MENTAL MODELS AND THE ACQUISITION OF A COMPLEX SKILL ACROSS INDIVIDUALS AND TEAMS: A MULTILEVEL STUDY

A Dissertation

by

GONZALO JAVIER MUNOZ GALVEZ

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Winfred Arthur, Jr.
Committee Members,	Christopher M. Berry
	Stephanie C. Payne
	Victor Willson
Head of Department,	Douglas W. Woods

May 2014

Major Subject: Psychology

ABSTRACT

A mental model reflects the structural relationships between concepts within a specified knowledge domain. Measuring the structure of knowledge is important because it offers the possibility of capturing expert knowledge which is often difficult to assess using traditional declarative knowledge measures. The concept of mental model has been extensively studied over the last decades and it is often acknowledged in the training literature as one of the key antecedents of performance in complex tasks, particularly in the context of teams where the construct of shared mental models has received ample attention. Whereas the training literature has established the validity of mental models for predicting individual and team performance using single-level studies, the extant literature has not yet tested the validity of mental models as a multilevel construct. Consequently, the purpose of the present study was to assess the extent to which the relationships between mental models and performance generalizes across individuals and teams, that is to test a homologous multilevel model.

Participants in this study completed a dynamic, networked computer-based simulation. Three-person teams operated the simulator collectively (through specialized roles) and as individuals (performing all roles simultaneously) over the course of a 2-day 48-hour-interval protocol. The sample consisted of 243 individuals nested in 81 3-person teams. Consistent with the multilevel nature of the problem under study, multilevel analyses were conducted to test the study hypotheses.

Consistent with theory and previous research on individual and team cognition, it was hypothesized that stronger relationships between mental models and performance would exist at the individual level compared to the team level. In essence, processes occurring at the team level were expected to attenuate the relationship between mental models and performance compared to the individual level of analysis. Contrary to this expectation, the magnitude of the relationship between mental models and performance was similar across levels of analysis. Additionally, consistent with previous research on the effectiveness of declarative knowledge measures for predicting complex performance, the present results indicated that declarative knowledge was more predictive of individual performance than team performance.

In addition to performance, an objective measure of behaviors was utilized to further understand of the processes through which mental models translate into effective individual and team performance. It was hypothesized that the relationship between mental models and behaviors would be stronger for individual tasks than team tasks as a function of the additional interaction requirements associated with team tasks. However, contrary to this expectations, mental models and behaviors were more strongly associated at the team level than the individual level.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, Dr. Winfred Arthur Jr. for his continuous support and guidance during the completion of my doctoral studies. Dr. Arthur’s dedication to research and his unyielding commitment to quality have inspired me to want to be a better scholar. I would also like to thank my committee members, Dr. Berry, Dr. Payne, and Dr. Willson, for their thoughtful comments and for helping to create a supporting environment for me to develop my ideas.

The data used for the present study are part of a larger data collection effort undertaken by Dr. Arthur’s research lab. The amount of time and effort put forth on this project was immense and I cannot thank enough all of the graduate and undergraduate students who made this study possible. Particularly, I would like to thank all of the graduate students who were involved in the development and implementation of the present study: Andrew Naber, Jennifer McDonald, Steven Jarrett, Ryan Glaze, Ira Schurig, Olabisi Asenuga, Craig White, Nate Keiser, and Inchul Cho.

I would not have been able to reach this point without the support of the Dean of the Psychology Department of University Adolfo Ibáñez, Dr. Jorge Sanhueza. Since I joined the psychology department Dr. Sanhueza has helped me to grow professionally in numerous ways. Importantly, he encouraged me to pursue a doctoral degree and played a pivotal part in marshalling the critical resources to complete this program. For that I will always be in debt to Dr. Sanhueza.

It is hard to imagine how this journey would have been possible without my wife, Tamara. She was always there for me and stood by me through the good and the bad times. In addition to her company and moral support, she was always willing to listen to my ideas and help me work through the many issues that I encountered while conducting my research. In my opinion, she should be entitled to an honorary degree in Industrial and Organizational Psychology, at the very least.

I would like to thank the Fulbright Program for awarding the grant that sponsored my graduate studies in the United States and the financial support of the Ministry of Education of Chile through its program “MECE Educación Superior 2”. Lastly, I would like to thank University Adolfo Ibáñez for its financial support and for facilitating this process through its financial department.

