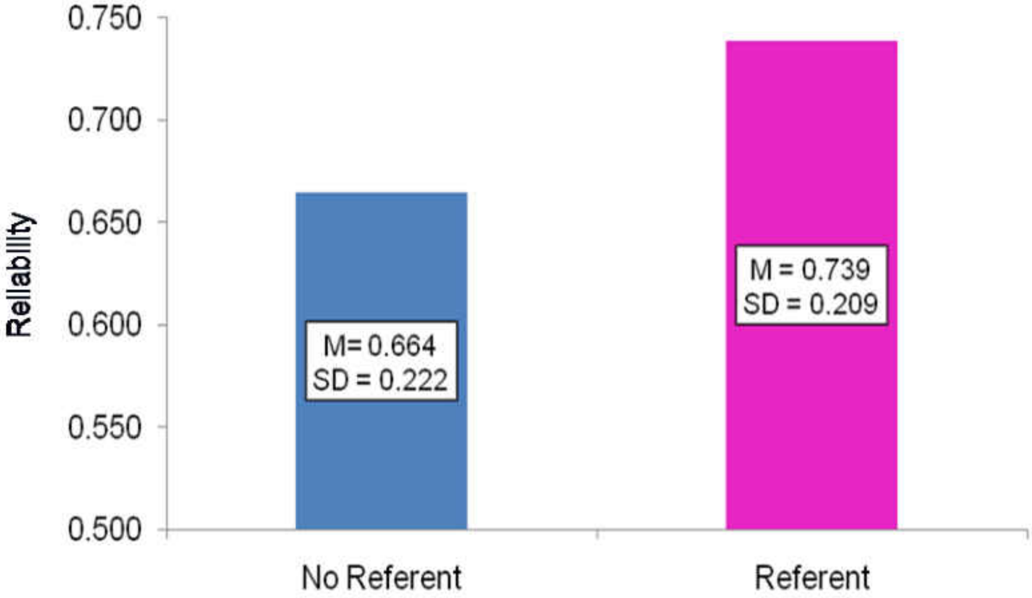


Figure 31. Mean reliability scores for the No Referent (left) and Referent (right) groups.

Analysis of Hypothesis 5

Raters with domain knowledge were proposed to be able to active more information about the domain, thus form a consistent mental model across multiple evaluations. Therefore, Hypothesis 5 stated that participants who had more knowledge of steering would have more consistent ratings than participants with limited to no knowledge of steering. A median split was used to group participants into high- and low-domain knowledge categories. As mentioned previously, the domain knowledge test consisted of 15 questions. The number of correct answers was totaled for each participant. Participants correctly answered anywhere between 2 to 14 questions. The mean score was 7.95 with a standard deviation of 2.138. The distribution of this data allowed for a nice median split. Using 8-correct as the criterion, the result was 52 participant in the low-domain knowledge group and 41 participants in the high-domain knowledge group.

The interaction between Set 1 and domain knowledge was not significant; therefore, Set 1 was used as a covariate. A one-way between-subjects analysis of covariance with Set 1 as the covariate, domain knowledge level (high vs. low) as the IV and Set 4 reliability as the DV was conducted. The result indicated that more domain knowledge did not increase the reliability of the ratings, *F* (1, 90) = 0.002, *p* = .963 (See Figure 32).

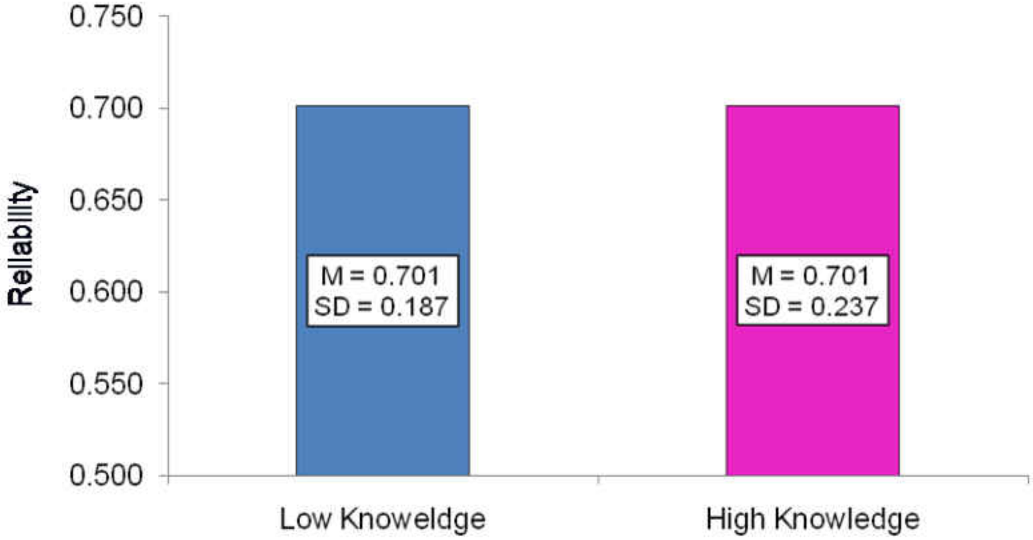


Figure 32. Mean reliability scores for the low domain knowledge group (left) and the high domain knowledge group (right).

Analysis of Hypothesis 6

Hypothesis 6 stated that domain knowledge would increase the accuracy of a rater's ratings. A one-way between-subjects analysis of variance with domain knowledge level (high vs. low) as the IV and the accuracy scores for the initial KS evaluation as the DV was conducted.

The result indicated that more domain knowledge did not increase the accuracy ratings, *F* (1, 90) = 0.002, *p* = .963 (See Figure 33).

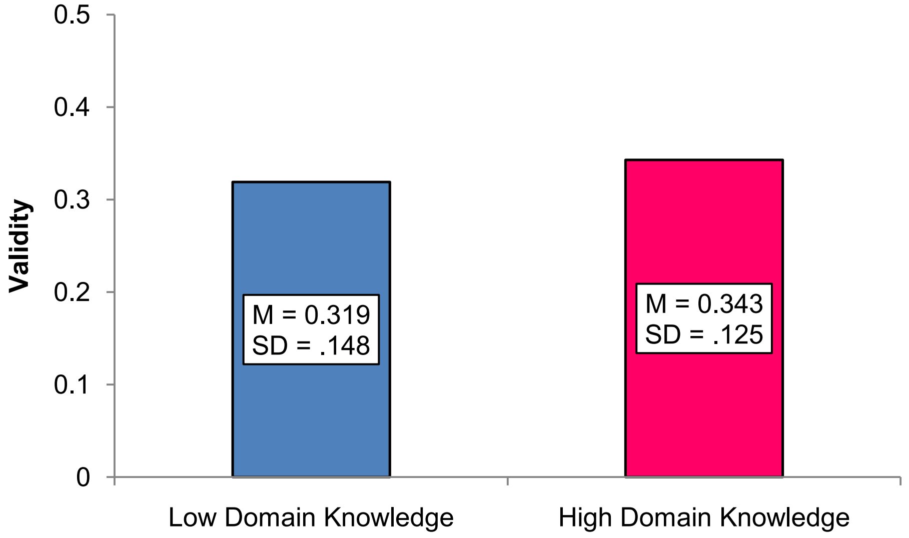


Figure 33. Mean reliability scores for the low domain knowledge group (left) and the high domain knowledge group (right).

Supplemental Analyses

Supplemental analyses were conducted to explore how the participants' reliability and validity scores were influenced by different levels of the manipulated variables including, low complexity/high complexity, referent/no referent, and less domain knowledge/more domain knowledge.

A repeated measures ANCOVA was used to conduct these analyses. The repeated measures included the reliability scores or validity scores for Set 1, Set 2, and Set 4 (Set 3 was left out because of the potential memorization effect discussed above). Furthermore, the complexity level (7-concepts vs. 10-concepts) and the referent condition (referent vs. no referent) were the IVs, and the scores from the steering knowledge test served as the CV. There were no significant interactions for validity; however, there was a significant interaction between the complexity and referent conditions for the reliability scores, *F* (1, 88) = 16.92, $p < .0001$, η^2 = .192 (see Figure 34). There was no significant crossover effect, however, there was a significant simple effect between the referent vs. no referent groups at low complexity, *F* (1, 88) = 4.503, $p = .0462$, indicating that the referent material significantly increased reliability when participants had less complex KSs to rate.

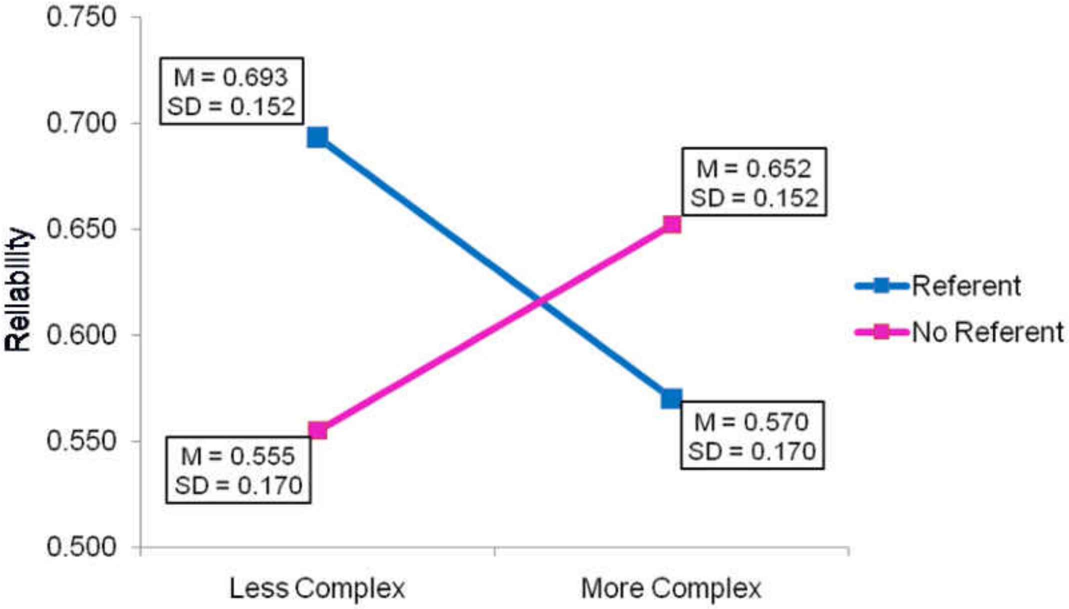


Figure 34. Depicts the simple effect indicating that those who scored less complex KSs had more reliable KS ratings when they were provided referent material.

Given this result, I looked for more instances where the mitigation methods or factor was effective within different groups of participants. Using a between-subject ANOVAs, I found that raters within the low domain knowledge condition had significantly more reliable rating when scoring less complex KSs (*M* = 0.636, *SD* = 0.215), than when scoring more complex KSs (*M* = 0.742, *SD* = 0.155), (*F* = 4.096, *p* = .048) (see Figure 35).

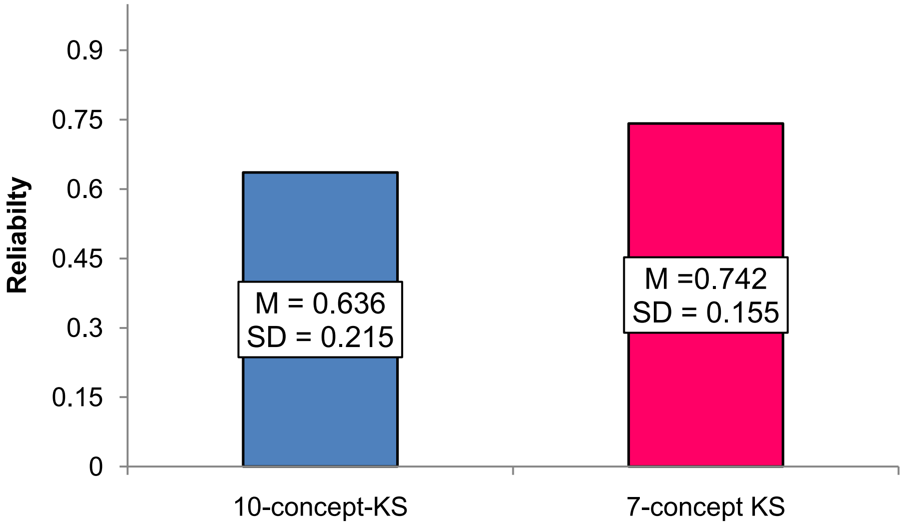


Figure 35. Depicts the significantly higher reliability scores for those who scored less complex KSs (right) as opposed to those who scored more complex KSs (left), within the low domain knowledge group.

Furthermore, participants in the high domain knowledge group had significantly more reliable ratings when assigned to the referent material group (M = X, SK + X), than when assigned to the control (M = X, SD = X) (see Figure 36).

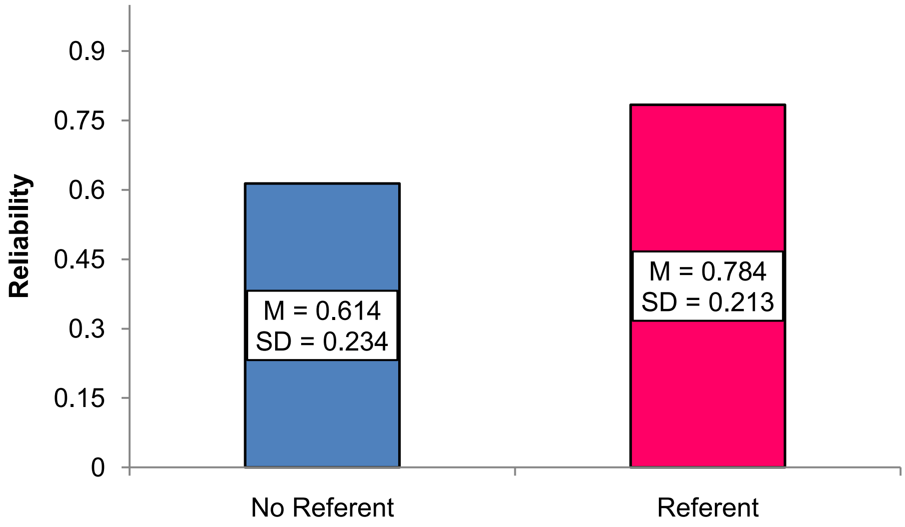


Figure 36. Depicts the significantly higher reliability scores for those who received referent material (right), as opposed to those who did not receive referent material (left), within in the more domain knowledge group.

CHAPTER SIX: DISCUSSION AND CONCLUSIONS

Discussion of Results

The primary purpose of this study was to identify methods that may mitigate the negative influence of a rater's limitations on the reliability and validity of knowledge structure (KS) evaluations. The methods and factors studied here included providing frame-of-reference (FOR) training, reducing the complexity of a KS, providing referent material, and having domain knowledge. Figure 37 represents the KS process framework used to guide this study (also seen in Chapter Three, Figure 9). The results of this study lead to an iteration of the KS process model depicted in Figure 37 (see Figure 38). The updated model depicts which mitigation techniques work together to affect the relationship between a rater's limitation and its respective cognitive process. This model is described in more detail in the following sections.

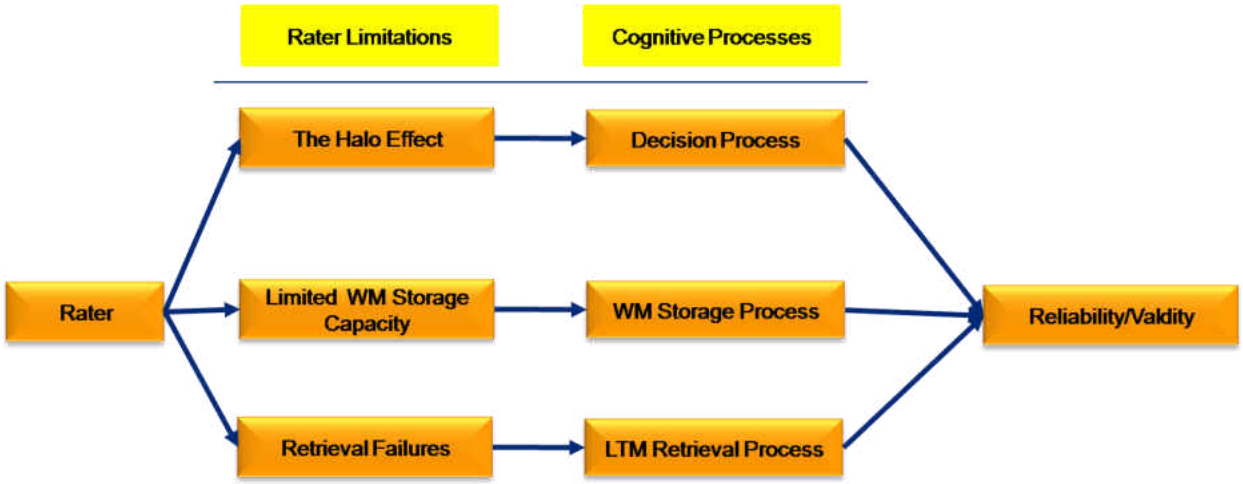


Figure 37. Depicts the original KS process framework that guided this study.

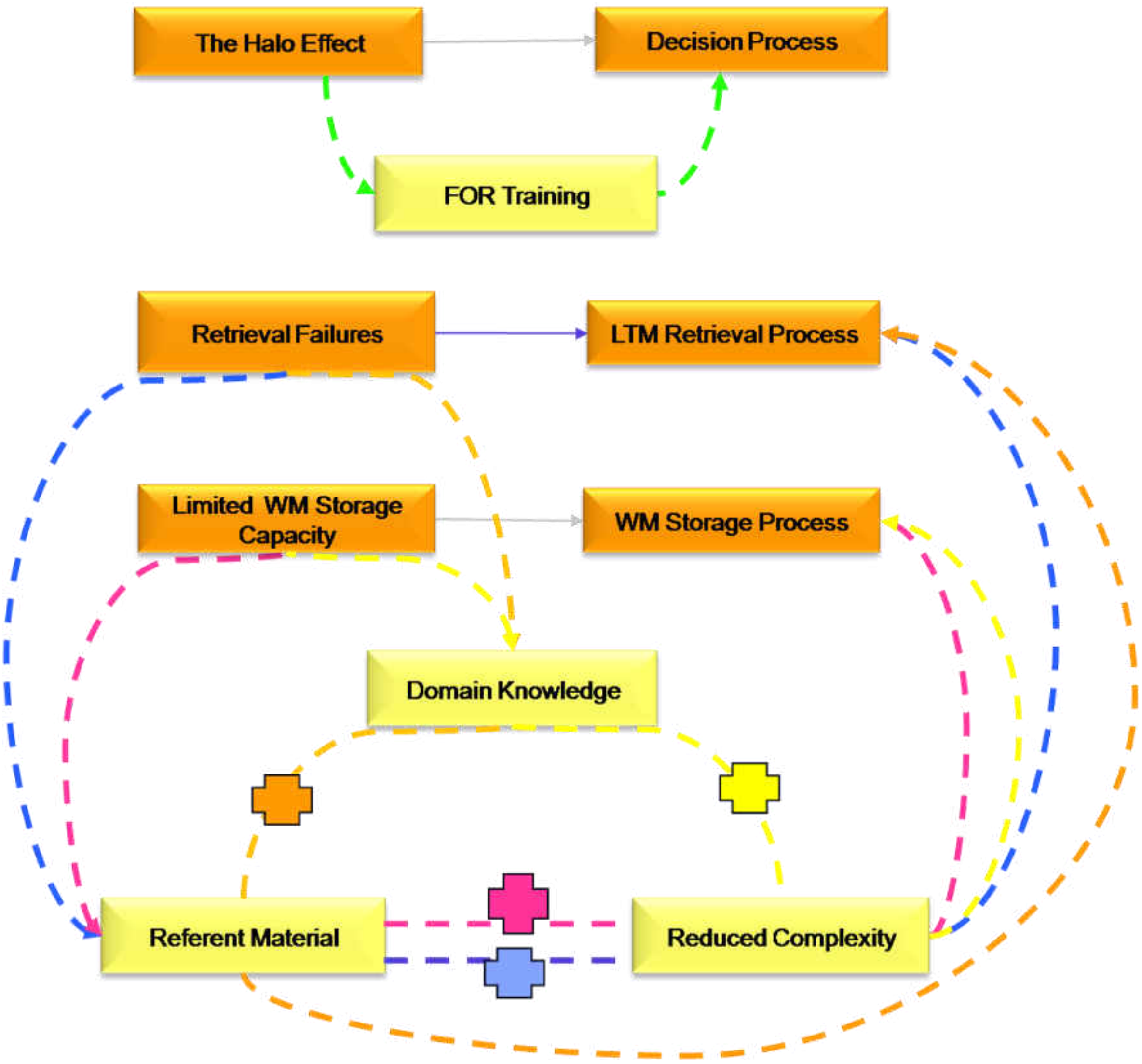


Figure 38. Depicts the revised KS process framework derived from the results of the study.