TABLE OF CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

"Intuition is nothing more and nothing less than recognition"
-Herbert Simon

A mental model is a higher-order cognitive construct for representing the global relations among relevant concepts within a domain. In the context of complex skills acquisition, experts' facility to articulate complex responses in a seemingly effortless manner has been linked to their ability to recognize meaningful patterns in their knowledge domain which would indicate a superior organization of the knowledge base (Glaser & Chi, 1988), that is, a superior mental model. In an attempt to reconcile some formal definitions of the construct, Rouse and Morris (1986) defined mental models as "the mechanisms whereby humans are able to generate descriptions of system purpose and form, explanations of system functioning and observed system states, and predictions of future system states" (p. 351). Advancements from the fields of cognitive psychology, computer science, and artificial intelligence have contributed to the development of psychological scaling algorithms to represent knowledge from data collected from human subjects (Schvaneveldt, 1990; Schvaneveldt, Dearholt, & Durso, 1988). The extant literature has established the validity of mental models for predicting performance on a variety of individual and team tasks, which suggests the presence of a multilevel construct—that is, a construct that is meaningful across multiple levels of analysis (Chen, Mathieu, & Bliese, 2005; Kozlowski & Klein, 2000).

The validity of several constructs that are meaningful across individual- and group-levels (e.g., teams, groups, organizations) has been demonstrated by employing *single-level* studies. Devine and Philips' (2001) meta-analysis demonstrated that team-level cognitive ability and team performance are positively correlated ($\rho = .29$), a finding that parallels the well-established relationship between cognitive ability and performance at the individual level (Hunter, 1986). Likewise, self- and team-efficacy have been found to exhibit positive correlations with individual ($\rho = .20$) and team performance ($\rho = .39$), respectively (Gully, Incalcaterra, Joshi, & Beaubien, 2002). In general, the impetus for scaling an individual-level construct to the group-level is that the specified construct is assumed to be homologous across levels—that is, the construct is said to operate similarly across levels of analysis. Yet, only a few studies have demonstrated the validity of constructs across levels of analysis using a multilevel framework that *directly* tests homologous models. Chen, Thomas, and Wallace (2005) tested a multilevel model relating training outcomes (team knowledge, collective efficacy, and team skills) to training transfer at both the individual and team levels. More recently, Beus, Muñoz, Arthur, and Payne (2013) tested a proportional theory of homology to describe safety climate's relationships with safety incidents and safety behaviors across construct levels. Specifically, Beus et al. found that safety climate demonstrated proportionately stronger associations with safety incidents at the workgroup level relative to the individual level. The point is that although the assumption of homology is rarely tested, "tests of homology can and should play an integral role in the validation of multilevel constructs and theories" (Chen, Bliese, & Mathieu, 2005, p. 378).

Direct empirical evidence supporting the multilevel validity of mental models is lacking. That is, although single-level research has been valuable in establishing the relationship between mental models and the acquisition of complex skills within individual and team contexts, a multilevel study of mental models is needed. One reason for undertaking such an effort is to more precisely compare the effectiveness of individual and team mental models as predictors of performance. To do so, it is critical to obtain an individual-level estimate of the mental model-performance relationship and evaluate the extent to which this estimate differs from its group-level counterpart. In addition, the appropriate research design to test a theory of homology involves having the same individuals performing a similar task both as individuals and teams. And because in this design individuals are nested in teams, in order to obtain unbiased estimates of the relationship between the constructs of interest at the individual level of analysis, the statistical procedures implemented must also account for the hierarchical structure of the data (e.g., hierarchical linear modeling). Thus, the main contribution of the proposed study is to test a multilevel theory of mental models—the extent to which the relationships between mental models and performance generalizes across individuals and teams—using current principles and analytical tools for testing homologous multilevel models (Chen, Bliese, & Mathieu, 2005).

Whereas the overall goal of training is to enhance organizational effectiveness, training interventions customarily focus on individual-level models and methods. This creates a levels paradox (Kozlowski, Brown, Weissbein, Cannon-Bowers, & Salas, 2000) or a problem of misspecification (Rousseau, 1985) that occurs when researchers draw conclusions at a higher-order level based on individual-level models, or fail to take into account the potential role of contextual factors on lower-level relationships. Kozlowski and Salas (1997) distinguished between transfer of training occurring across different settings at the same level (*horizontal transfer)* from upward transfer across different levels of the organization (*vertical transfer*). From a training perspective, a direct comparison of individual and team mental models is concerned with vertical transfer, or the link between training outcomes across levels of the organizational system. Consistent with the premise that there is a conceptually staggering difference between building an expert team versus developing a group of expert individuals, the present study contributes to the training literature by examining the difference between team and individual knowledge acquisition and their relationships with performance.

Lastly, previous studies have examined the relationship between team mental models and team processes. For instance, Mathieu, Heffner, Goodwin, Salas, and Cannon-Bowers (2000) demonstrated that the relationship between shared mental models and team performance was mediated by team processes. Furthermore, evidence has shown that the use of structural assessment techniques for measuring mental models is critical for predicting team processes (DeChurch & Mesmer-Magnus, 2010a). However, the manner in which team processes are measured is often based on subjective rather than objective measures of team processes. The use of raters, for instance, is not uncommon in this literature (e.g., Mathieu et al., 2000). In contrast, for the present study an objective measure of team behavioral processes was developed in which specific actions performed during the performance task (e.g., putting out a fire collectively using two distinct platforms) was recorded. These action sequences were then mapped on to specific conceptual links in the mental model networks. The use of an objective measure of behaviors is important because the overlap between mental models and behaviors indexes the extent to which mental models are “used” effectively. Furthermore, differences between individuals and teams in terms of their ability to implement their mental models speaks to the issue of the effective integration of cognitions during team performance (Cooke, Salas, Kiekel, & Bell, 2004). Consequently, yet another contribution of the present study is to bridge the gap between cognitions and their implications for action in individual and team contexts.

2. VALIDITY OF PSYCHOLOGICAL CONSTRUCTS ACROSS LEVELS OF ANALYSIS

Within the broader context of the unified concept of validity (Messick, 1995), construct validity represents the all-encompassing focus of validation efforts. Hence, different forms of evidence are subsumed as different *aspects* of construct validity. Construct validation can thus be seen as a continuing effort of appraising the meaningfulness of test scores, that is, the extent to which empirical evidence and theoretical rationales support the appropriateness of score interpretation and its implications for action. Specifically, Messick distinguished six aspects of construct validity comprising content, substantive, structural, generalizability, convergence and discriminant, and consequential validity.

Evidence related to the content, substantive, structural, generalizability, and consequential validity of mental models is presented here to highlight some methodological and theoretical issues related to the mental model construct, and as a background to contextualize the unique contribution of the present study.

Evidence in support of various aspects of the construct validity of mental models is available. From a content validation perspective, the development of mental model measures involves conducting a detailed job or task analysis which typically entails interviewing subject matter experts (SME). Because mental models are typically used to assess expertise in specific knowledge domains (e.g., avionics troubleshooting) and often in the context of synthetic tasks in laboratory settings, researchers who make use of the mental model construct must expend considerable effort constructing ad-hoc measures.

Mohammed, Klimoski, and Rentsch (2000) advocated for the use of multiple methods to adequately sample the underlying dimensions of the specified domain. Cooke and McDonald's (1986) study on driving mental models as well as Smith-Jentsch, Mathieu, and Kraiger's (2005) study with air traffic controllers, are excellent examples of the careful consideration involved in the selection of stimuli for mental model assessment.

The substantive aspect of construct validity refers to "theoretical rationales for the observed consistencies in test responses . . . along with empirical evidence that the theoretical processes are actually engaged by respondents in the assessment task" (Messick, 1995, p. 745). Rowe, Cooke, Hall, and Halgren (1996), for instance, had 19 Air Force technicians complete various system knowledge measures—a laddering interview, concept relatedness ratings, a diagramming task, and a think aloud task—and assessed their comparative effectiveness as predictors of the technicians' scores on a verbal troubleshooting task. In the troubleshooting task technicians were asked to describe the steps that should be taken to repair a fault that had occurred with the equipment. Mental model scores were based on the concept relatedness ratings which were analyzed via the Pathfinder algorithm (Schvaneveldt, 1990). Rowe et al.'s results indicated that participants' scores on the troubleshooting task were positively correlated to the quality of their mental models, demonstrating the substantive validity of mental models. That is, Rowe et al.'s study demonstrated that mental model scores were associated with the technicians' cognitive processes underlying their performance.