The Effectiveness of Frame-of-Reference Training

To reiterate, the halo effect in conceptual evaluations manifests itself as a tendency of a rater to derive ratings using only a subset of dimensions, rather than providing a rating that represents the quality of all relevant dimensions (Eckes, 2008; G. Engelhard, 1994). Prior to this investigation, minimal to no research, to my knowledge, had investigated how the halo effect influenced KS evaluations. From past research on job performance evaluations, it was assumed that KS ratings are dependent on whether the halo effect influenced the decision process that occurs while evaluating KSs.

FOR training was identified as a method for mitigating the negative effects that halo may have on KS ratings. It was chosen based on its success at reducing halo within job performance evaluations, and furthermore, its focus on teaching raters how to identify relevant dimensions within a KS, and assign a score that represent the quality of those ratings. As hypothesized, FOR training was able to increase the reliability of the KS ratings, as indicated by the approximate 25% increase in the reliability of the ratings both immediately and distally following the training. In sum, FOR training was effective at facilitating the development of a deeply encoded conceptualization of evaluating KSs that was consistently applied to immediate and future (one-day following training) KS evaluations.

The Effectiveness of Referent Material and Reduced Complexity

As discussed above, researchers who previously studied various KS evaluation techniques had found that (a) the complexity of the evaluation and (b) not having referent material to use when conducting the evaluation increases the cognitive demands of the KS evaluation process. Although this study did not directly duplicate prior research results, the findings showed that referent material did effectively increase reliability among those who evaluated less complex KSs. Based on this result, it is assumed that the influence that retrieval failures have on the retrieval process, and the influence that having a limited storage capacity has on the storage process is mitigated by combining referent material with less complex knowledge structures (shown as the blue and pink lines, respectively, in Figure 38 above).

Additional simple effects showed that those with more knowledge of the operation benefited from the referent material. As a result, it is assumed that the influence that retrieval failures had on the retrieval process is mitigated when those with more domain knowledge use referent material (shown as the orange line in Figure 38 above). Furthermore, those with less domain knowledge benefited from reducing the complexity of a KS; therefore, the influence that having a limited storage capacity has on the storage process is mitigated among those with low domain knowledge when evaluating less complex knowledge structures (shown as the yellow line in Figure 38 above).

Overall, it was concluded that the effectiveness of a mitigation technique may often be dependent on its use with other mitigation techniques, or on the characteristics of the rater (e.g., domain knowledge).

Study Limitations

It is important to note the difficulties faced when conducting evaluations that rely on subjective assessments, such as KS evaluations. Research, dating back a half a century ago on performance ratings, evidences the endless issues that may be encountered during the evaluation process. Although KS evaluations do not share the same complexity associated with evaluating human behaviors, they do present challenges that arise from attempting to evaluate information that has been stored in the complex, ever-acquiring structures within memory. This study is among the first to methodically investigate, and attempt to mitigate these challenges.

Training Design

One limitation of the study was the design used to investigate whether frame-of-reference (FOR) training increased the reliability of the KS ratings. Given the constraints with the data collection process, a within-subjects design was used to test the effectiveness of the training. This design, as opposed to using a control condition, does not allow for the elimination of confounding variables. Therefore, it may be the case that the 25% increase in reliability from pre to post training was related to the participants becoming more practiced at, or more familiar with the evaluation process. Future studies must further investigate whether FOR training accounts for a significant portion of the increase in reliability, and determine what other factors may have contributed to the 25% increase seen here.

Generalizability

Another limitation of the study is its lack of generalization to the targeted population. As discussed in the introduction, the goal of the study is to identify methods and procedures for implementing KS evaluations in environments where knowledge of complex systems and procedures are being learned. The participants in this study do not represent the targeted population. This study, however, provides researchers with a framework and methods for studying KS evaluations in more complex learning environments.

Additional Rater Error

Finally, aside from the halo effect, this study did not specifically attempt to mitigate other types of common rater errors. As a result, some of the participants' ratings tended to be restricted to the higher end of the scale. This range restriction may have actually inflated the reliability of the participants' ratings, thus masking the true effects of the mitigation techniques. For example, the overall tendency of the rater's to elicit biases may have prevented them from effectively using their domain knowledge.

Implications

Despite its limitations, this study provides both immediate and future implications.

Future Implications

There are several avenues of research to follow based on the results of this study including: (a) exploring the boundaries of the mitigation methods, (b) exploring the effects of mixing mitigation methods on cognitive processes, and (c) further exploring the influence of domain knowledge on the evaluation process.

Explore the Boundaries of the Mitigation Methods. This study showed that reducing the complexity of a Knowledge Structure (KS) was effective at increasing reliability among rater's with low knowledge of the domain. Researchers should investigate at what point a rater with more knowledge of the domain is affected by the complexity of the KS. Furthermore, the referent model assisted those with more knowledge of the domain, even among participants who had a minimal understanding of the domain. Researchers should determine whether this effect is true at expert levels of domain knowledge; or, whether raters with more domain knowledge encounter more demands due to conflicts between what is represented in the referent material, and what they have stored in memory.

Exploring the Effects of Mixing Mitigation Methods on Cognitive Processes. The interaction found among complexity and referent material is indicative of a storage-by-retrieval interaction. Essentially, whether the retrieval process was affected by retrieval failures was dependent on whether the capacity of the episodic buffer was exceeded. Future studies must further investigate the interactions between the cognitive processes that occur during KS evaluations, and specify conditions under which a mitigation method may or may not be effective.

Further Exploring the Influence of Domain Knowledge. The finding that referent material was effective at increasing reliability among those with more domain knowledge suggests that a certain amount of domain knowledge is necessary for the referent material to be effective. This finding, however, may be related to the referent material used. Raters with less knowledge of a domain may not successfully use referent material that only presents an ideal KS. Future research should investigate whether other forms of referent material (e.g., an inventory of propositions or power point slides containing domain information) will assist those with minimal knowledge of the domain.

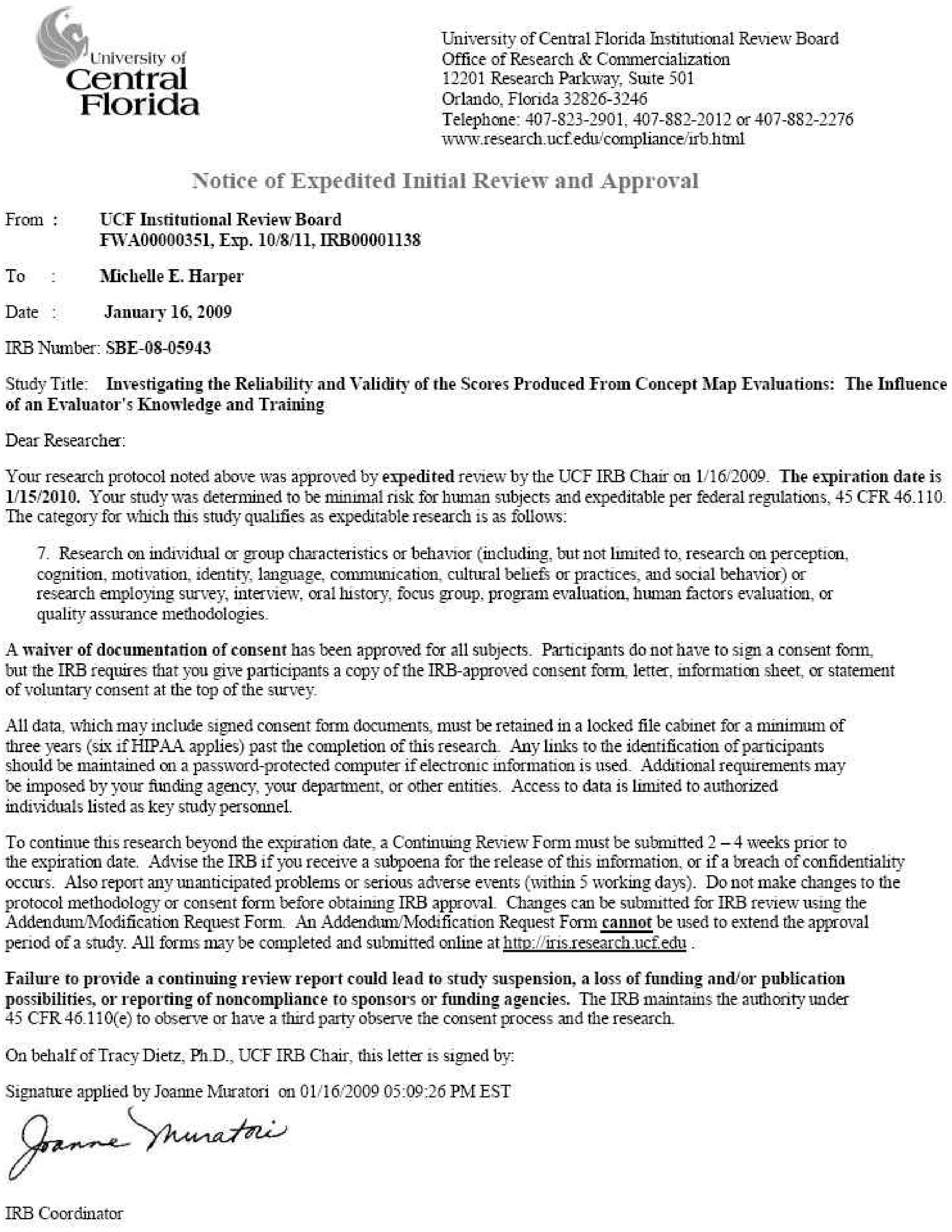
Immediate, Practical Implications

Several practical implications were identified in the form of guidelines to follow when implementing the KS evaluation method.

> *Guideline 1:* Provide raters with training that explains what dimensions are important to evaluate within a KS, and how to provide a rating that represents those dimensions.
>
> *Guideline 2:* If a KS consists of a lower number of concepts (7 or less here), and then provide the rater with a referent KS containing propositions that the organization sees as effectively defining the operation.
>
> *Guideline 3:* If a rater has little to no knowledge of the operation being assessed, then only assign KSs that have fewer concepts.

Guideline 4: If raters have knowledge of the operation, then provide them with a referent KS.

APPENDIX A: IRB APPROVAL FORM



University of Central Florida

University of Central Florida Institutional Review Board
Office of Research & Commercialization
12201 Research Parkway, Suite 501
Orlando, Florida 32826-3246
Telephone: 407-823-2901, 407-882-2012 or 407-882-2276
www.research.ucf.edu/compliance/irb.html

Notice of Expedited Initial Review and Approval

From : **UCF Institutional Review Board**
FWA00000351, Exp. 10/8/11, IRB00001138

To : **Michelle E. Harper**

Date : **January 16, 2009**

IRB Number: **SBE-08-05943**

Study Title: **Investigating the Reliability and Validity of the Scores Produced From Concept Map Evaluations: The Influence of an Evaluator's Knowledge and Training**

Dear Researcher:

Your research protocol noted above was approved by **expedited** review by the UCF IRB Chair on 1/16/2009. **The expiration date is 1/15/2010.** Your study was determined to be minimal risk for human subjects and expeditable per federal regulations, 45 CFR 46.110. The category for which this study qualifies as expeditable research is as follows:

> 7. Research on individual or group characteristics or behavior (including, but not limited to, research on perception, cognition, motivation, identity, language, communication, cultural beliefs or practices, and social behavior) or research employing survey, interview, oral history, focus group, program evaluation, human factors evaluation, or quality assurance methodologies.

A **waiver of documentation of consent** has been approved for all subjects. Participants do not have to sign a consent form, but the IRB requires that you give participants a copy of the IRB-approved consent form, letter, information sheet, or statement of voluntary consent at the top of the survey.

All data, which may include signed consent form documents, must be retained in a locked file cabinet for a minimum of three years (six if HIPAA applies) past the completion of this research. Any links to the identification of participants should be maintained on a password-protected computer if electronic information is used. Additional requirements may be imposed by your funding agency, your department, or other entities. Access to data is limited to authorized individuals listed as key study personnel.

To continue this research beyond the expiration date, a Continuing Review Form must be submitted 2 – 4 weeks prior to the expiration date. Advise the IRB if you receive a subpoena for the release of this information, or if a breach of confidentiality occurs. Also report any unanticipated problems or serious adverse events (within 5 working days). Do not make changes to the protocol methodology or consent form before obtaining IRB approval. Changes can be submitted for IRB review using the Addendum/Modification Request Form. An Addendum/Modification Request Form **cannot** be used to extend the approval period of a study. All forms may be completed and submitted online at http://iris.research.ucf.edu .

Failure to provide a continuing review report could lead to study suspension, a loss of funding and/or publication possibilities, or reporting of noncompliance to sponsors or funding agencies. The IRB maintains the authority under 45 CFR 46.110(e) to observe or have a third party observe the consent process and the research.

On behalf of Tracy Dietz, Ph.D., UCF IRB Chair, this letter is signed by:

Signature applied by Joanne Muratori on 01/16/2009 05:09:26 PM EST

Joanne Muratori

IRB Coordinator

APPENDIX B: WAIVER OF CONSENT FORM

INFORMED VOLUNTARY CONSENT TO PARTICIPATE

Please read this consent document carefully before you decide to participate in this study You must be 18 years of age or older to be included in the research study.

1. You are being asked to voluntarily participate in a research study titled "Investigating the Validity of Concept Map Evaluations." As a volunteer, you are asked to participate in our approximately 3-hour study that will take place over 3 consecutive days on both the internet and in the laboratory. For the first and third day you will complete the study on the internet from a location of your choosing. For the second day you must complete the study in our laboratory. On the first day (internet) you will complete two tasks including (a) filling out a demographics form and (b) completing a multiple choice test on steering an automobile. The tasks for Day 1 should take approximately 35 minutes to complete. For the second day, which will be done at the Psychology Building in room 303G, you will complete four tasks including (a) evaluating a concept map, (b) evaluating a second concept map, (c) viewing training slides on how to score concept maps, (d) evaluating a third concept map, and (e) evaluating a fourth concept map. The second day should take approximately 2 hours to complete. Finally, on third day (internet) you will complete two tasks including (a) evaluating a concept map, and then (b) evaluating a second concept map. The third day should take approximately 30 minutes to complete.

You do not have to answer any questions that you do not wish to answer on any of the questionnaires, and have the right to learn more about the study before signing this informed consent form.

2. The purpose of this study is to determine under what conditions participants can reliably and validly evaluate concept maps.

3. The investigator believes there is a slight risk of breach of confidentiality associated with participation. We must link your name with your username in order to give you extra credit. Although a link between your name and your username is recorded, it is stored completely separate from your responses to the tasks in this study. We assure you that ever possible procedure is being taken to maintain your confidentiality.

4. You understand that you will receive no direct benefit other than:
   - An opportunity to learn about concept mapping
   - A copy of any publications resulting from the current study, if requested

5. You understand that participation in face-to-face studies (Day 2) earns more points than participation in online studies (Day 1 and Day 3). Each half hour in a face-to-face study counts as a half (.50) percentage point, whereas each half hour in an online study counts as a quarter (.25) of a percentage point. Points are rounded up. For Day 2, a half (.50) percentage point is awarded for 30 minutes or less whereas 1 percentage point is awarded for 30 minutes or more. Thus, if on Day 2 you participate for 2 hours and 15 minutes you will receive 2.50 percentage points. If you participate on Day 1 or Day 3 for 20 minutes you will receive a quarter (.25) of a percentage point. If you participate on Day 1 or Day 3 for 40 minutes you will receive a half (.50) percentage point.

6. Your identity will be kept confidential. The researcher will make every effort to prevent anyone who is not on the research team from knowing that you gave us information, or what that information is. For example, your name will be kept separate from the information you give, and these two things will be stored in different places. Your information will be assigned a code number. The list connecting your name to this number will be stored on a password protected computer in the Psychology Building Room 303G. Only the experimenter will have access to this computer. When you have completed the study, your code number and name will be permanently removed from the computer. The information we collect from you will be combined with information from other people who took part in this study. When the researcher writes about this study to share what was learned with other researchers, she will write about this combined information. Your name will not be used in any report, so people will not know how you answered or what you did.

7. If you have any questions about this study you should contact the following individual:

   Principal Investigator: Michelle Harper (407) 882-0305

   E-mail: Mharper@ist.ucf.edu

   The person doing this research is Michelle Harper, a Ph.D. student in the Psychology department at UCF. Because the researcher is a graduate student she is being guided by Dr. Florian Jentsch, a UCF faculty supervisor in the Psychology department. If you have any questions about the study or would like to report a problem, please contact Florian Jentsch at 407-882-0304; fjentsch@mail.ucf.edu

8. Your participation in this study is completely voluntary and will not affect your grade or status in any program or class.

9. Your participation in this study may be stopped by the investigator at any time without my consent if it is believed the decision is in your best interest. There will be no penalty or loss of benefits to which you are otherwise entitled at the time your participation is stopped.

10. No out of pocket costs to you may result from your voluntary participation.

11. If you decide to withdraw from further participation in this study, there will be no penalties. To ensure your safely and orderly withdrawn from the study, you should inform the Principal Investigator, Michelle Harper

12. Official government agencies may have a need to inspect the research records from this study, including yours, in order to fulfill their responsibilities.

13. You have been informed that your consent form will be stored under lock and key.

14. If you have any questions about your rights in the study, you may contact:

    Institutional Review Board, University of Central Florida

Office of Research & Commercialization
12201 Research Parkway, Suite 501
Orlando, FL 32826-3246
(407) 823-2901

15. You are aware that if you have any questions about this study and its related procedures and risks, as well as any of the other information contained in this consent form, or would like to review the study materials prior to completing the study you may contact the experimenter at cmresearchstudy2@yahoo.com prior to signing this consent. Your signature below indicates that (a) all your questions have been answered to your satisfaction, (b) you understand what has been explained to you in this consent form about your participation in this study, (c) you feel you do not need any further information to make a decision about whether or not to volunteer as a participant in this study, (d) you give your voluntary informed consent to participate in the research as it has been explained to you, and (e) you acknowledge that you may receive a copy of this form from the experimenter for your own personal records.

If you do not agree with the following statements, please do not submit your responses to the task associated with this study.

By submitted my responses to the tasks associated with this study I am indicating that:

□ I have read and completely understand the information contained within this document

□ I voluntarily agree to take part in this study

□ I am at least 18 years of age or older

APPENDIX C: BIOGRAPHICAL FROM

Demographic Questions

(1) What is your gender?