Messick (1995) postulated that structural validity (or structural fidelity; Loevinger, 1957) is demonstrated when the scoring model is consistent with the theory of the construct domain. The use of structural assessment techniques and scaling algorithms that capture the *configural* aspect of knowledge is aligned with the mental model construct. A study conducted by DeChurch and Mesmer-Magnus (2010a) demonstrated that "only methods that model the structure or organization of knowledge are predictive of [team] process" (p. 1). Although the way in which team mental models are operationalized differentially predicts team processes, differences in operationalization do not seem to relate to team performance. Somewhat similar results have been observed using individual performance as criteria. For example, Schuelke et al. (2009) examined the relative criterion-related validity of coherence (an index of the quality of mental models' internal organization) and two indices of mental model accuracy (closeness and correlation) for predicting individual performance. Although not acknowledged by Schuelke et al. one could argue that unlike coherence and closeness, a simple correlation between novice and expert ratings does not really assess the underlying *structure* of knowledge. In this sense, Schuelke et al.'s finding that the different indices (coherence, closeness, and correlation) were highly similar for predicting individual performance is not really at odds with an extensive body of literature in the team domain which suggests that different methods for representing knowledge can be used to predict performance but only structural assessment methods are effective for predicting processes.

Generalizability refers to the degree to which construct interpretation generalizes across tasks and contexts (Messick, 1995). Meta-analyses are typically employed to support the generalizability aspect of construct validity. The general purpose of meta-analytic studies (e.g., Hunter & Schmidt, 2004) is to determine the extent to which X-Y relationships can be generalized across samples and identify the conditions under which said relationships hold true. Although meta-analyses are useful for establishing the boundaries of score meaning across samples, a different approach is needed to demonstrate the generalizability of constructs *across levels*. In contrast to validity generalization studies which focus on generalizing empirical results across samples, multilevel tests of homology focus on generalization across levels (Chen, Bliese, & Mathieu, 2005). Specifically, homologous multilevel theories consider "whether processes and relationships among variables at one level (e.g., the individual) are consistent with analogous processes and relationships at another level (e.g., the team)" (Chen, Bliese, & Mathieu, 2005, p. 376). Whereas the accumulated evidence demonstrates that mental models are valid predictors of team processes and performance (DeChurch & Mesmer-Magnus, 2010b), the unique contribution of the present study is to evaluate the generalizability of mental models across levels of analysis.

Issues pertaining to the consequential validity of mental models are clearly underdeveloped in the scholarly literature. One reason for this lack of attention may be that structural assessment techniques are rarely included in high-stakes settings in which issues of test bias and fairness are more salient. At first glance, the failure of structural assessment techniques to enter the applied world seems at odds with mounting evidence suggesting that mental models are valid predictors of performance. Research has shown, for instance, that computer programmers can be correctly classified as naïve, novice, intermediate, or expert, based on the quality of their mental models (Cooke & Rowe, 1993). Thus, it does not seem unreasonable to use mental models as an employment decision making tool (e.g., personnel selection).

In addition to potential technical barriers for conducting structural assessments, there is at least one explanation for this apparent disconnect between research and practice. Mental model scoring often relies on SMEs whose mental models act as a normative or referent structure. However, evidence shows that it is not unusual for experts to hold different mental models (e.g., Acton, Johnson, & Goldsmith, 1994) which complicates the interpretation of mental model scores and their implications for action. If there are multiple ways to accomplish a task, how should one rank order candidates based on their mental model scores? One reason for the observed differences between experts' knowledge structures is the complex nature of the tasks involved and the fact that for some complex tasks, there are multiple ways to effectively perform the same task (i.e., equifinality; Edwards, Day, Arthur, & Bell, 2006; Mathieu, Heffner, Goodwin, Cannon-Bowers, & Salas, 2005).