Female ○ Male ○

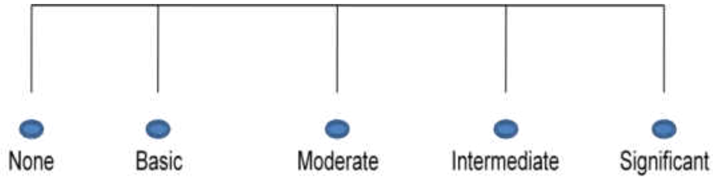
(2) What is your age? _________

(3) What year are you in college?

Freshman ○ Sophomore ○ Senior ○ Graduate ○

Automotive Mechanics Experience

(1) Please rate your level of experience with automotive mechanics by circling one of the points on the scale below. The descriptions of each point on the scale below should assist you with making your rating.



NONE = I have no experience in performing Automotive mechanic tasks
BASIC = I have performed basic activities related to automotive mechanics in a limited number of different situations
MODERATE = I have performed basic activities related to automotive mechanics in a wide variety of situations
INTERMEDIATE = I have performed complex activities related to automotive mechanics in a limited number of different situations
SIGNIFICANT = I have performed complex activities related to automotive mechanics this task in a wide variety of situations

Automotive Steering Experience

(2) How many times have you conducted mechanical tasks associated with an automobile steering system?

0 times_______ 1 to 20 times _______ 20 or more times _____

(3) How long have you been conducting mechanical tasks associated with an automobile steering system?

N/A _______ Less than 5 years _______ More than 5 years _______

(4) Have you conducted mechanical tasks associated with an automobile steering system in any of your jobs? Yes _______ No _______

(4a) If yes, for how many years did you do this job? _______

Driving Experience Questions

(1) How many years have you been driving with or without a permit/license? _________

(2) How long have you held a driver's permit/ driver's license? __________

(3) How many days do you drive a car in a typical week? (circle one)

Less than once a week 1 2 3 4 5 6 7

(4) What kind of vehicle do you drive most often?

a. Car
b. Van or minivan
c. Sport utility vehicle
d. Pickup truck
e. Other truck
f. Motorcycle
g. Other (SPECIFY) __________

(5) How many accidents have you been involved in over the past year when you were the driver? _________

(6) How would you rate the quality of your driving?

Poor Excellent

○ ○ ○ ○ ○

APPENDIX D: STEERING KNOWLEDGE TEST

Steering Knowledge Test

Please only circle *one* answer

1. A car is steered directly by the driver with a ______________________.

   A. steering shaft
   B. steering linkage
   C. steering wheel
   D. tires

2. What type of force is needed from a driver to initiate the steering process?

   A. rotational
   B. lateral
   C. vertical
   D. spinning

3. What is the purpose of power steering?

   A. it aids in acceleration
   B. it helps to charge the car's battery
   C. it makes it easier to turn the steering wheel
   D. it makes it easier to go in reverse

4. If the power steering stops working it will?

   A. make it difficult to accelerate
   B. make it difficult to open the door
   C. make the steering feel heavy
   D. make the steering feel light

5. When you want to turn left or right when driving a car, you will directly turn the _______________.

   A. steering column
   B. steering wheel
   C. steering gear
   D. accelerator

6. Which unit is a steering shaft a part of?

   A. steering linkage assembly
   B. steering gear assembly
   C. power steering pump assembly
   D. steering idler assembly

7. What is meant by the term “under steer”?

   A. the tendency for a vehicle to steer on a smaller turning circle than is expected
   B. the steering wheel is situated in front of the first axle
   C. the steering wheel is situated directly over the first axle
   D. the tendency for a vehicle to steer on a larger turning circle than is expected

8. What steering system does a car commonly have?

   A. recirculating-ball
   B. power assist
   C. active steering
   D. rack-and-pinion

9. In a power steering system, what is the purpose of the pump?

   A. it pumps air through the system
   B. it prevents the system from locking
   C. it provides hydraulic power
   D. it cools the system

10. What is steering ratio?

   A. the ratio of how much power the drive must use on the steering wheel, to the power of the steering system itself
   B. the ratio of how far the driver turns the steering wheel to how far the car’s wheels turn
   C. the ratio of how far the driver turns the steering wheel to how much resistance the wheels have
   D. the number of turns the steering wheel can do before it locks

11. What is steering castor?

    A. a steering joint lubricant
    B. a front hub grease slinger
    C. a steering geometry feature
    D. a constant velocity joint component

12. The camber angle setting of a road wheel determines?

    A. the plane of a road wheel in relation to the vertical
    B. the wheel bearing type
    C. the maximum amount that the steering can be turned towards locks
    D. the maximum rebound action

13. Which force is a steering column most subjected to in normal use?

    A. torsion
    B. sheer
    C. bending
    D. decompression

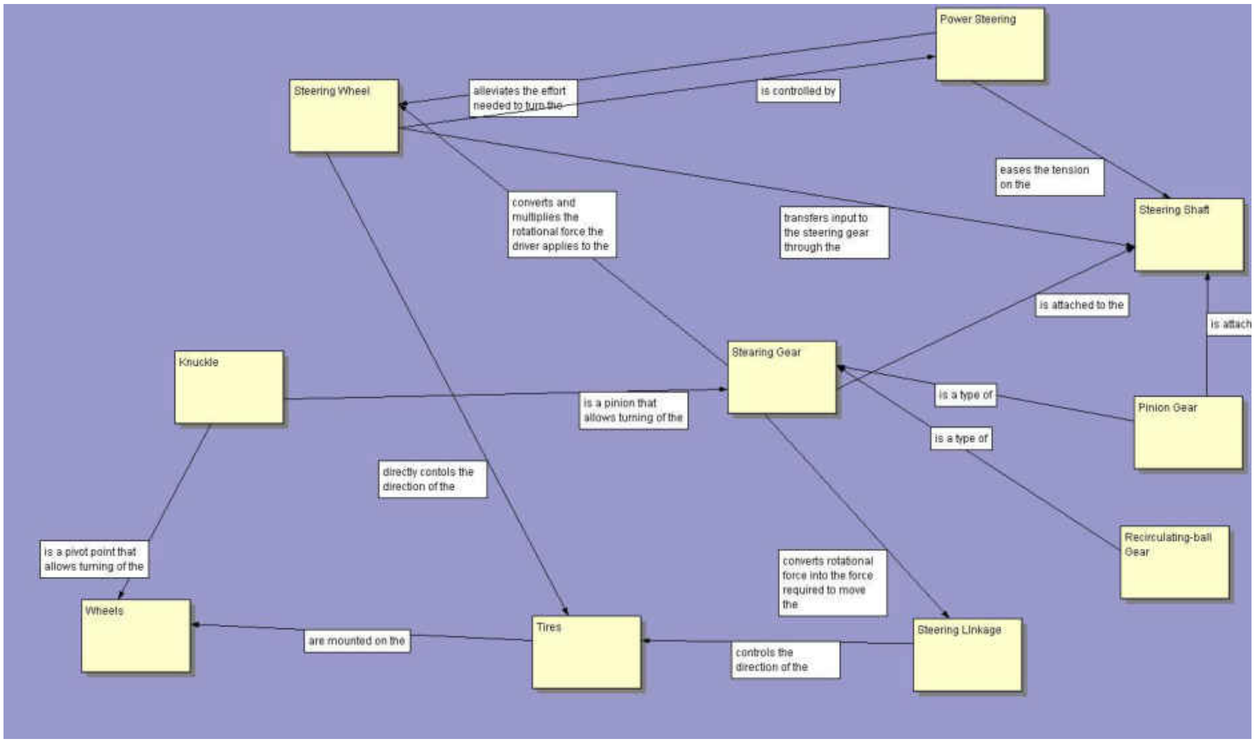
14. Which one of the following determines the amount of steering 'toe out' (Ackerman effect) on locks?

    A. steering arm to stub axle angle
    B. castor
    C. rack and pinion gear ratio
    D. Toe

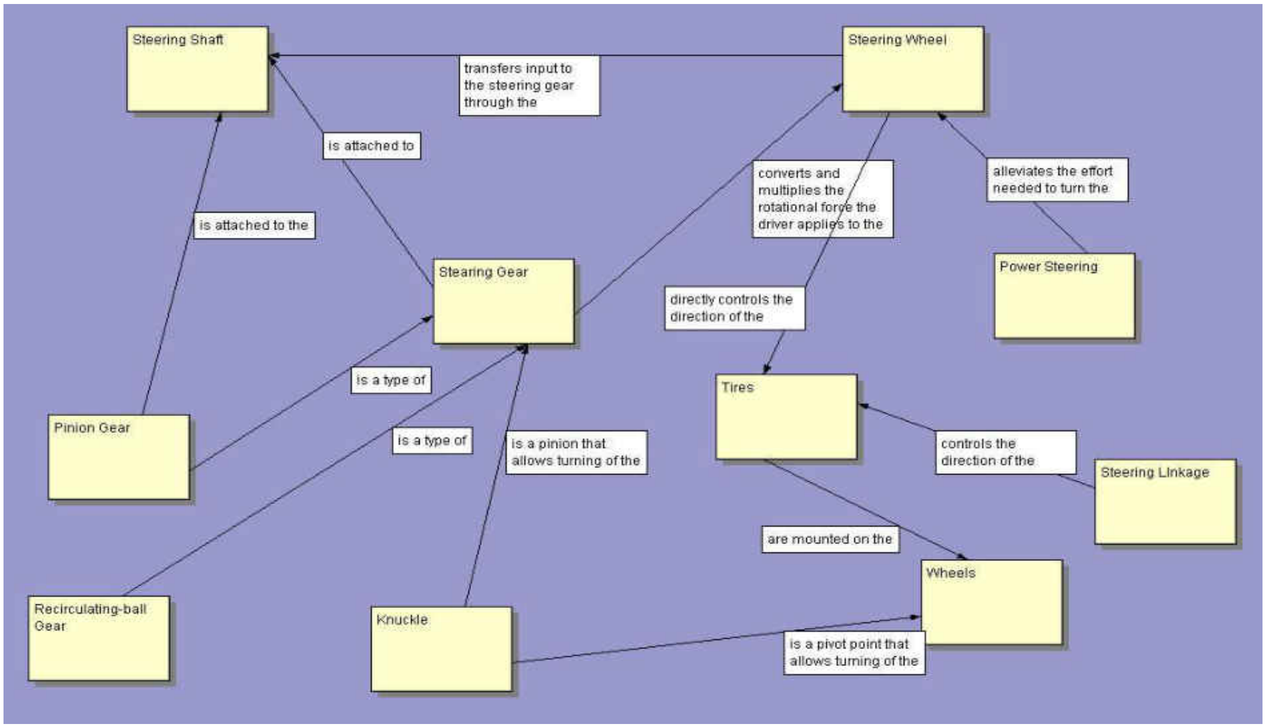
15. What best describes how a rack-and-pinion steering system works?

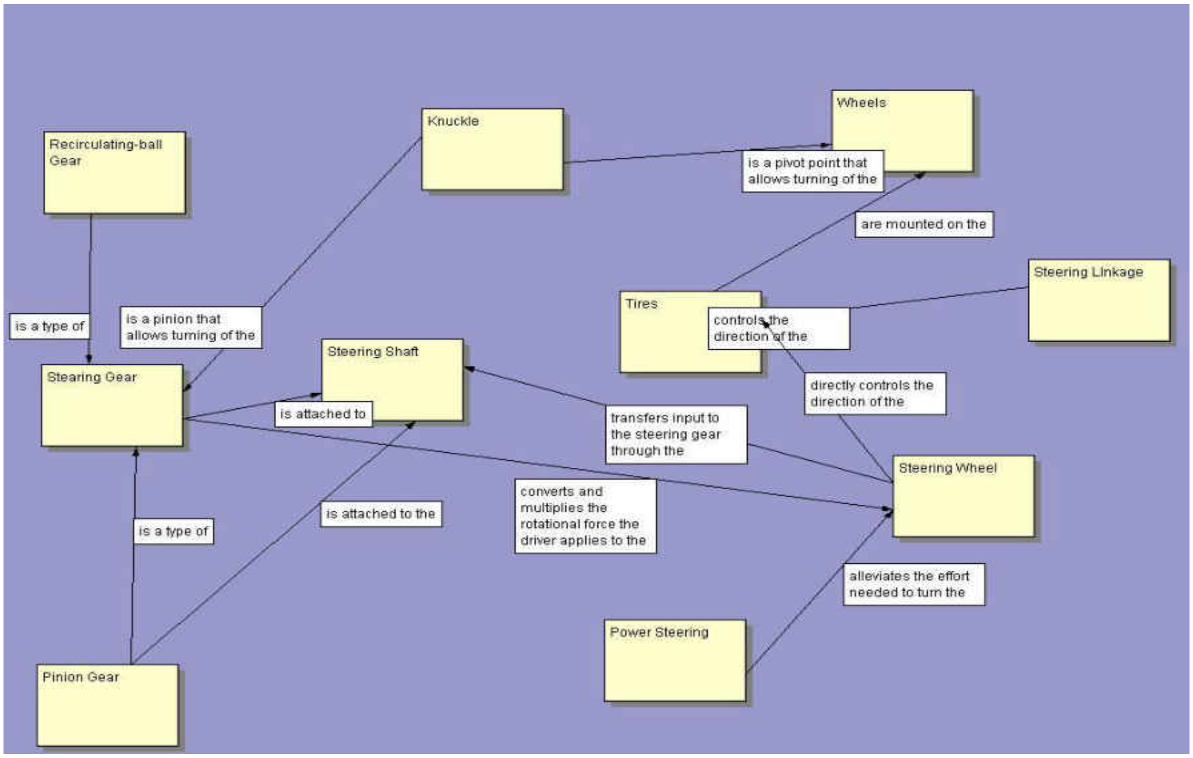
    A. A rack-and-pinion gear set is enclosed in a tube that turns a tire rack. The tire rack turns the cars wheels
    B. A rack is connected to the steering wheel, which turns the pinion which turns the car wheels.
    C. A spindle moves the pinion, which sits on a rack.
    D. The steering arm moves the pinion, which powers the rack to turn the wheels.

APPENDIX E: EVALUATED KNOWLEDGE STRUCTURES

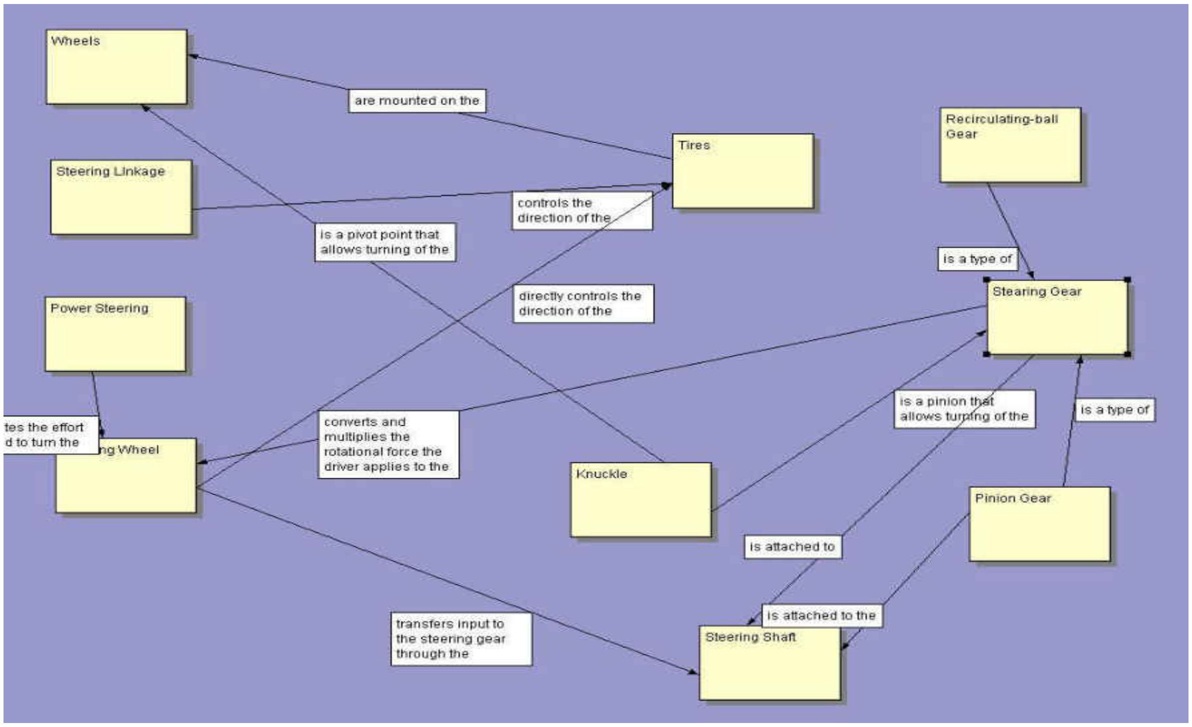

Power Steering
Steering Wheel
alleviates the effort needed to turn the
is controlled by
eases the tension on the
Steering Shaft
converts and multiplies the rotational force the driver applies to the
transfers input to the steering gear through the
is attached to the
is attach
Knuckle
Stearing Gear
is a type of
Pinion Gear
is a pinion that allows turning of the
is a type of
directly contols the direction of the
Recirculating-ball Gear
is a pivot point that allows turning of the
converts rotational force into the force required to move the
Wheels
are mounted on the
Tires
controls the direction of the
Steering Linkage

KS A and B

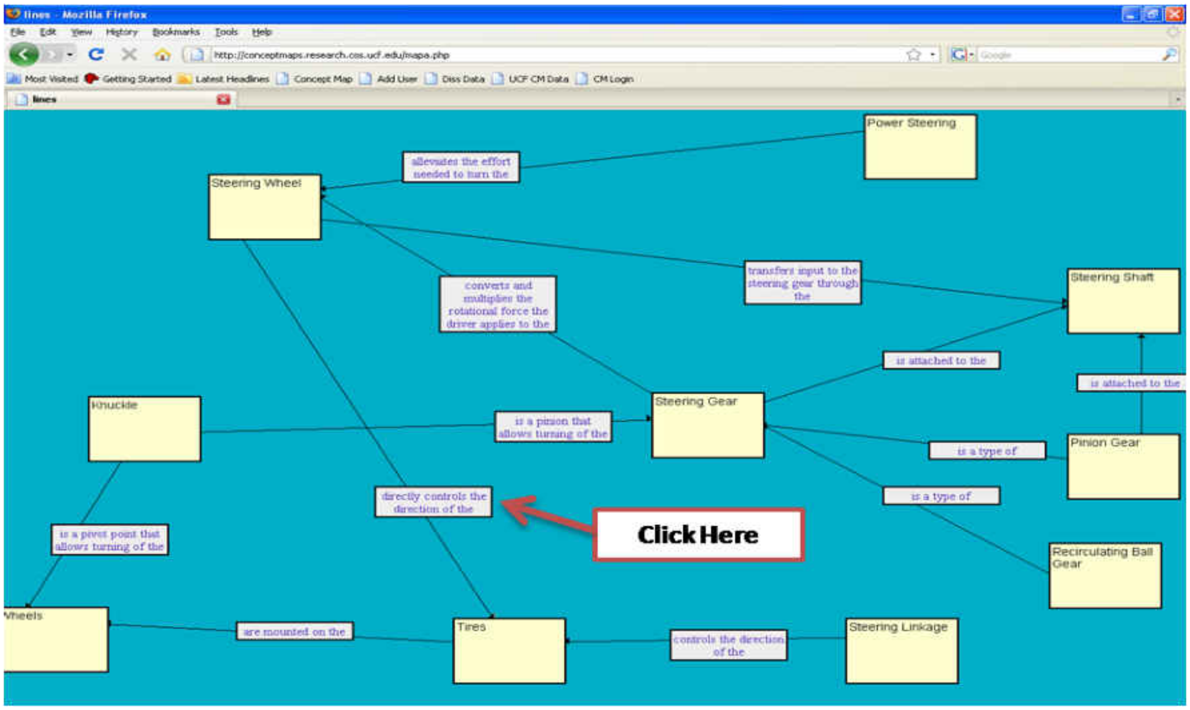

Steering Shaft
transfers input to the steering gear through the
Steering Wheel
is attached to
is attached to the
converts and multiplies the rotational force the driver applies to the
alleviates the effort needed to turn the
Stearing Gear
Power Steering
directly controls the direction of the
is a type of
Tires
Pinion Gear
is a type of
is a pinion that allows turning of the
controls the direction of the
Steering Linkage
are mounted on the
Wheels
Recirculating-ball Gear
Knuckle
is a pivot point that allows turning of the