In sum, mounting evidence supports the construct validity of mental models. However, as highlighted in multilevel theories of training (e.g., Kozlowski et al., 2000, Kozlowski & Salas, 1997), evidence on the validity of mental models *across* levels of analysis is needed. In the next section, a discussion of multilevel theories and methods is presented to further explicate the issues that arise when dealing with multilevel phenomena.

2.1. Why a Multilevel Perspective?

As highlighted by Rousseau (1985), assessing multilevel effects becomes the inevitable subject of study once one recognizes that organizations are multilevel systems. Multilevel research is a conceptual and methodological approach germane to organizational science and other disciplines (e.g., education) where the phenomena under study involve potential multilevel effects. Thus, the need for a multilevel approach stems from the nature of organizations as multilevel systems and the need to clarify and specify complex conceptual and methodological issues that arise in the study of multilevel phenomena.

Although historically the influence of organizational systems theory had been merely metaphorical (Kozlowski & Klein, 2000), the multilevel approach has evolved towards a well-developed conceptual and methodological framework for studying multilevel issues. Numerous journal articles and books (including a dedicated book series; Yammarino & Dansereau, 2002-2009) attest to the presence of an emergent paradigm in organizational science (Kozloski & Klein, 2000) that is likely to influence the study of a broad range of phenomena that occur in organizations, impacting subfields of industrial and organizational psychology such as personnel selection (Schneider, Smith, & Sipe, 2000) and training (Kozlowski et al., 2000) that have customarily treated psychological phenomena as an individual-level issue.

By taking issues of levels into account one reduces the risk of model misspecification. A problem of misspecification occurs when the validity of the constructs employed has not been established at the focal level of analysis; when relationships occurring at one level of analysis (e.g., individual level) are interpreted at another level (e.g., team level); or when contextual effects are not accounted for when assessing lower-level relationships (Rousseau, 1985). A researcher interested in the relationship between individual differences (e.g., general mental ability, personality traits) and performance at the individual level may erroneously infer that similar functional relationships exist at higher levels, and falsely conclude that, for instance, implementing a selection system using said constructs will positively impact the performance of the organization as a whole. What is problematic about this claim is that in order to appraise the value of a selection system in terms of organizational-level performance, one must also consider the influence of processes occurring at higher-order levels and their impact on lower-level variables. For instance, because employees rarely perform their tasks in isolation, interactions between organizational members may hinder (thorough process losses; Steiner, 1972) or facilitate (through synergistic processes; Larson, 2010) individual performances in a group context, which may question the utility estimates of a selection system based exclusively on individual performance criteria. Steiner's (1972) theory of group productivity, for example, states that the potential productivity of a team will be offset by process losses associated with motivational processes occurring at the individual level in group contexts, such that greater opportunities for social loafing are present in relatively large groups which, in turn, will result in lower individual effort and decreased group performance.

From another perspective, the utility of selection systems as a human resource management *strategy* conflicts with evidence from strategy research demonstrating that gains from employing high-ability individuals can be offset by the compensation premiums required to hire them (Molloy, Ployhart, & Wright, 2010). Unfortunately, researchers typically assess the effectiveness of selection systems using individual performance as criteria instead of organizational-level criteria (Schneider, Smith, & Sipe, 2000), which illustrates how researchers sometimes overlook or underestimate the complexity of organizations as multilevel systems. Thus, broadly speaking, the cautionary tale of the multilevel approach is that researchers should not interpret the parts without understanding the whole, and vice versa.

2.2. Multilevel Models

As previously noted, theoretical frameworks and methodological guidelines for conducting multilevel research have been developed over the past decades. The purpose of this section is to identify the main themes that characterize this research stream with the purpose of illustrating the complexity of the issues surrounding multilevel research while providing a set of basic principles and terminology that will be used henceforth.

A number of scholars have examined the issue of multilevel theory development (Chan, 1998; Chen, Mathieu, & Bliese, 2005; House, Rousseau, & Thomas-Hunt, 1995; Klein, Dansereau, & Hall, 1994; Morgeson & Hofmann, 1999; Roberts, Hulin, & Rousseau, 1978; Rousseau, 1985). Rousseau was among the first to develop a typology of mixed-level models to describe the types of models and effects that characterize multilevel phenomena. According to Rousseau, there are three basic forms that mixed-level models can take, that is composition, cross-level, and multilevel.