Recirculating-ball Gear
Knuckle
Wheels
is a pivot point that allows turning of the
are mounted on the
Steering Linkage
Tires
controls the direction of the
is a type of
is a pinion that allows turning of the
Steering Shaft
Stearing Gear
is attached to
directly controls the direction of the
transfers input to the steering gear through the
Steering Wheel
converts and multiplies the rotational force the driver applies to the
is a type of
is attached to the
alleviates the effort needed to turn the
Power Steering
Pinion Gear

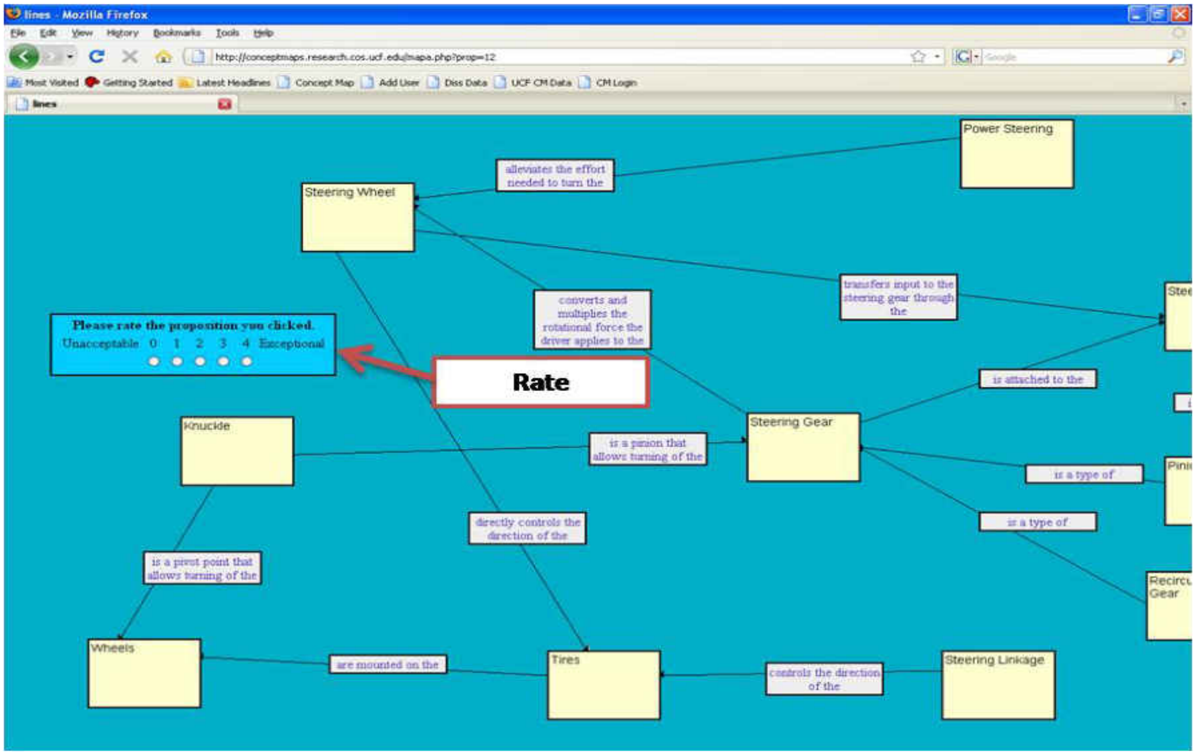
KS C and D


Wheels
are mounted on the
Tires
Recirculating-ball Gear
Steering Linkage
controls the direction of the
is a pivot point that allows turning of the
is a type of
directly controls the direction of the
Power Steering
Stearing Gear
tes the effort d to turn the
ng Wheel
converts and multiplies the rotational force the driver applies to the
is a pinion that allows turning of the
is a type of
Knuckle
Pinion Gear
is attached to
is attached to the
transfers input to the steering gear through the
Steering Shaft

APPENDIX F: EVALUATION PROCESS


lines - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://conceptmaps.research.cos.ucf.edu/mapa.php
Most Visited Getting Started Latest Headlines Concept Map Add User Diss Data UCF CM Data CM Login
lines
Power Steering
alleviates the effort needed to turn the
Steering Wheel
transfers input to the steering gear through the
Steering Shaft
converts and multiplies the rotational force the driver applies to the
is attached to the
is attached to the
Knuckle
is a pinion that allows turning of the
Steering Gear
is a type of
Pinion Gear
directly controls the direction of the
is a type of
is a pivot point that allows turning of the
Recirculating Ball Gear
Wheels
are mounted on the
Tires
controls the direction of the
Steering Linkage

Click Here

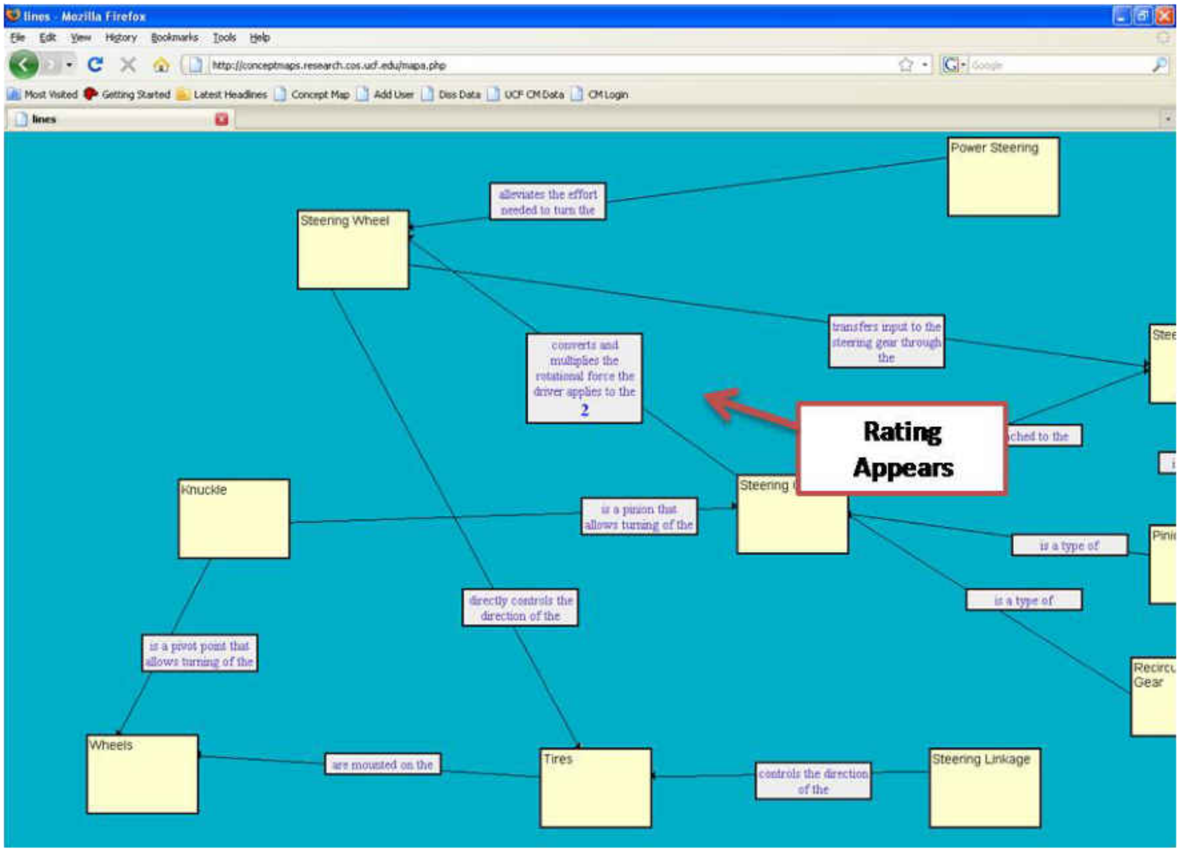

lines - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://conceptmaps.research.cos.ucf.edu/mapa.php?prop=12
Most Visited Getting Started Latest Headlines Concept Map Add User Diss Data UCF CM Data CM Login
lines
Power Steering
alleviates the effort needed to turn the
Steering Wheel
transfers input to the steering gear through the
Stee
converts and multiplies the rotational force the driver applies to the
is attached to the
Knuckle
is a pinion that allows turning of the
Steering Gear
is a type of
Pini
directly controls the direction of the
is a type of
is a pivot point that allows turning of the
Recircu Gear
Wheels
are mounted on the
Tires
controls the direction of the
Steering Linkage

Please rate the proposition you clicked.

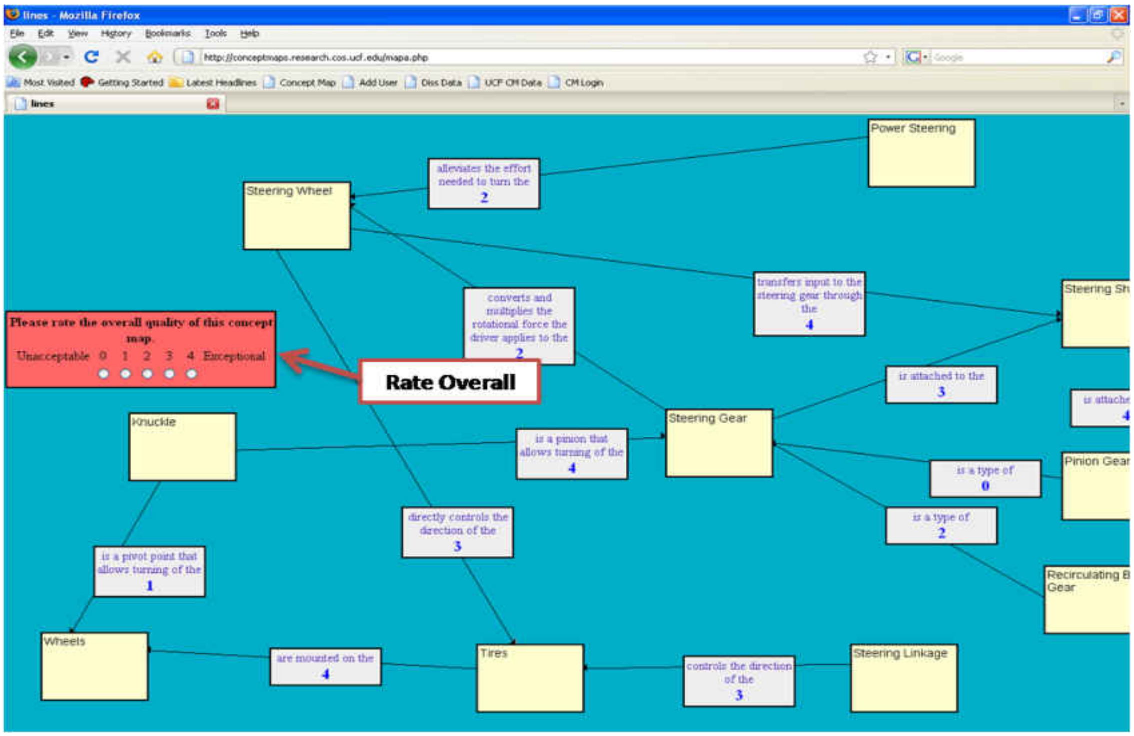
Unacceptable 0 1 2 3 4 Exceptional

○ ○ ○ ○ ○

Rate


lines - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://conceptmaps.research.cos.ucf.edu/mapa.php
Most Visited Getting Started Latest Headlines Concept Map Add User Diss Data UCF CM Data CM Login
lines
Power Steering
alleviates the effort needed to turn the
Steering Wheel
transfers input to the steering gear through the
Stee
converts and multiplies the rotational force the driver applies to the
2
ched to the
Knuckle
is a pinion that allows turning of the
Steering (
is a type of
Pini
is a type of
directly controls the direction of the
is a pivot point that allows turning of the
Recircu Gear
Wheels
are mounted on the
Tires
controls the direction of the
Steering Linkage

Rating Appears


lines - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://conceptmaps.research.cos.ucf.edu/mapa.php
Most Visited Getting Started Latest Headlines Concept Map Add User Diss Data UCF CM Data CM Login
lines
Power Steering
alleviates the effort needed to turn the
2
Steering Wheel
transfers input to the steering gear through the
4
Steering Sh
converts and multiplies the rotational force the driver applies to the
2
is attached to the
3
is attache
4
Knuckle
is a pinion that allows turning of the
4
Steering Gear
Pinion Gear
is a type of
0
is a type of
2
directly controls the direction of the
3
Recirculating B Gear
is a pivot point that allows turning of the
1
Wheels
are mounted on the
4
Tires
controls the direction of the
3
Steering Linkage

Please rate the overall quality of this concept map.

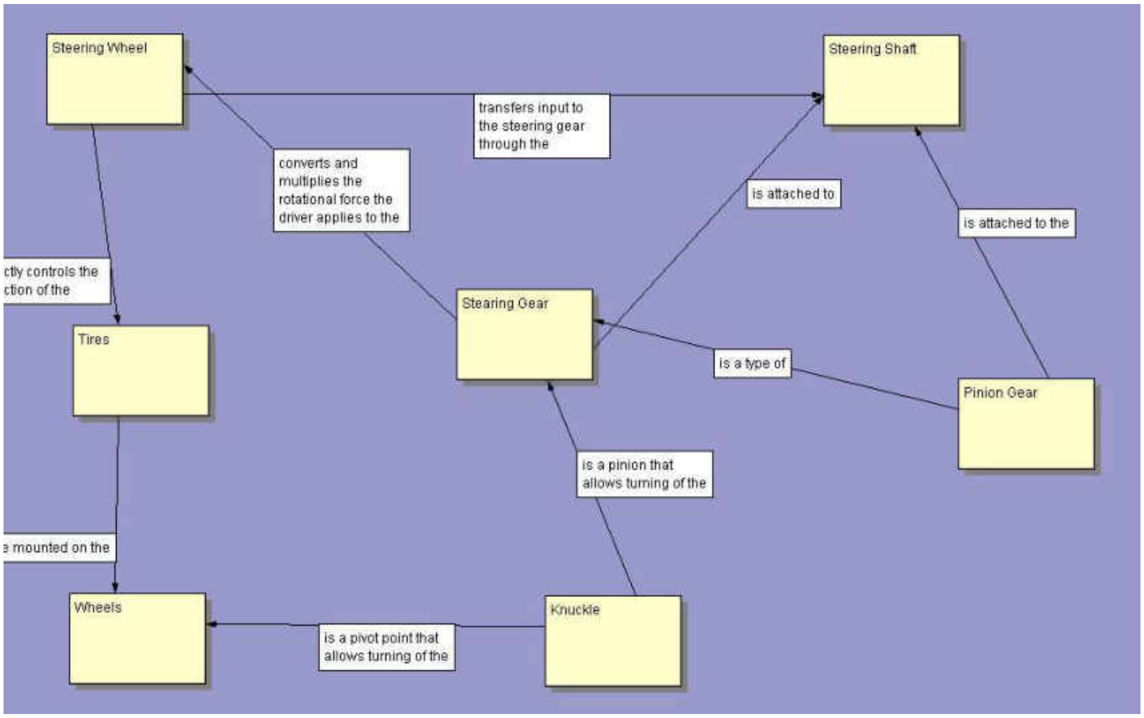
Unacceptable 0 1 2 3 4 Exceptional

○ ○ ○ ○ ○

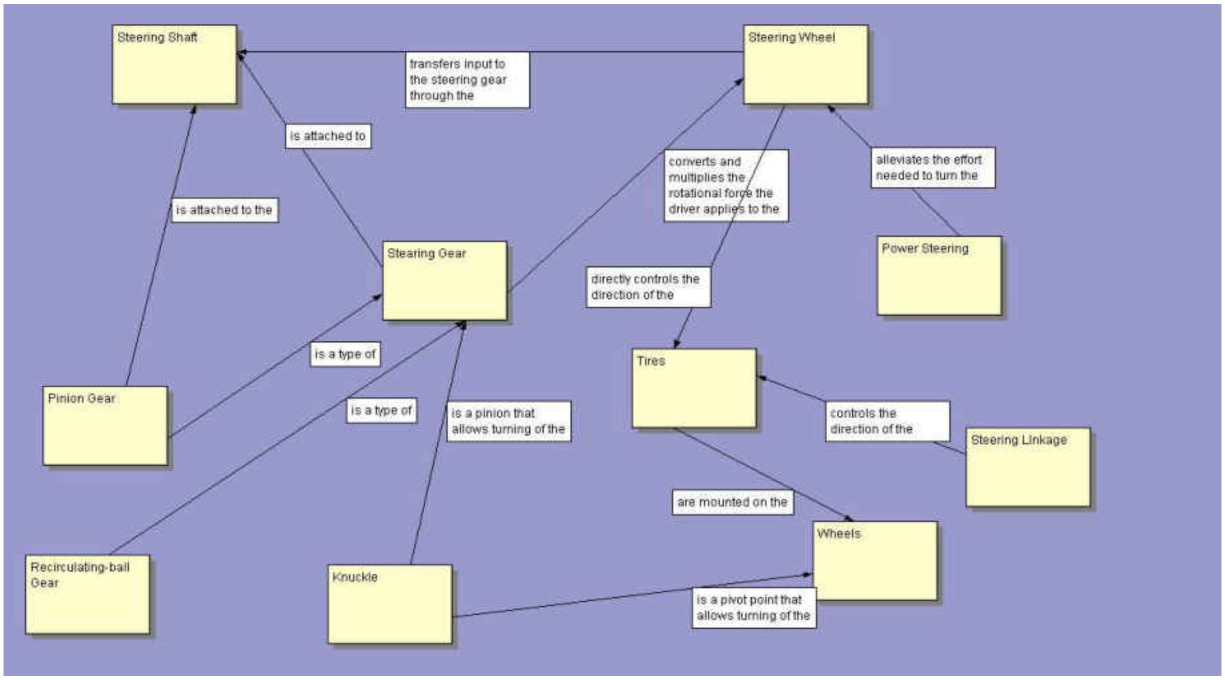
Rate Overall

APPENDIX G: 10 CONCEPT KS VS. 7 CONCEPT KS

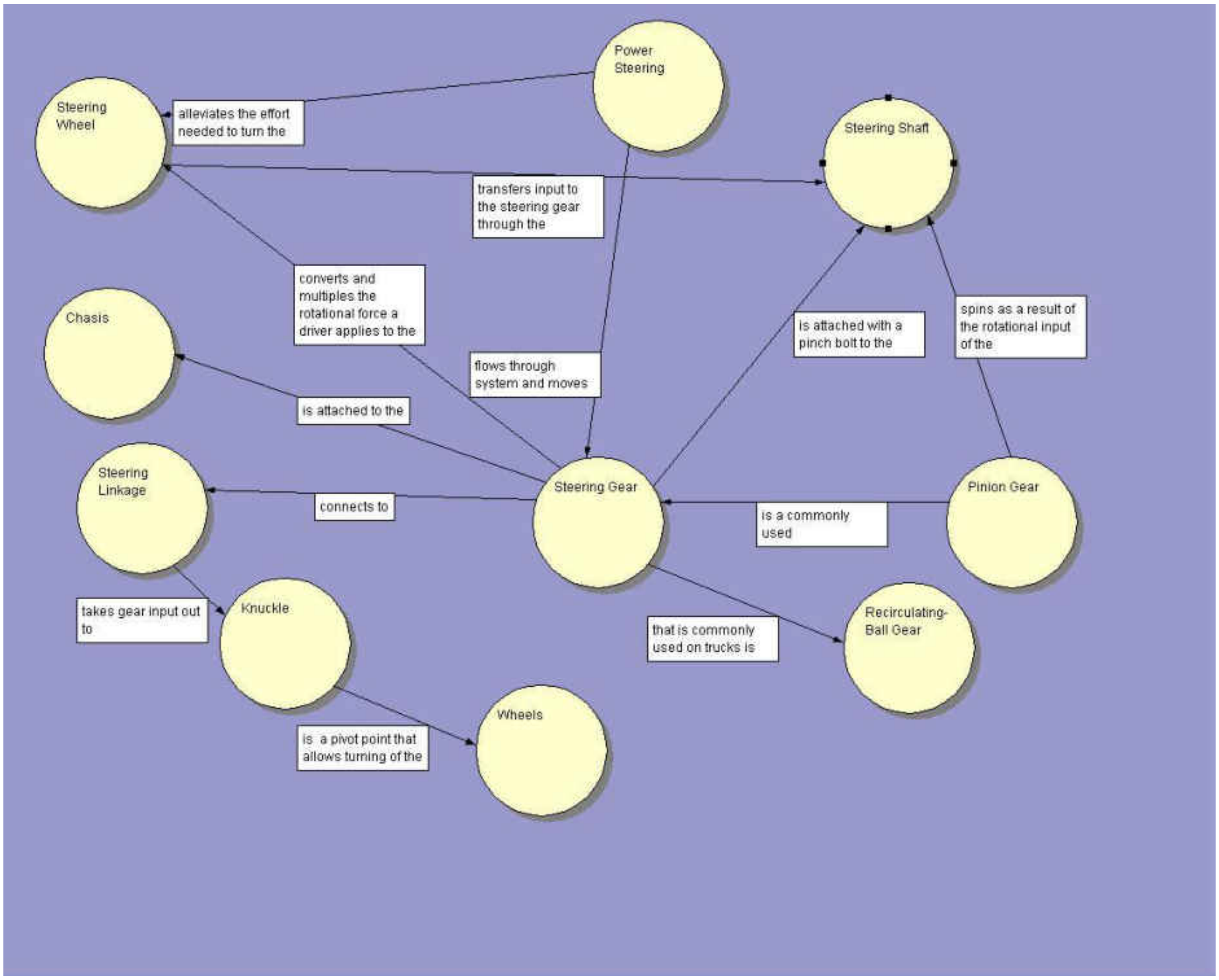
7 Concepts


Steering Wheel
transfers input to the steering gear through the
Steering Shaft
converts and multiplies the rotational force the driver applies to the
is attached to
is attached to the
ctly controls the ction of the
Stearing Gear
is a type of
Pinion Gear
Tires
is a pinion that allows turning of the
e mounted on the
Wheels
is a pivot point that allows turning of the
Knuckle

10 Concepts

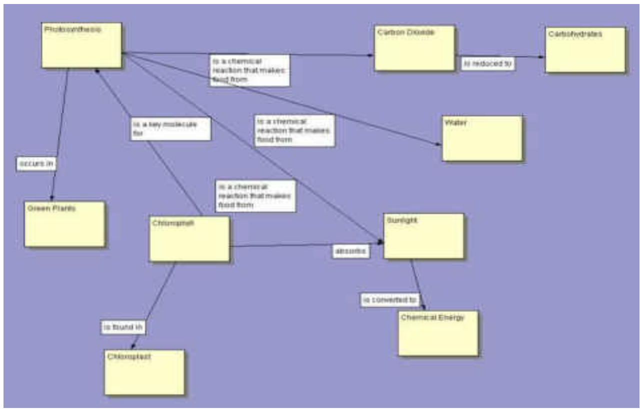

Steering Shaft
transfers input to the steering gear through the
Steering Wheel
is attached to
converts and multiplies the rotational force the driver applies to the
alleviates the effort needed to turn the
is attached to the
Stearing Gear
Power Steering
directly controls the direction of the
is a type of
is a type of
Pinion Gear
Tires
is a pinion that allows turning of the
controls the direction of the
Steering Linkage
are mounted on the
Wheels
Recirculating-ball Gear
Knuckle
is a pivot point that allows turning of the

APPENDIX H: REFERENT MATERIAL


Power Steering
Steering Wheel
alleviates the effort needed to turn the
Steering Shaft
transfers input to the steering gear through the
converts and multiples the rotational force a driver applies to the
Chasis
is attached with a pinch bolt to the
spins as a result of the rotational input of the
flows through system and moves
is attached to the
Steering Linkage
Steering Gear
Pinion Gear
connects to
is a commonly used
takes gear input out to
Knuckle
Recirculating-Ball Gear
that is commonly used on trucks is
Wheels
is a pivot point that allows turning of the

APPENDIX I: TRAINING MATERIAL

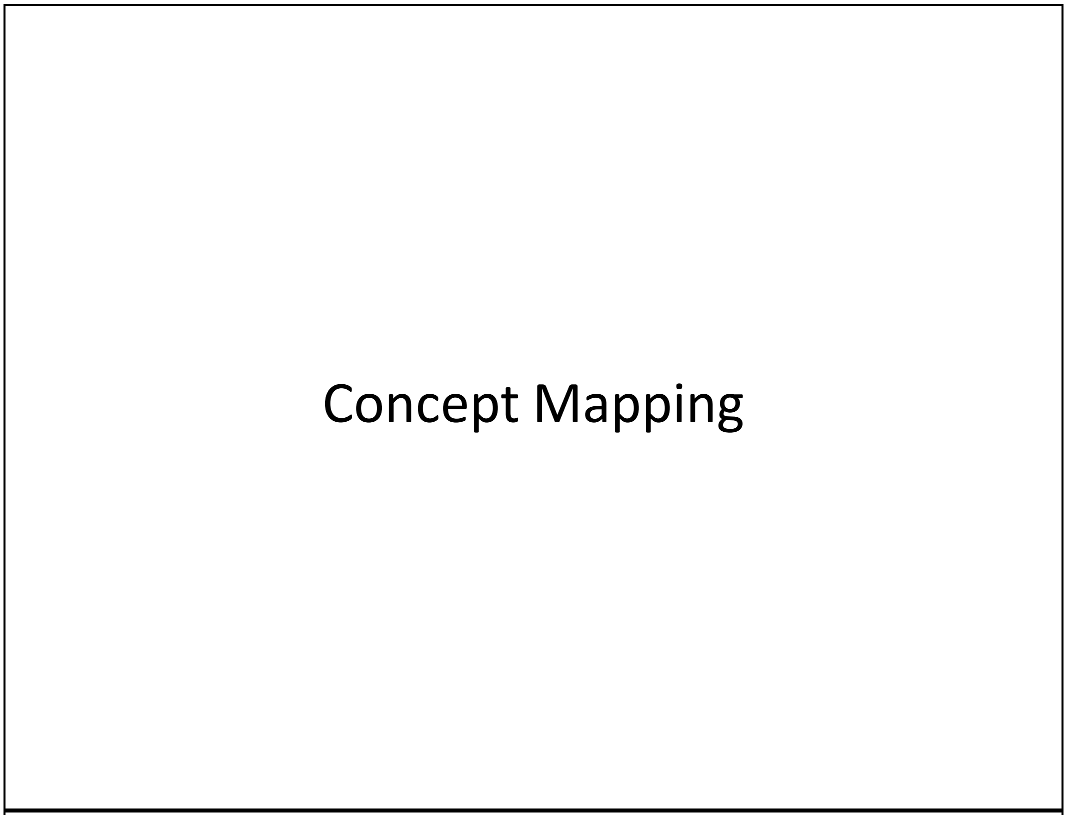
Concept Map Evaluation Training



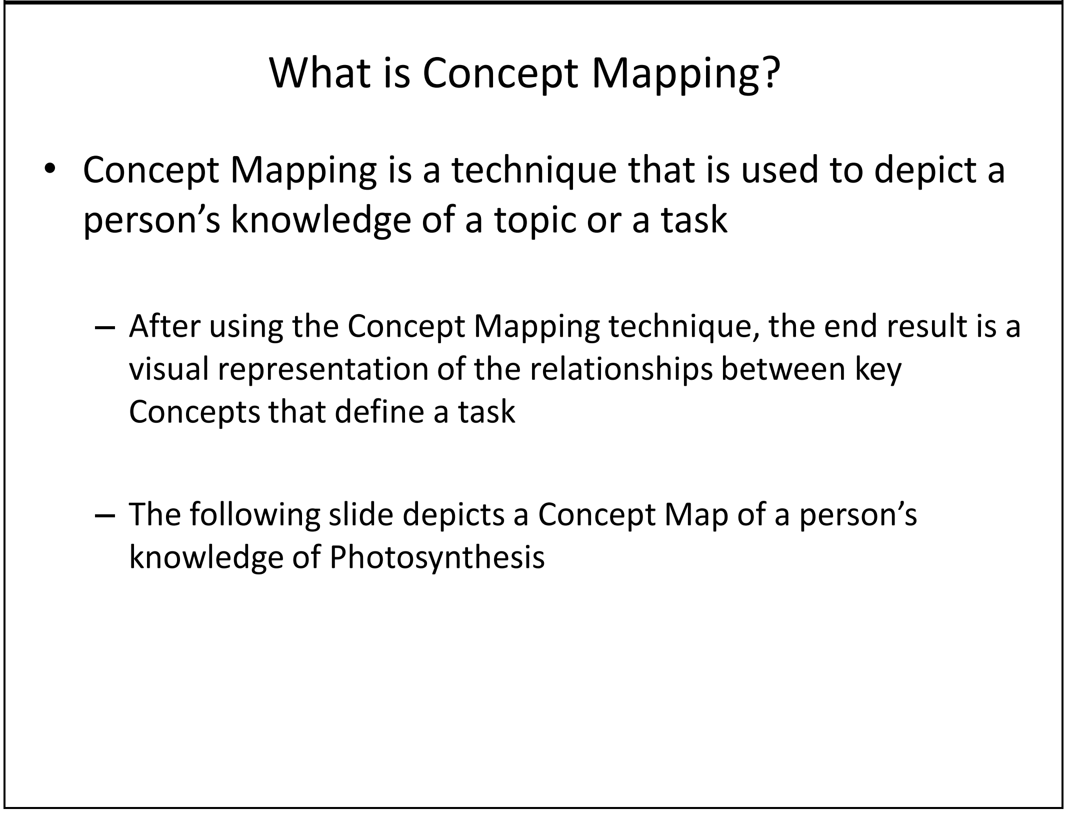
Overview

- In this training, you will learn about Concept Maps and be given specific instructions on how to score Concept Maps

- Following this training you should be able to answer the following questions:
  - What is the Concept Mapping technique?
  - What does a Concept Map look like?
  - How are Concept Maps evaluated?

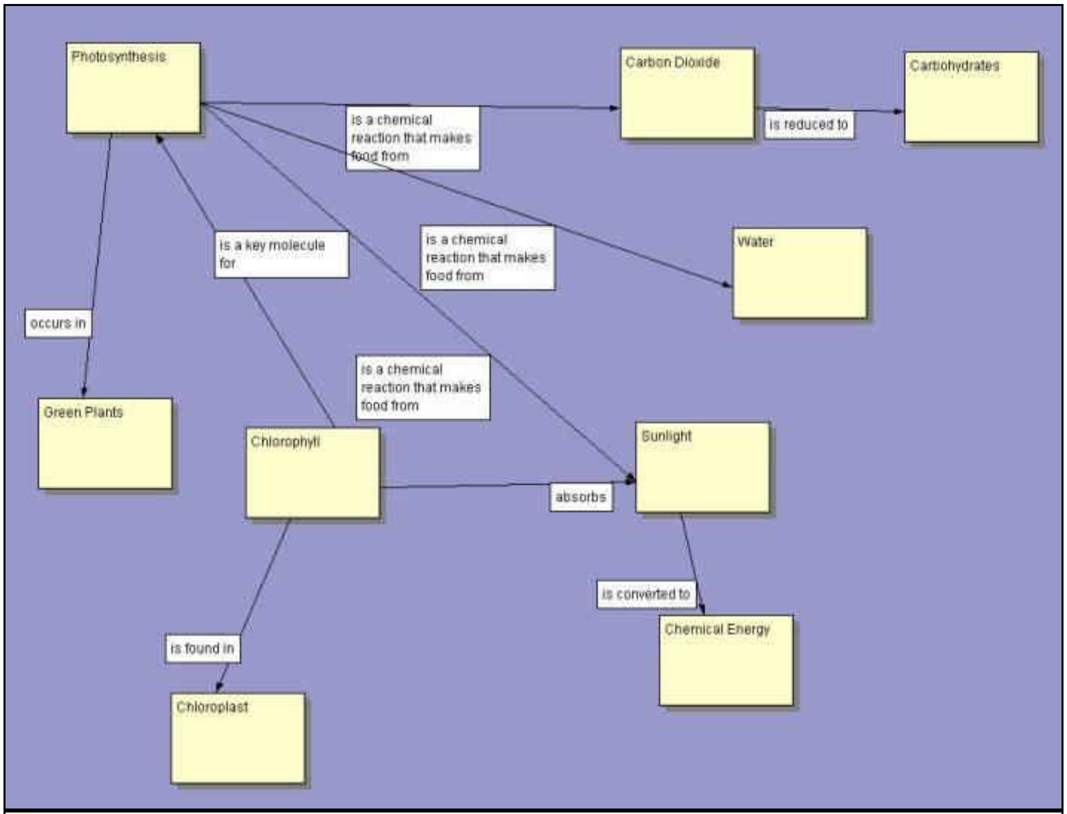


Concept Mapping



What is Concept Mapping?

- Concept Mapping is a technique that is used to depict a person’s knowledge of a topic or a task
  - After using the Concept Mapping technique, the end result is a visual representation of the relationships between key Concepts that define a task
  - The following slide depicts a Concept Map of a person’s knowledge of Photosynthesis

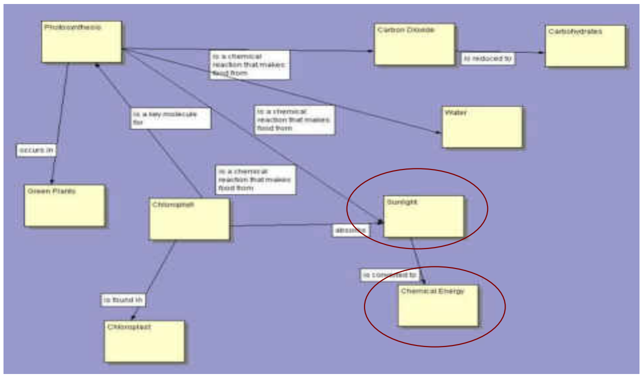


Components of a Concept Map

- As seen on the previous slide, a Concept Map is made up of several components including
  - Concepts, which are the terms that are relevant to the topic or task
  - Links, which are the arrows connecting the Concepts
  - Labels, which describe the relationship between the connected Concepts
- Together, the Concepts, Links, and labels form a Proposition

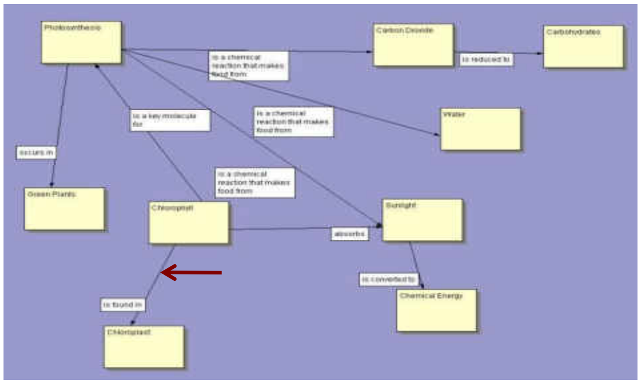
Components of a Concept Map

- Concepts
  - Concepts are key terms that define a topic; without them, it would be difficult to fully define a topic
  - This Concept Map has Concepts that are necessary for defining photosynthesis such as “sunlight” and “chlorophyll”
  - The Concepts within a Concept Map are often pre-selected and given to the creator



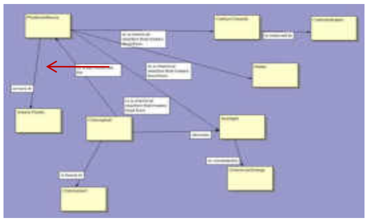
Components of a Concept Map

- Links
  - A Concept Map is created by placing Links between Concepts that are believed to be related
  - For example, in this Concept Map the creator has indicated that there is a relationship between Chlorophyll and Chloroplast

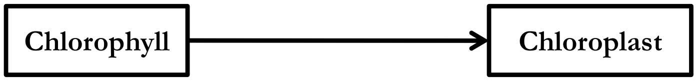


Components of a Concept Map

- Links Cont’d
  - Take note that the Links in the CM have arrows which indicate there is a directional relationship between the concepts



  - As you will learn, the connected concepts form statements about a topic; therefore, the direction of the arrow indicates the flow of the statement

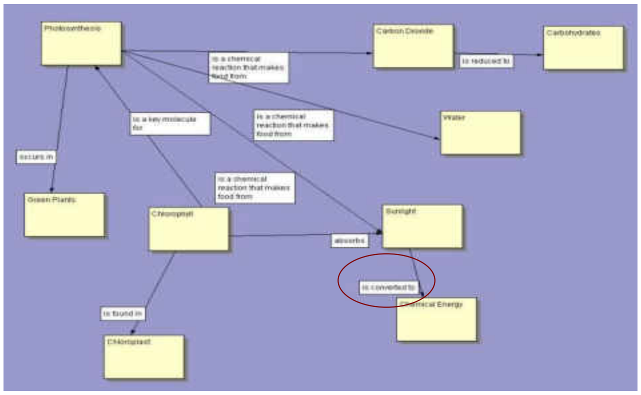


Chlorophyll

Chloroplast

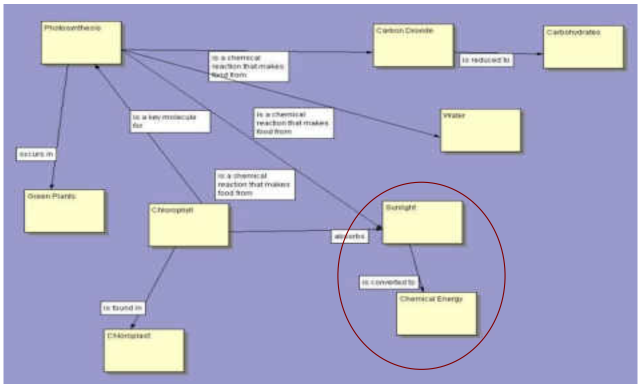
Components of a Concept Map

- Labels
  - Once a Link is created, the user will then provide a Label that describes the relationship between the linked Concepts
  - In this Concept Map the creator has indicated that, in terms of Photosynthesis, the relationship between sunlight and chemical energy is best described using the Label “is converted to”