Composition models specify the nature of the relationship between similar constructs presumed to exist at different levels. The relevance of delineating a composition model is that once the composition model is specified, it becomes clear how to combine the lower level unit data to establish the higher level construct (e.g., mean of some personality trait).[1] Building on Rousseau's (1985) work, Chan (1998) developed a typology of composition models to guide the development and validation of constructs in multilevel research. Chan's typology is comprised of five basic forms of composition models: (a) additive, (b) direct consensus, (c) referent-shift consensus, (d) dispersion, and (e) process composition. Table 1 summarizes Chan's typology of composition models. For each composition model there is a typical operation by which the lower level construct is combined to form its higher level counterpart. For instance, compared to additive models which involve a simple summation (e.g., mean) of the lower level data, direct consensus models require within-group consensus of the lower level units to justify aggregation. However, provided that within-group agreement has been established in direct consensus models, the validity of the aggregate index provides empirical support for the constructs of interest.

[1] Chan (1998) stated that global indices of higher level constructs (e.g., group size) are not as prevalent in multilevel research and that the variables of interest in multilevel research often rely on aggregated data from lower level units. Consonant with this, the following discussion focuses on shared and configural unit properties rather than global unit properties.

Table 1

Chan's (1998) Typology of Composition Models

Functional relationships	Typical operation combination	Empirical support
	Additive model	
Higher level unit is a summation of the lower level units regardless of the variance among these units	Summing or averaging lower level scores	Validity of additive index (e.g., mean of lower level units)
	Direct consensus model	
Meaning of higher level construct is in the consensus among lower level units	Within-group agreement to index consensus and justify aggregation	Value of within-group agreement index (e.g., r_{wg}); validity of aggregated scores
	Referent-shift consensus model	
Lower level units being composed by consensus are conceptually distinct though derived from the original individual-level units	Within-group agreement of new referent lower level units to index consensus and justify aggregation	Value of within-group agreement index (e.g., r_{wg}); validity of aggregated scores
	Dispersion model	
Meaning of the higher level construct is in the dispersion or variance among lower level units	Within-group variance (or its derivative) as operationalization of the higher level construct	Absence of multimodality in within-group distributions of lower level scores; validity of dispersion index
	Process model	
Process parameters at higher level are analogues of process parameters at lower level	No simple algorithm; ensure analogues exist for all critical parameters	Nomological validity for source and target constructs at their respective levels to distinguish shared core content from level-specific aspects

In an attempt to further extend previous work on compositional emergent phenomena, Kozlowski and Klein's (2000) typology explicitly considered the possibility of compilational forms of emergence. In contrast to composition models in which the type and amount of elemental content is similar among group members (i.e., isomorphism), compilational models assume discontinuity in either the type or amount of elemental content. In baseball, for instance, "the pitcher pitches, fielders field, and batters hit" (Kozlowski & Klein, 2000, p. 62). That is, whereas the elemental content in compilational forms of emergence comes from a common domain (e.g., baseball), the nature of individual contributions can be quite different.

In sum, a model of emergence (compositional or compilational) reflects the theoretical rationale and specifies the methodological implications (e.g., aggregation operations) for establishing multilevel constructs. For instance, do groups and other collectives possess characteristics such as "personalities"? Does personality have the same meaning across levels? And if so, what are the implications for measurement and data aggregation operations?

It is important to note that the emergence model is mainly concerned with the function of the construct of interest across levels, and not so much with its underlying structure. The structure of a construct refers to the actions and interactions between group members that result in the emergence of a collective phenomenon; in contrast, a functional analysis of a collective construct refers to the outputs or effects of a given construct (Morgeson & Hofmann, 1999). For instance, Chen, Bliese, and Mathieu (2005) argued that individual personality and group personality are not isomorphic in terms of the processes whereby individual and group personality develops. Whereas the former is based on genetic makeup and developmental experiences, the latter would be largely a function of social processes. However, individual and group personality may be *functionally* equivalent in that correlates of personality may be comparable across levels (Chen, Bliese, & Mathieu, 2005). One implication of the functionalist approach for understanding collective constructs is its usefulness for integrating constructs across levels at least during the first stages of multilevel theory development. However, in order to fully develop multilevel theories one also needs to explicate and identify the unique processes underlying the construct's structure at each level of analysis, thus integrating the functional and structural analysis of the construct of interest (Morgeson & Hofmann, 1999).