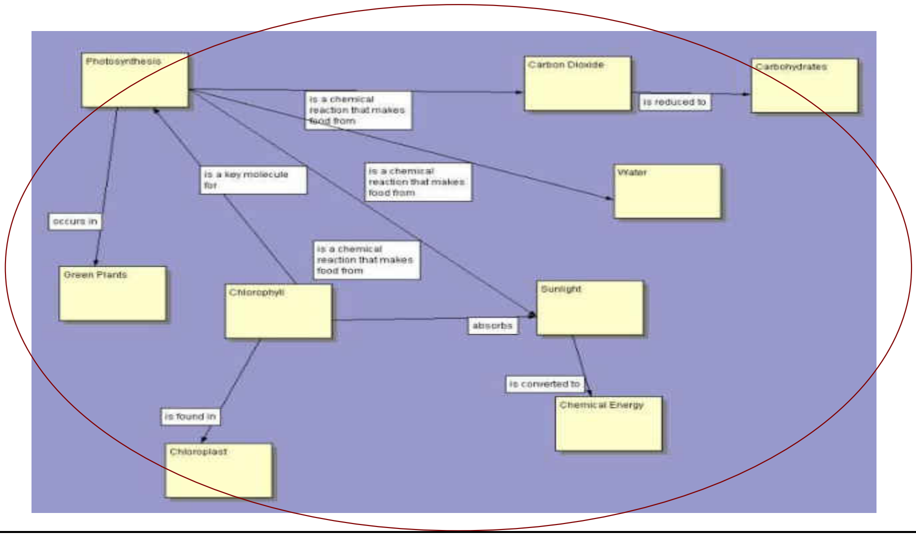
Components of a Concept Map

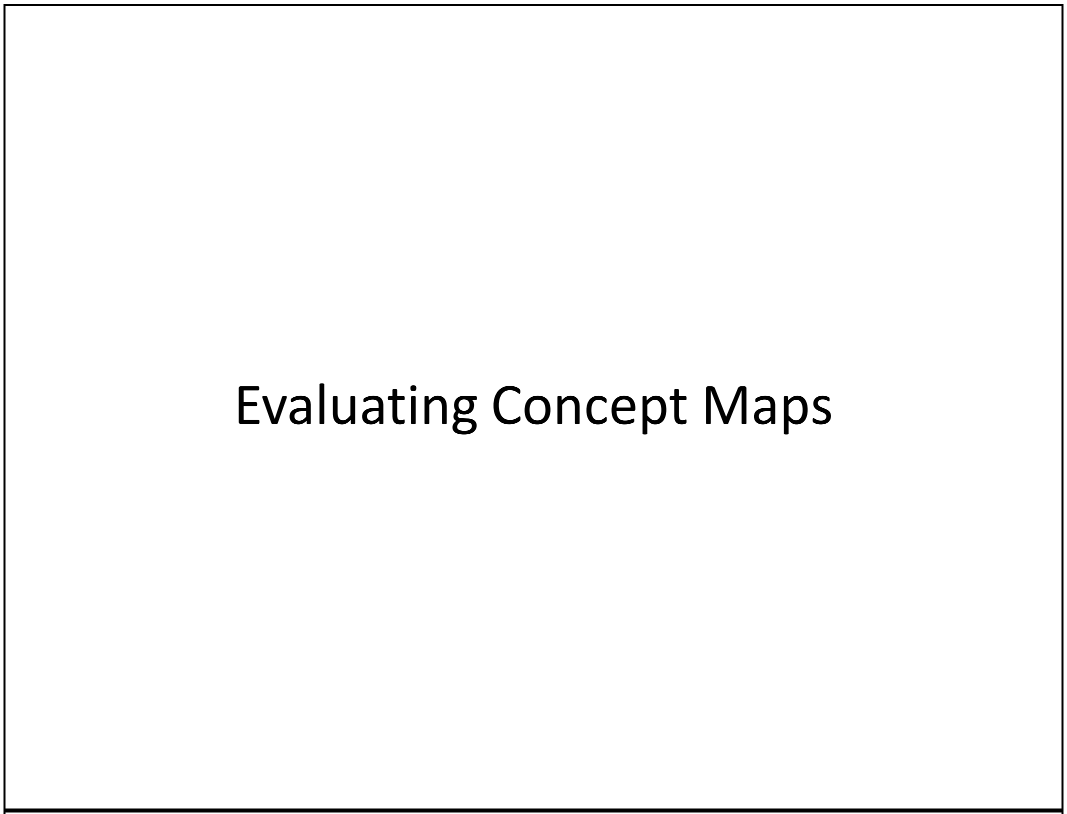
- Propositions
  - Together Concepts, Links, and Labels form what is referred to as a Proposition
  - A Proposition is a statement about some object or event that defines a topic
  - Concept Maps are essentially networks of statements or Propositions that define a topic



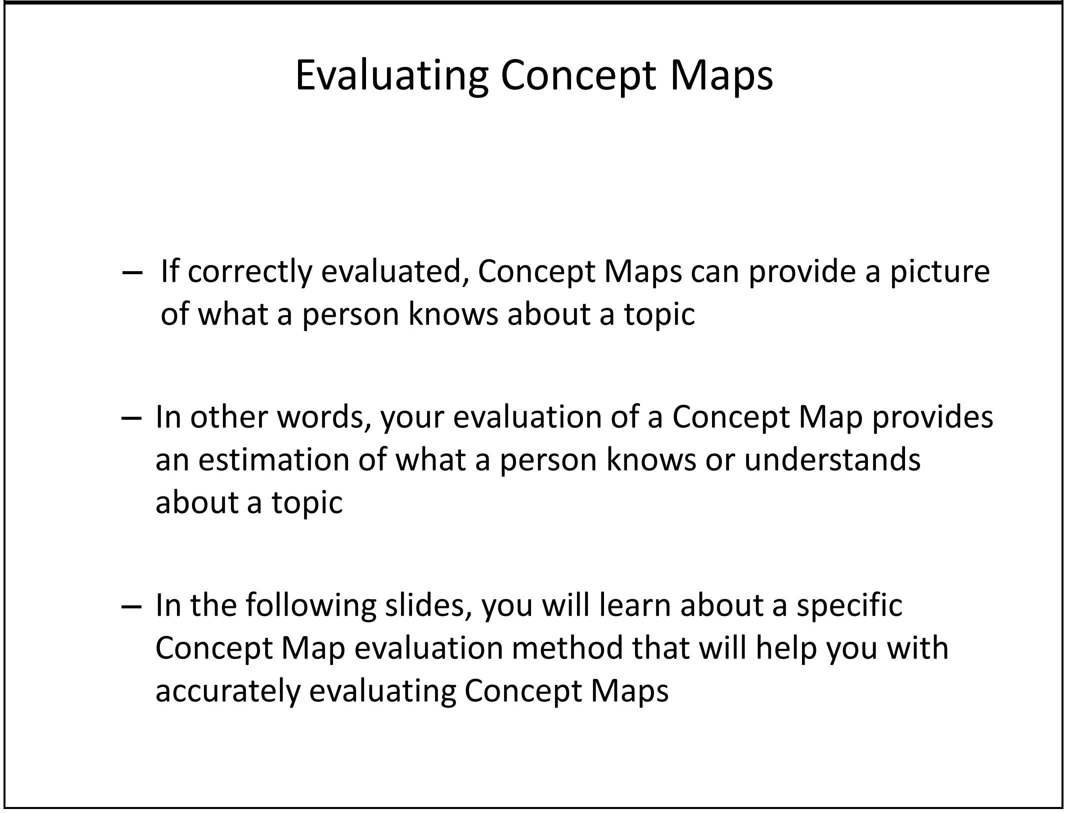
Components of a Concept Map

- Concept Map
  - As a whole, the Concepts, Links, and Labels or Propositions should provide a meaningful definition of the topic (i.e., photosynthesis)





Evaluating Concept Maps



Evaluating Concept Maps

- If correctly evaluated, Concept Maps can provide a picture of what a person knows about a topic

- In other words, your evaluation of a Concept Map provides an estimation of what a person knows or understands about a topic

- In the following slides, you will learn about a specific Concept Map evaluation method that will help you with accurately evaluating Concept Maps

Evaluating Concept Maps

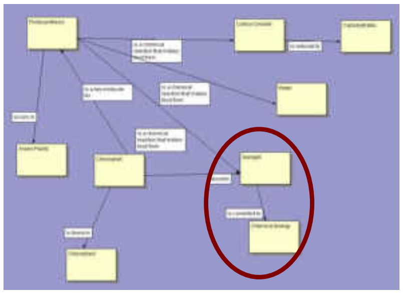
- For the remainder of this study, we ask that you apply the procedures you learn from the following slides to your later Concept Map evaluations

- By using this procedure, you can obtain scores that accurately represent the quality of the Proposition within a Concept Map, and the Concept Map as a whole

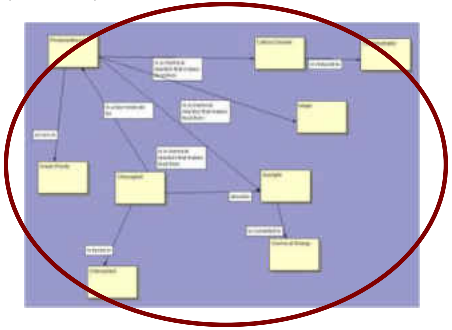
Evaluating Concept Maps

- Proposition vs. Whole Concept Map Evaluations
  - The Concept Map evaluations you completed prior to this training required you to provide ratings for both the individual Proposition and for the Concept Map as a whole

  - Here, you will learn different approaches to Proposition evaluations and Whole Concept Map evaluations



Proposition

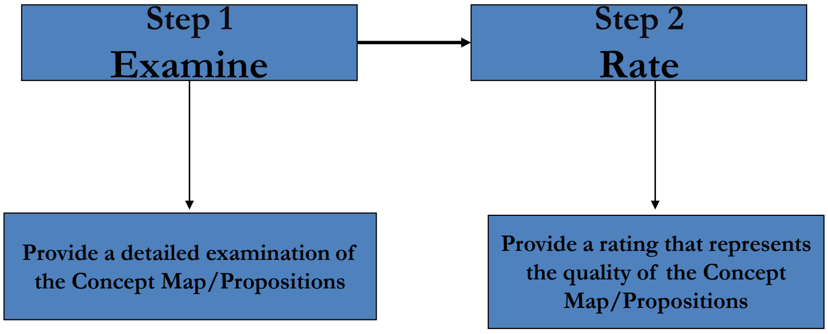


Whole Concept Map

Procedures for Evaluating Concept Maps

Procedures for Evaluating Concept Maps

- This training will present a two step procedure for evaluating Concept Maps
- When using this procedure you must consider the quality of specific dimensions within the Concept Maps



Step 1
Examine

Step 2
Rate

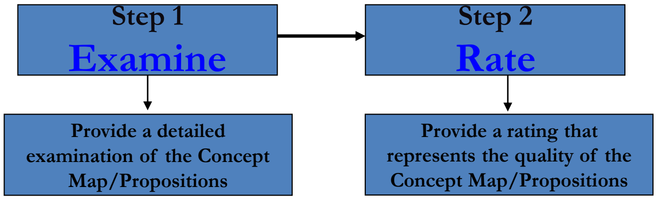
Provide a detailed examination of the Concept Map/Propositions

Provide a rating that represents the quality of the Concept Map/Propositions

Procedures for Evaluating Concept Maps

- Concept Map Dimensions
  - The dimensions that will be evaluated include Accuracy, Relevancy, and Density
  - In the following slides, you will learn how to apply the evaluation procedure to, first, the Proposition evaluations, then, the Whole Concept Map evaluations in terms of the dimensions

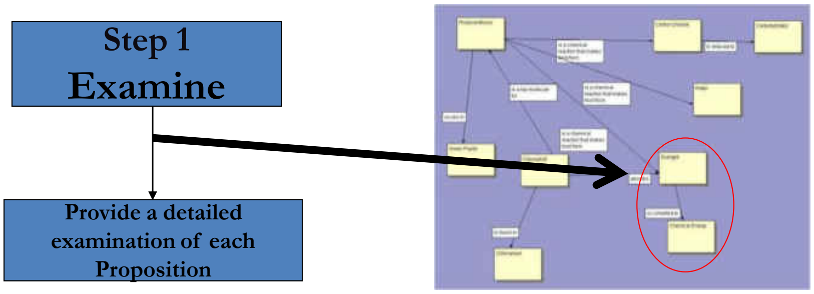
Accuracy Relevancy Density


Step 1
Examine
Provide a detailed examination of the Concept Map/Propositions
Step 2
Rate
Provide a rating that represents the quality of the Concept Map/Propositions

Proposition Evaluation

Examining Propositions

- Step 1: Examine
  - Step 1 requires an examination of the Propositions in terms of the dimensions: Accuracy and Relevancy
  - Therefore, the first step in the Proposition evaluation is to Examine, in detail, the Accuracy and Relevancy of each individual Proposition



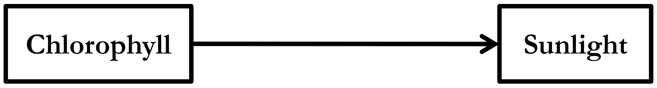
Step 1
Examine

Provide a detailed examination of each Proposition

Examining Propositions

- Examining the Accuracy of the Proposition
  - When examining the Accuracy of a Proposition, first consider whether two connected Concepts share a relationship
  - Ask Yourself: *Do these Concepts share a relationship?*

- For Example,



Chlorophyll

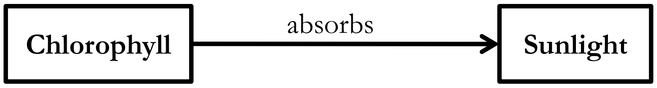
Sunlight

  - The answer should be yes

Examining Propositions

- Examining the Relevancy of the Proposition
  - Since the Concepts share a relationship you must then consider whether the label between the Concepts explains a relationship that is relevant to the topic (i.e., Photosynthesis)
  - Ask Yourself: *Is the relationship relevant to Photosynthesis?*

- For Example,



Chlorophyll absorbs **Sunlight**

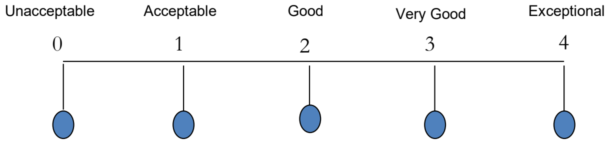
  - The answer should be yes, therefore you would proceed to Step 2, rating the quality of the relationship

Rating Propositions

Rating Propositions

- Step 2: Rating
  - Once you have Examined a Proposition for Accuracy and Relevancy, you will then assign a rating that represents the quality of that Proposition

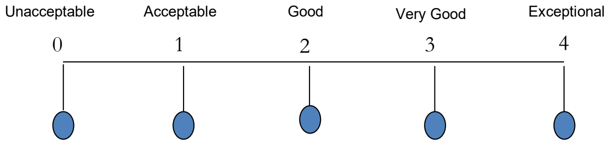
- Rating Scale
  - To rate the individual Propositions you will use the same rating scale as before



Rating Propositions

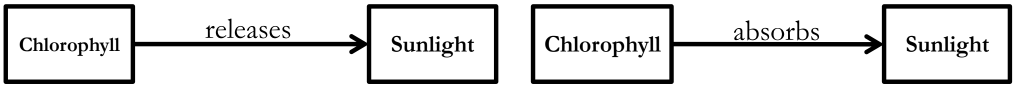
- Rating for Accuracy and Relevancy
  - Keeping your examination from Step 1 in mind, provide a rating that represents both the Accuracy of the Proposition and the Relevancy of the Proposition to the domain

  - In the following slides you will learn how to consider both Accuracy and Relevancy when assigning ratings



Rating Propositions

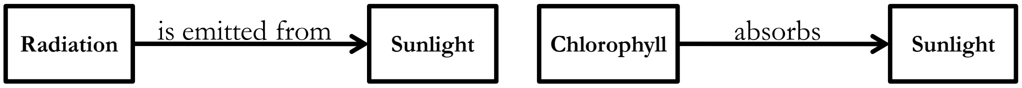
- Accuracy
  - Consider the Propositions,



  - The first Proposition should be rated a “0” given that it is an inaccurate statement
  - Because the second statement is accurate, the rating may therefore be a “1” or above (up to “4”) depending on its level of Accuracy and Relevancy

Rating Propositions

- Relevancy
  - Consider the Propositions,

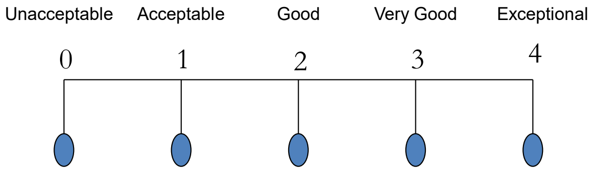


  - The first Proposition should be rated a “0” given that it is not relevant to photosynthesis
    - In other words, the relationship does not help define photosynthesis
  - The second statement is relevant, therefore a rating of a “1” or above (up to “4”) may be assigned depending on its level of Relevancy

Rating Propositions

- Rating Propositions

  – Once you have determined that a Proposition is both accurate and relevant, then you must consider what rating represents the Proposition’s appropriate level of Accuracy and Relevancy



  – The following slides provides examples of what each point on the rating scale may represent

Rating Concept Map Proposition Scores

– A rating of “4” may indicate that the Proposition is both accurate and relevant and the best possible explanation of the relationship between the Concepts

– A rating of “3” may indicate that the Proposition is both accurate and relevant and above average but not the best possible explanation

– A rating of “2” may indicate that the Proposition is both accurate and relevant and of average quality

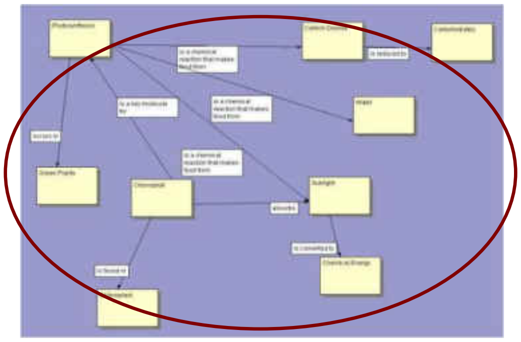
– A rating of “1” may indicate that the Proposition is both accurate and relevant and of below average quality