In contrast to composition models' which are concerned with *similar* constructs across levels, cross-level and multilevel models postulate relationships between *distinct* constructs. Figure 1 illustrates the broad type of mixed-level models identified in prior work (Klein & Kozlowski, 2000; Rousseau, 1985). *Direct cross-level* models describe a direct relationship between different constructs at different levels of analysis (Rousseau, 1985). The effect of organizational differences (e.g., human resource management practices) on individual performance is an exemplar of a cross-level direct effect. In addition to cross-level direct effects, it is also possible to hypothesize cross-level moderator effects in which higher-level constructs (e.g., job complexity) moderate the relationship between lower-level constructs (e.g., individual differences and individual performance). In statistical terms, the difference between direct and moderator cross-level effects is that the former involves intercept differences whereas the latter involves differences in slopes (Schneider et al., 2000).

Multilevel models (Rousseau, 1985) or *homologous multilevel models* (Kozlowski & Klein, 2000) postulate relationships among constructs that generalize across two or more levels. Thus, homology models assume that "X-Y relationships observed at one level of analysis are comparable to those obtained between similar variables at different levels of analysis" (Chen, Bliese, & Mathieu, 2005, p. 378). Because constructs in homology models are presumed to be similar in meaning across levels, researchers must first evaluate the extent to which the constructs are sufficiently similar across levels and, therefore, must pay special attention to the composition model. Thus, homologous multilevel models examine parallel relationships *and* parallel constructs at different levels of analysis.

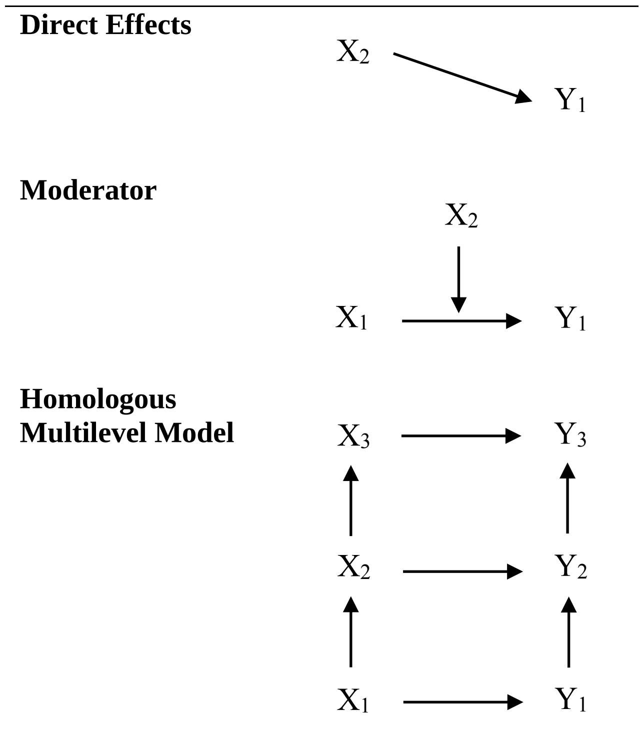

Figure 1. A Conceptual typology of mixed-level models (Rousseau, 1985)

2.2. Multilevel Analytical Strategies

Klein and Kozlowski (2000) identified several general approaches for the analysis of multilevel data. At the onset it is important to realize that “there is no one, all-encompassing multilevel data-analytic strategy that is appropriate to all research questions” (Klein & Kozlowski, 2000, p. 51), and that ultimately the choice of analytic strategy is dictated by the researchers’ questions and hypotheses. Some of the analytical strategies for testing multilevel models identified by Klein and Kozlowski (2000)—analysis of covariance (ANCOVA) and contextual analysis, cross-level and multilevel regression, and multilevel random-coefficient modeling—will be briefly discussed here.

However, because the present study proposes a multilevel theory of homology, Chen, Bliese, and Mathieu's (2005) framework for testing homologous multilevel theories is discussed in more detail.