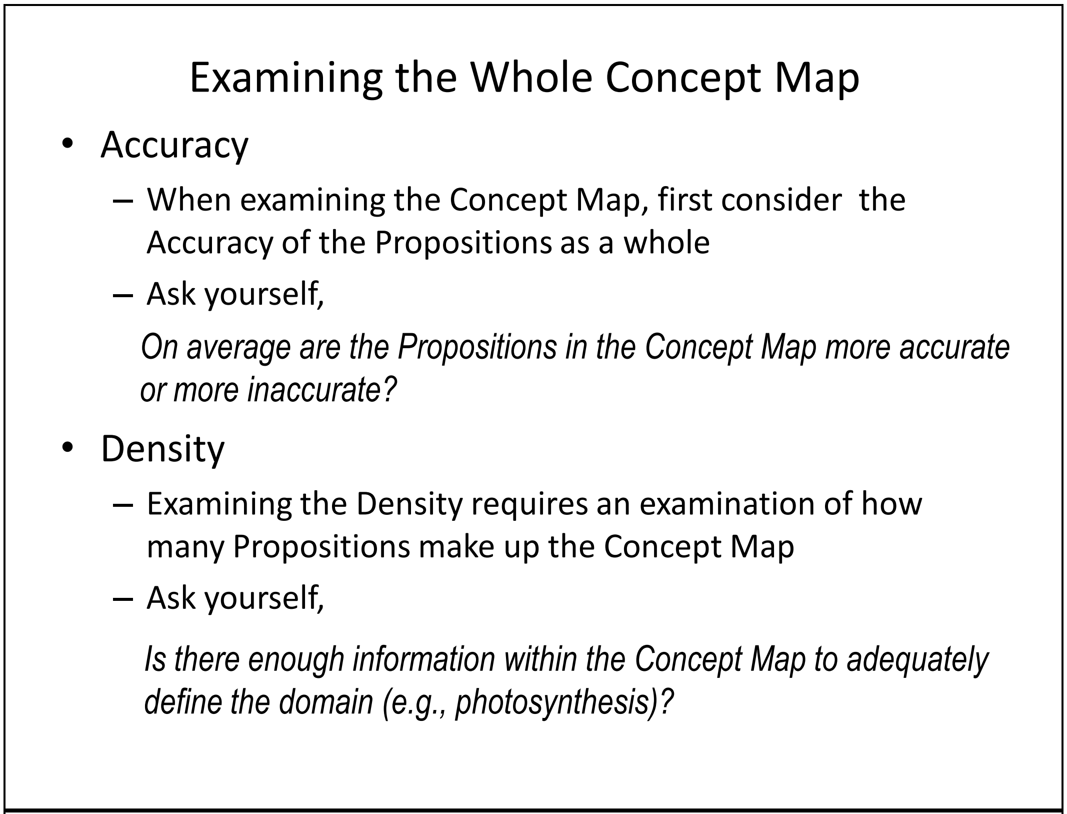
– A rating of “0” may indicate that the Proposition is either inaccurate and/or irrelevant

Examining the Whole Concept Map

Examining the Whole Concept Map

- Examining the Concept Map
  - To examine the Concept Map as a whole you must consider the Accuracy and Density of the entire Concept Map
  - In other Words, you must provide a detailed examination of the Accuracy of the Concept Map and the Density of the information within the Map



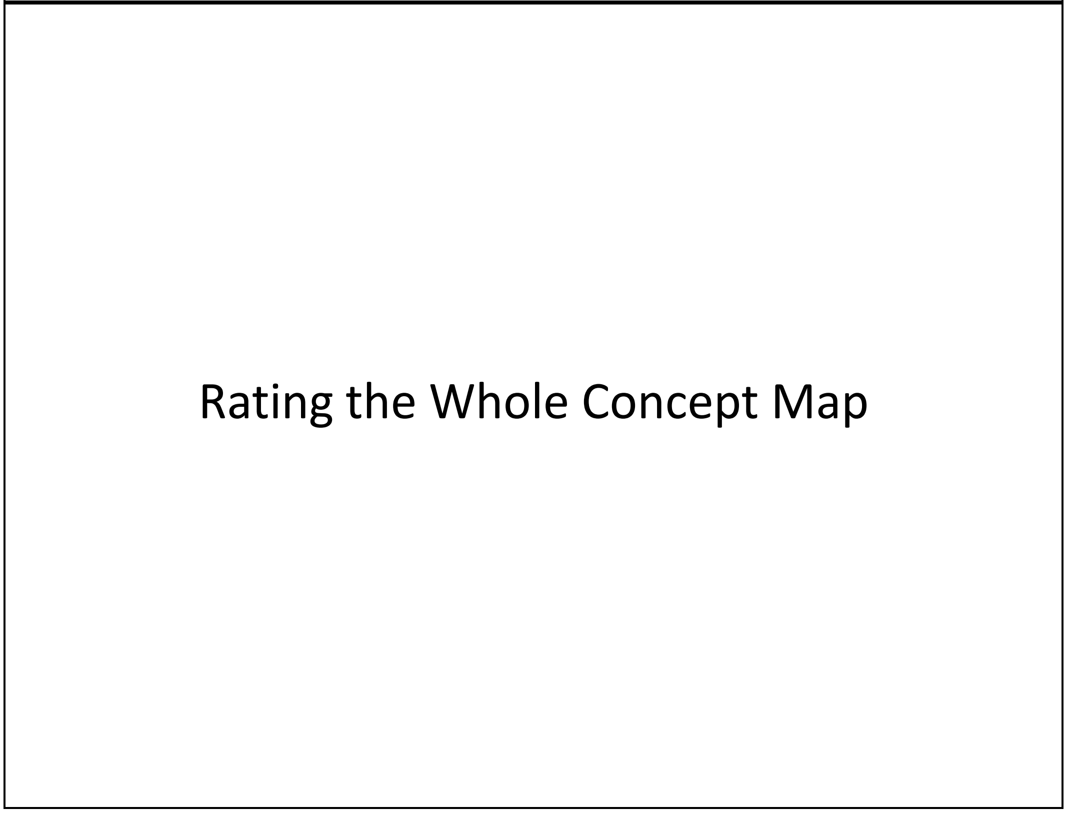


Examining the Whole Concept Map

- Accuracy
  - When examining the Concept Map, first consider the Accuracy of the Propositions as a whole
  - Ask yourself,

    On average are the Propositions in the Concept Map more accurate or more inaccurate?
- Density
  - Examining the Density requires an examination of how many Propositions make up the Concept Map
  - Ask yourself,

    Is there enough information within the Concept Map to adequately define the domain (e.g., photosynthesis)?

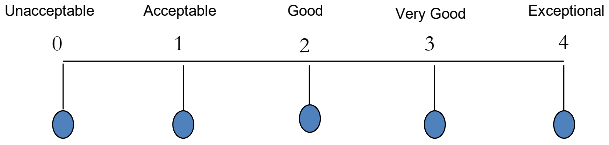


Rating the Whole Concept Map

Rating the Whole Concept Map

- Step 2: Rating
  - Once you have Examined the Concept Map for Accuracy and Density, you will then assign a rating that represents the quality of the Concept Map

- Rating Scale
  - To rate the Concept Map you will use the same rating scale as before



Rating the Whole Concept Map

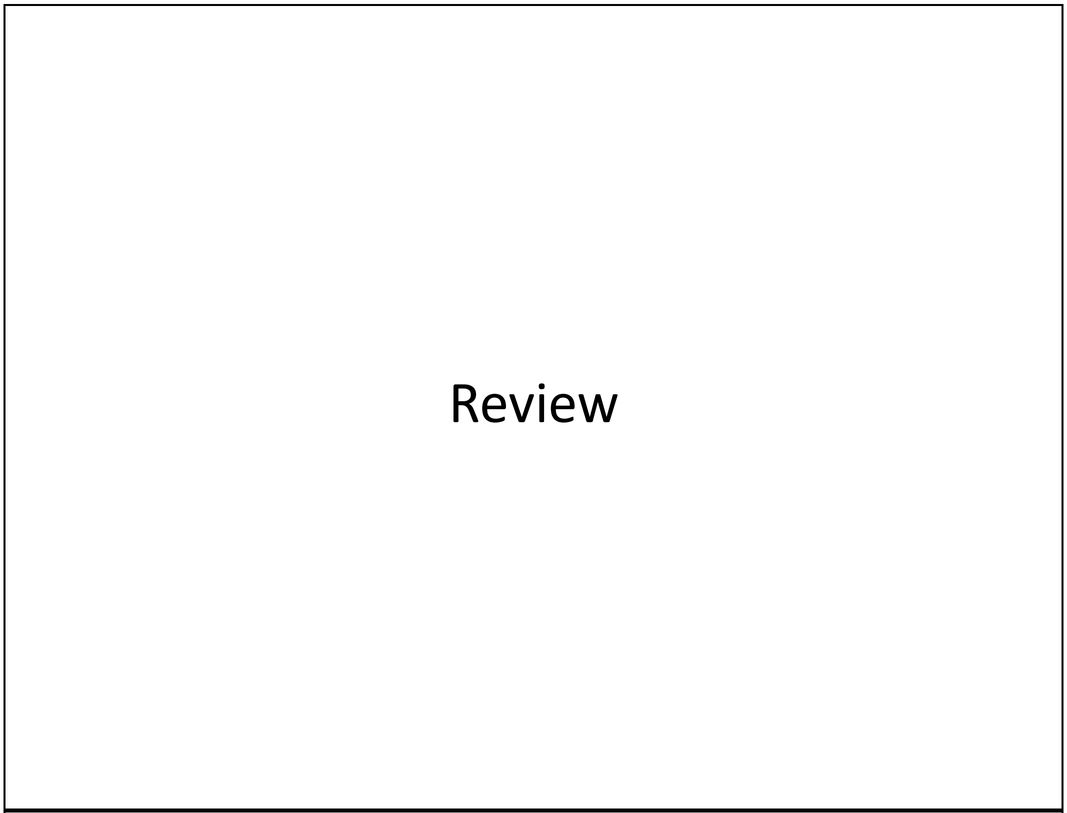
- A rating of “4” may indicate that the Concept Map contains exceptional Propositions and contains an exceptional amount of information pertaining to photosynthesis

- A rating of “3” may indicate that the Concept Map contains enough above average information to define photosynthesis

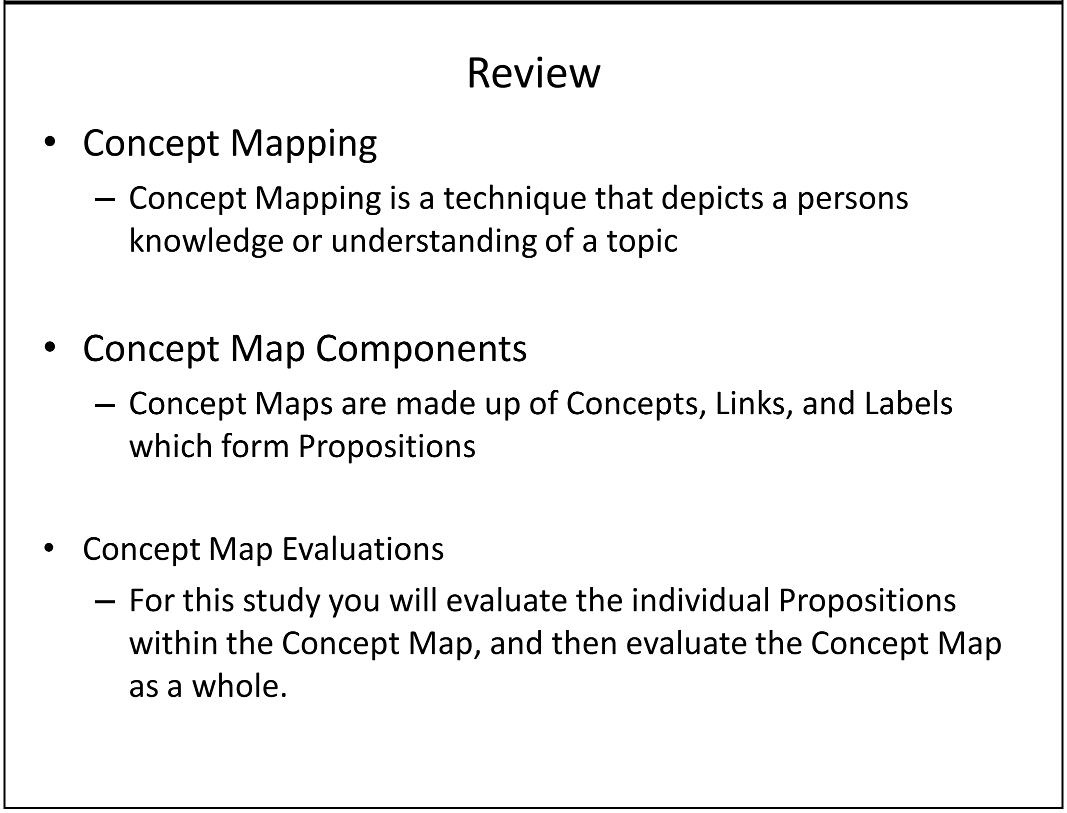
- A rating of “2” may indicate that the Concept Map contains enough information of average quality to define photosynthesis

- A rating of “1” may indicate that the Concept Map contains very little accurate information pertaining to photosynthesis

- A rating of “0” may indicate that the Concept Map has no accurate information pertaining to photosynthesis

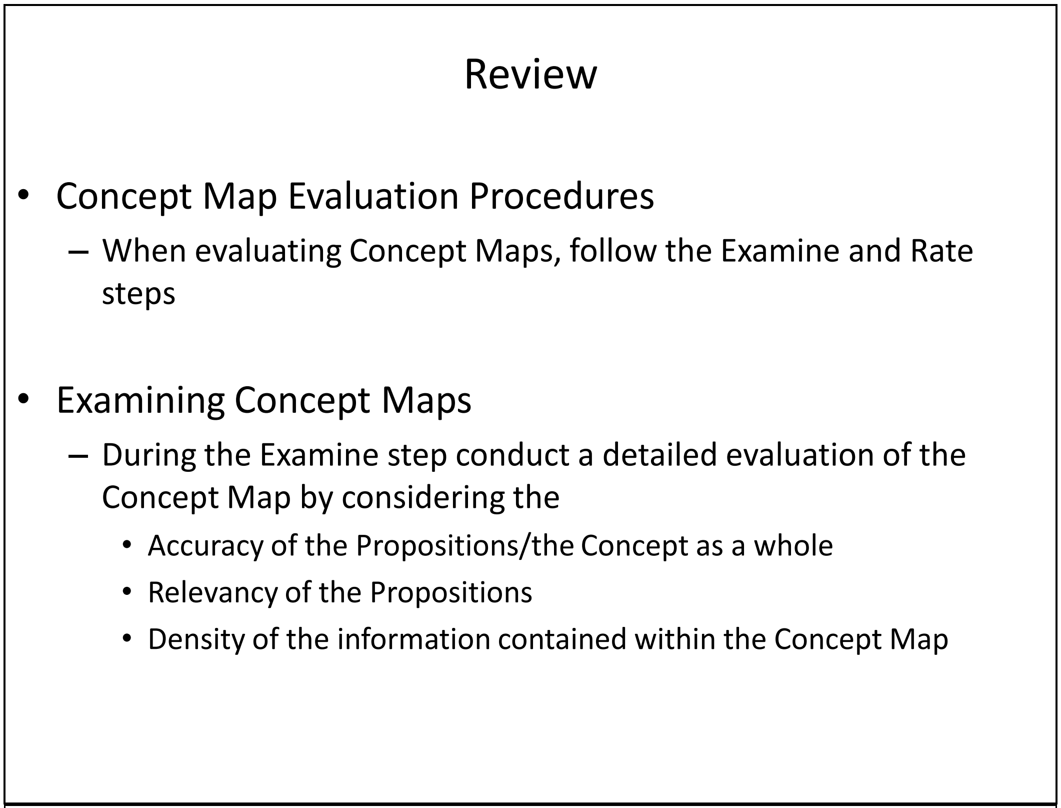


Review



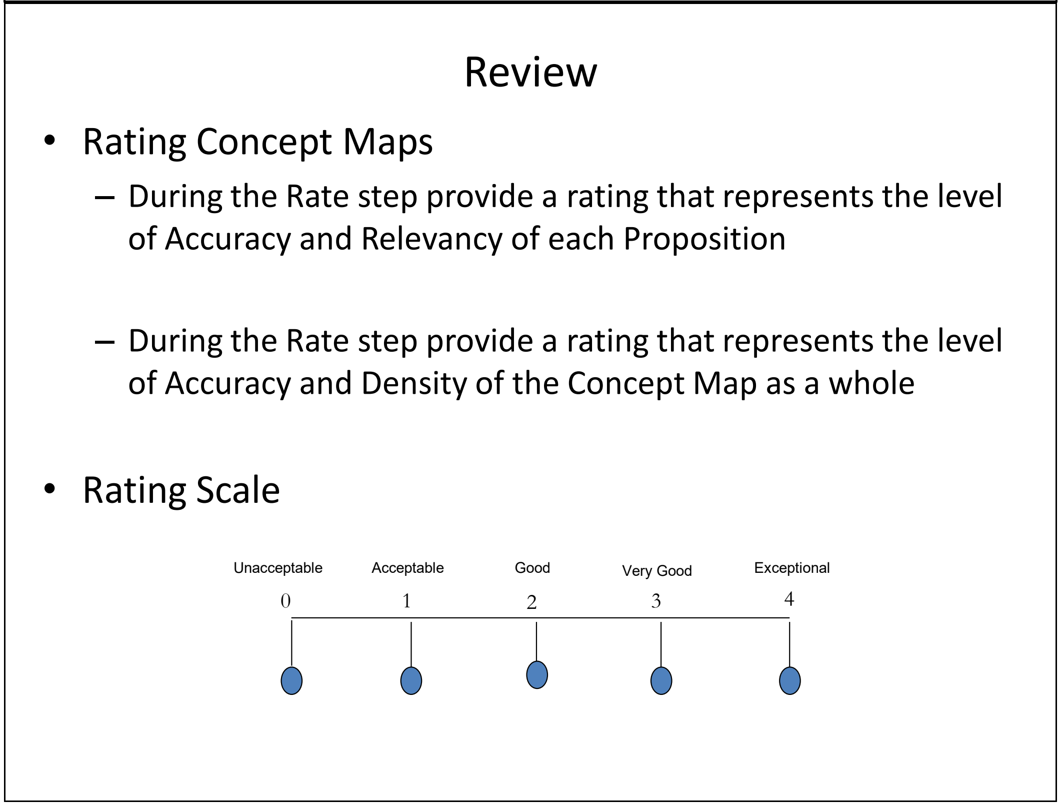
Review

- Concept Mapping
  - Concept Mapping is a technique that depicts a persons knowledge or understanding of a topic
- Concept Map Components
  - Concept Maps are made up of Concepts, Links, and Labels which form Propositions
- Concept Map Evaluations
  - For this study you will evaluate the individual Propositions within the Concept Map, and then evaluate the Concept Map as a whole.



Review

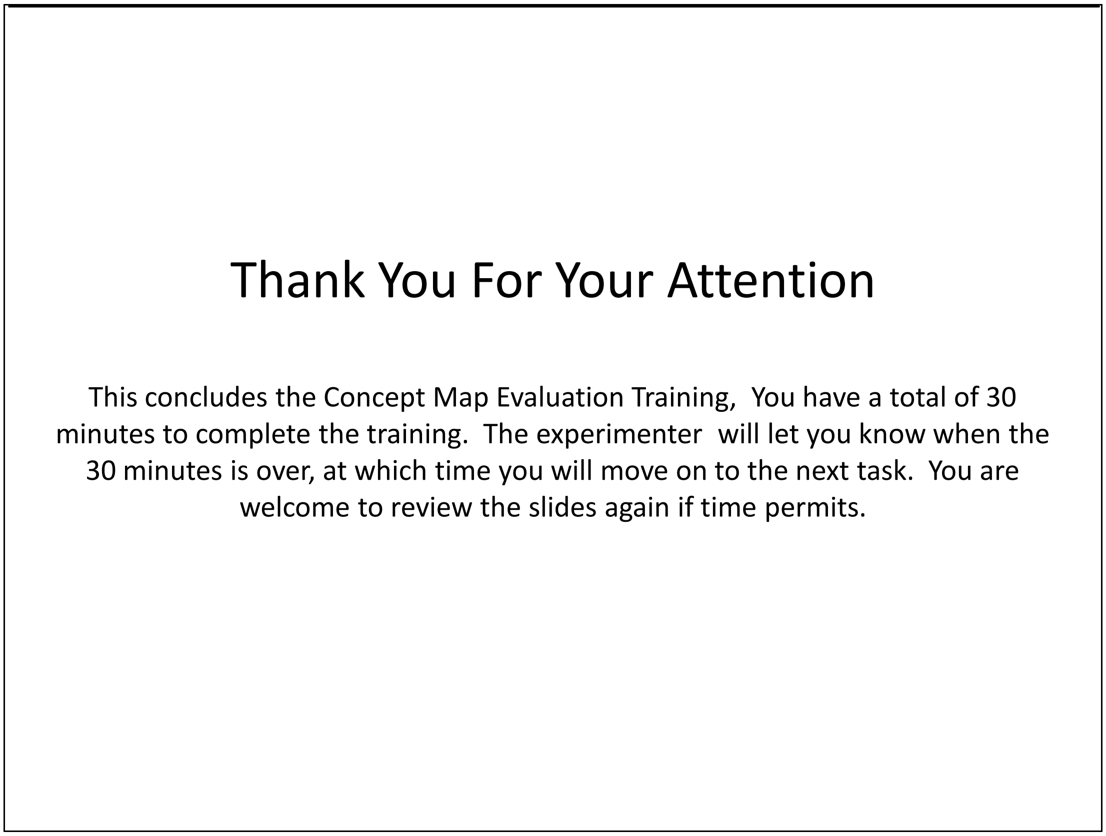
- Concept Map Evaluation Procedures
  - When evaluating Concept Maps, follow the Examine and Rate steps
- Examining Concept Maps
  - During the Examine step conduct a detailed evaluation of the Concept Map by considering the
    - Accuracy of the Propositions/the Concept as a whole
    - Relevancy of the Propositions
    - Density of the information contained within the Concept Map



Review

- Rating Concept Maps
  - During the Rate step provide a rating that represents the level of Accuracy and Relevancy of each Proposition
  - During the Rate step provide a rating that represents the level of Accuracy and Density of the Concept Map as a whole
- Rating Scale

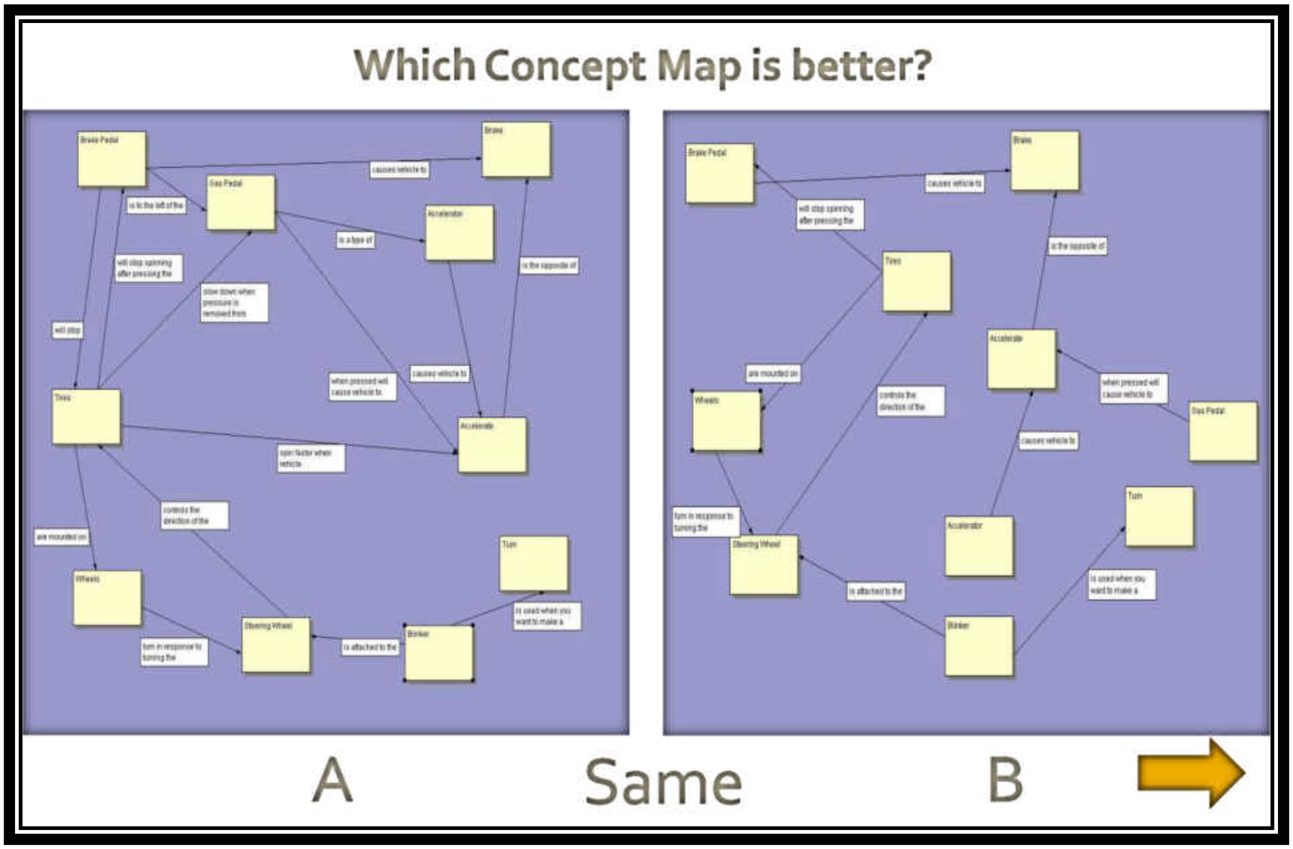
Unacceptable
0
Acceptable
1
Good
2
Very Good
3
Exceptional
4



Thank You For Your Attention

This concludes the Concept Map Evaluation Training, You have a total of 30 minutes to complete the training. The experimenter will let you know when the 30 minutes is over, at which time you will move on to the next task. You are welcome to review the slides again if time permits.

APPENDIX J: DICRIMINATION TASK



Which Concept Map is better?

A Same B

REFERENCES

Acton, W. H., Johnson, P. J., & Goldsmith, T. E. (1994). Structural knowledge assessment: Comparison of referent structures. *Journal of Educational Psychology, 86*(2), 303.

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

Athey, T. R., & McIntyre, R. M. (1987). Effect of rater training on rater accuracy: Levels-of-processing theory and social facilitation theory perspectives. *Journal of Applied Psychology, 72*(4), 567.

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences, 4*(11), 417.

Bedard, J., & Chi, M. T. (1992). Expertise. *Current Directions in Psychological Science, 1*(4), 135.

Bernardin, H. J. (1978). Effects of rater training on leniency and halo errors in student ratings of instructors. *Journal of Applied Psychology, 63*(3), 301-308.

Bernardin, H. J., Bernardin, H. J., & Walter, C. S. (1977). Effects of rater training and diary-keeping on psychometric error in ratings. *Journal of Applied Psychology, 62*(1), 64.

Boehm-Davis, D. A. (1989). Knowledge elicitation and representation. In *Human performance models for computer-aided engineering* (pp. 291-298): National Academy Press.

Borman, W. C. (1978). Explaining the Upper Limits of Reliability and Validity in Performance Ratings, *Journal of Applied Psychology* (Vol. 63, pp. 135-144).

Cannon-Bowers, J. A., Salas, E., & Converse, S. (1993). Shared mental models in expert team decision making. 221.

Carter, L., Haythorn, W., Meirowitz, B., & Lanzetta, J. (1951). The relation of categorizations and ratings in the observation of group behavior. *Human Relations, 4*, 239.

Cheatham, D. B., & Lane, S. C. (2002). Differential Access Hypothesis: The Effects of Task and Information Type on the Validity of Knowledge Acquisition Methods. *Human Factors and Ergonomics Society Annual Meeting Proceedings, 46*, 487-491.

Collins, A., & Gentner, D. (1987). How people construct mental models. In D. Holland & N. Quinn (Eds.), *Cultural models of Language and thought* (pp. 243-265). Cambridge: Cambridge University Press.

Collins, A., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning & Verbal Behavior, 8*(2), 240.

Cooke, N. (1994). Varieties of knowledge elicitation techniques. *International Journal of Human-Computer Studies, 41*(6), 801.

Cooke, N. (1999). Knowledge Elicitation. In F. T. Durso (Ed.), *Handbook of Appllied Cognition*. Chichester: Wiley.

Cooke, N., & McDonald, J. E. (1987). The application of psychological scaling techniques to knowledge elicitation for knowledge-based systems. *International Journal of Man-Machine Studies, 26*(4), 533.

Cooper, W. H. (1981). Ubiquitous halo. *Psychological Bulletin, 90*(2), 218.

Curtis, M., Harper-Sciarini, M. E., Jentsch, F., Schuster, D., & Swanson, R. (2007). *Filling the gaps: An Investigation of the knowledge needed for effective human-automation* *interaction.* Paper presented at the Proceedings of the International Symposium on Aviation Psychology,, Dayton, Ohio.

Davis, M. A., Curtis, M. B., & Tschetter, J. D. (2003). Evaluating cognitive training outcomes: Validity and utility of structural knowledge assessment. *Journal of Business and Psychology, 18*(2), 191.

Day, E. A., Arthur, W., Jr., & Gettman, D. (2001). Knowledge structures and the acquisition of a complex skill. *Journal of Applied Psychology, 86*(5), 1022-1033.

Deese, J. (1961). From the Isolated Verbal Unit to Connected Discourse. In C. N. Cofer (Ed.), *Verbal learning and verbal behavior* (pp. 11). New York: McGraw-Hill.

Dennis, I. (2007). Halo effects in grading student projects. *Journal of Applied Psychology, 92*(4), 1169.

Dismukes, R. K., Berman, B., A., & Loukopoulos, L., D. (2007). *The Limits of Expertise: Rethinking Pilot Error and the Causes of Airline Accidents*. Burlington: Ashgate.

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.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25*(2), 155.

Engelhard, G. (1994). Examing rater errors in the assessment of written composition with a many-faceted Rasch Model. *Journal of Educational Measurement, 31*(2), 93-112.

Engelhard, G., Jr. (1994). Examining Rater Errors in the Assessment of Written Composition with a Many-Faceted Rasch Model. *Journal of Educational Measurement, 31*(2), 93-112.

Evans, A. W., Jentsch, F., Hitt, J. M., & Bowers, C. (2001). Mental Model Assessments: Is There Convergence Among Different Methods. *Human Factors and Ergonomics Society Annual Meeting Proceedings, 45*, 293-296.

Fisher, R. P., & Craik, F. I. (1977). Interaction between encoding and retrieval operations in cued recall. *Journal of Experimental Psychology: Human Learning and Memory, 3*(6), 701.

Glaser, R., & Bassok, M. (1989). Learning theory and the study of instruction. *Annual Review of Psychology, 40*, 631.

Goldsmith, T. E., Johnson, P. J., & Acton, W. H. (1991). Assessing structural knowledge. *Journal of Educational Psychology, 83*(1), 88-96.

Goldsmith, T. E., & Kraiger, K. (1997). Application of Structual Knowledge Assessment to Training Evaluations. In S. K. Kevin Ford, Kurt Kraiger, Eduardo Salas, Teachout (Ed.), *Improving Training Effectiveness in Work Organizations* (pp. 73-98). Mahwah: Lawrence Erlbaum Associates, Inc.

Halderman, J. D., & Mitchell, C. D. (2004). *ASE Test Preparation Series Steering and Suspension*. Uppersaddle, NJ: Prentice Hall.

Harper, Hoeft, R. M., Jentsch, F., & Boehm-Davis, D. A. (2005). *Predicting performance using concept mapping.* Paper presented at the American Psychological Association, Washington, D.C.

Harper, M., Evans, A. W., Hoeft, R., & Jentsch, F. (2004). *Practically scoring concept maps for the assessment of trainee's knowledge structures.* Paper presented at the Human factors and Ergonomics Society 48th Annual Meeting, New Orleans, Louisiana.

Harper, M. E., Hoeft, R. M., Evans, A. W., III, & Jentsch, F. (2004). *Practically scoring concept maps for the assessment of trainee's knowledge structures.* Paper presented at the Human factors and Ergonomics Society 48th Annual Meeting, New Orleans, Louisiana.

Harper, M. E., Schuster, D., Hoeft, R. M., & Jentsch, F. (2008). Scoring Concept Maps. University of Central Florida.

Hoeft, R. M., Jentsch, F. G., Harper, M. E., Evans, A. W., Bowers, C. A., & Salas, E. (2003). TPL-KATS--concept map: a computerized knowledge assessment tool. *Computers in Human Behavior, 19*(6), 653-657.

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.

Johnson-Laird, P. N. (1980). Mental models in cognitive science. *Cognitive Science: A Multidisciplinary Journal, 4*(1), 71-115.

Johnson, D. M., & O'Reilly, C. A. (1964). Concept attainment in children: Classifying and defining. *Journal of Educational Psychology, 55*(2), 71.

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

Kavanagh, M. J., MacKinney, A. C., & Wolins, L. (1971). Issues in managerial performance: Multitrait-multimethod analyses of ratings. *Psychological Bulletin, 75*(1), 34.

Klein, G. A., Calderwood, R., & MacGregor, D. (1989). Critical decision method for eliciting knowledge. *IEEE Transactions on Systems, Man, & Cybernetics, 19*(3), 462.

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*(2), 311-328.

Landy, F., & Farr, J. (1980). Performance Ratings. *Psychological Bulletin, 87*(1), 72-107.

Lievens, F. (2001). Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. *Journal of Applied Psychology, 86*(2), 255-264.

Lievens, F., & Sanchez, J. I. (2007). Can training improve the quality of inferences made by raters in competency modeling? A quasi-experiment. *Journal of Applied Psychology, 92*(3), 812.

Mathieu, J. E., Heffner, T. S., Goodwin, G. F., Salas, E., & Cannon-Bowers, J. A. (2000). The influence of shared mental models on team process and performance. *Journal of Applied Psychology, 85*(2), 273-283.

McClure, J. R., Sonak, B., & Suen, H. K. (1999). Concept map assessment of classroom learning: Reliability, validity, and logistical practicality. *Journal of Research in Science Teaching, 36*(4), 475.

Novak, J., Gowen, B., & Johansen, G. (1983). The use of concept mapping and knowldge vee mapping with junior high school science students. *Science Education, 67*(5), 625-645.

Novak, J. D. (1995). Concept mapping: A strategy for organizing knowledge. 229.

Novak, J. D., & Canas, A. (2006). *The Theory Underlying Concept Maps and How to Construct and Use Them* (No. IHMC CmapTools 2006-01). Pensacola, FL: Florida Institute for Human Machine Cogntiono. Document Number)

Novak, J. D., & Gowen, B. (1984). *Learning how to learn*. New York: Cambridge University Press.

Plummer, K. (2008). *Analysis of the psychometric properties of two different concept-map assessment tasks.* Unpublished Dissertation, Brigham Young University.

Rouse, W. B., & Morris, N. M. (1986). On looking into the black box: Prospects and limits in the search for mental models. *Psychological Bulletin, 100*(3), 349.

Rowe, A. L., & Cooke, N. J. (1995). Measuring mental models: Choosing the right tools for the job. *Human Resource Development Quarterly, 6*(3), 243-255.

Rowe, A. L., Cooke, 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.

Ruiz-Primo, M. A. (2004). *Examining Concepts Maps As An Assessment Tool.* Paper presented at the First International Conference on Concept Mapping, Pamplona, Spain.

Ruiz-Primo, M. A., Schultz, S. E., Li, M., & Shavelson, R. J. (2001). Comparison of the reliability and validity of scores from two concept-mapping techniques. *Journal of Research in Science Teaching, 38*(2), 260-278.

Ruiz-Primo, M. A., & Shavelson, R. J. (1996). Problems and issues in the use of concept maps in science assessment. *Journal of Research in Science Teaching, 33*(6), 569.

Ruiz-Primo, M. A., Shavelson, R. J., Li, M., & Schultz, S. E. (2001). On the Validity of Cognitive Interpretations of Scores From Alternative Concept-Mapping Techniques. *Educational Assessment, 7*(2), 99-141.

Ruiz-Primo, M. A., Shavelson, R. J., & Schultz, S. (1997). *On the Validity of Concept Map-Base Assessment Interpretation: An Experiment Testing the Assumption of Hierarchical Concept Maps in Science* (No. CSE Technical Report 455): Stanford Universityo. Document Number)

Saal, F. E., Downey, R. G., & Lahey, M. A. (1980). Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin, 88*(2), 413.

Schleicher, D. J., Schleicher, D. J., & Day, D. V. (1998). A cognitive evaluation of frame-of-reference rater training: Content and process issues. *Organizational Behavior and Human Decision Processes, 73*(1), 76.

Schleicher, D. J., Schleicher, D. J., Day, D. V., Mayes, B. T., & Riggio, R. E. (2002). A new frame for frame-of-reference training: Enhancing the construct validity of assessment centers. *Journal of Applied Psychology, 87*(4), 735.

Schvaneveldt, R. (1990). Proximities, networks, and schemata. In *Pathfinder associative networks: studies in knowledge organization* (pp. 135-148): Ablex Publishing Corp.

Shavelson, R. J. (1972). Some aspects of the correspondence between content structure and cognitive structure in physics instruction. *Journal of Educational Psychology, 63*(3), 225.

Shavelson, R. J., Ruiz-Primo, M. A., & Wiley, E. (2005). Window into the mind. *Higher Education, 49*(4), 413-430.

Smith, C. F., & Boehm-Davis, D. A. (2008). The effect of functional display use on novice performance and knowledge acquisition. Unpublished Journal Article. George Mason University.

Smith, C. F., Boehm-Davis, D. A., & Fadden, S. (2008). The use of functional displays to improve pilot performance. Unpublished Journal Article. George Mason University.

Stayanov, S., & Kirschner, P. (2004). Expert Concept Mapping Method for Defining the Characteristics of Adaptive E-Learning: ALFANET Project Case. *Educational Technology Research & Development, 52*(2), 41-56.

Sulsky, L. M., & Day, D. V. (1994). Effects of frame-of-reference training on rater accuracy under alternative time delays. *Journal of Applied Psychology, 79*(4), 535.

Tennyson, R. D., & Cocchiarella, M. J. (1986). An empirically based instructional design theory for teaching concepts. *Review of Educational Research, 56*(1), 40.

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4*(1), 25.

Tulving, E., & Pearlstone, Z. (1966). Availability versus accessibility of information in memory for words. *Journal of Verbal Learning & Verbal Behavior, 5*(4), 381.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review, 80*(5), 352.

Uggerslev, K. L., & Sulsky, L. M. (2008). Using frame-of-reference training to understand the implications of rater idiosyncrasy for rating accuracy. *Journal of Applied Psychology, 93*(3), 711.

Viswesvaran, C., Schmidt, F. L., & Ones, D. S. (2005). Is There a General Factor in Ratings of Job Performance? A Meta-Analytic Framework for Disentangling Substantive and Error Influences. *Journal of Applied Psychology, 90*(1), 108-131.

Weekley, J. A., & Gier, J. A. (1989). Ceilings in the reliability and validity of performance ratings: The case of expert raters. *Academy of Management Journal, 32*(1), 213.

West, D. C., Pomeroy, J. R., Park, J. K., Gerstenberger, E. A., & Sandoval, J. (2000). Critical Thinking in Graduate Medical Education: A Role for Concept Mapping Assessment? *JAMA, 284*(9), 1105-1110.

Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology, 67*(3), 189.

Yin, Y., Vanides, J., Ruiz-Primo, M. A., Ayala, C. C., & Shavelson, R. J. (2005). Comparison of two concept-mapping techniques: Implications for scoring, interpretation, and use. *Journal of Research in Science Teaching, 42*(2), 166-184.