ANCOVA and contextual analysis are among the earliest approaches to analyzing cross-level data (Klein & Kozlowski, 2000). To test for cross-level effects ANCOVA compares the variance explained by unit membership (i.e., the independent variable) to the variance explained by the individual-level predictors, which are treated as covariates in the model. In contrast to ANCOVA, contextual analysis is a regression-based approach that typically includes individual-level predictors and unit means on the same predictors. Thus, unlike ANCOVA which can only speak to differences attributable to the grouping variable, contextual analysis identifies the unit characteristic responsible for the observed differences. Because in contextual analysis the lower- and higher-level constructs are entered in the regression model concurrently, one implication of this procedure is that individual-level analogues of the contextual construct are controlled for.

Cross-level and multilevel regression proceeds by first evaluating the model of emergence, and subsequently tests the substantive study hypothesis. For example, prior to examining the relationship between group safety climate (a unit-level construct) and job satisfaction (an individual-level construct), the researcher should demonstrate that there is enough consistency or agreement within each group before aggregating to the higher-order level. Finally, multilevel random-coefficient modeling (or hierarchical linear regression; Hox, 2010) is mainly concerned with biases occurring at the individual-level of analysis that result from the violation of the assumption of independence in hierarchically structured data.

2.2.1. A Statistical Procedure for Testing a Multilevel Theory of Homology

Chen, Bliese, and Mathieu (2005) proposed a new analytical framework and method that can be used to test homologous multilevel theories. A discussion of Chen, Bliese, and Mathieu's analytical framework is germane to the present study because it is used to test the study's research hypotheses.

Chen, Bliese, and Mathieu (2005) stated that existing methods for testing homologous models such as within-and-between analysis (WABA; Dansereau, Alutto, & Yammarino, 1984; Dansereau & Yammarino, 2000) and multi-group or multilevel structural equation modeling (SEM; Muthén, 1994) are limited in flexibility. Among other things, because WABA analysis necessitates individual-level data, it cannot be used to test models that use primarily group consensus data or, more generally, models in which unit level data involve a single score at the group level that cannot be disaggregated to lower levels of analysis. In addition, WABA can only accommodate additive or direct consensus composition models but not referent-shift consensus measures because in the latter the higher-order construct (e.g., team-efficacy) is measured using a different metric than its individual-level analogue (e.g., self-efficacy). Chen, Bliese, and Mathieu's critique of multilevel SEM focuses on some statistical issues inherent to the SEM approach (e.g., identification problems, instability of estimates obtained from small sample sizes) and other challenges yet to be addressed such as the problem of standardizing SEM variables within versus between units.

However, perhaps the most problematic aspect of multilevel SEM highlighted by Chen, Bliese, and Mathieu relates to the fact that this technique treats higher-level units as independent from their lower-level counterparts, which precludes explicit tests of homology.

Chen, Bliese, and Mathieu's (2005) organizing framework is consistent with Widaman's (2000) levels of similarity (configural, scalar, and metric) and describes how the general linear model (GLM) and hierarchical linear modeling can be used to test sequential models to assess similarity across levels of analysis. Configural similarity is achieved when parameter estimates across different levels exhibit similar patterns of significance. For instance, consider the efficacy-performance relationship. If efficacy-performance associations are statistically significant across individual and team levels, then configural similarity is supported. In contrast to configural similarity, tests of scalar and metric similarity make specific predictions regarding the *magnitude* of the expected effects across levels of analysis. In the case of scalar similarity, parameter estimates from one level are related to parameters obtained at another level by a multiplicative function. For instance, one may hypothesize that the efficacy-performance association is twice as strong at the team level than at the individual level. Thus, whereas in tests of scalar similarity the expected pattern of effects is consistent across levels, the corresponding parameters are expected to differ in magnitude, although they are related by a scaling factor. Finally, metric similarity is a special case of scalar similarity using a rescaling factor of 1, that is, when the magnitudes of parameter estimates are not expected to differ significantly across levels.

2.3. Summary

As previously noted, construct validity encompasses various forms of evidence. In contrast to other forms of evidence, the generalizability aspect of mental models across levels of analysis has not been previously addressed. Although previous research suggests the presence of a multilevel construct, a multilevel approach is critical to evaluate the validity of mental models as predictors of performance across levels of analysis, that is, a multilevel test of homology. Finally, among the available methods for testing homologous multilevel models, Chen, Bliese, and Mathieu's (2005) analytic framework seems the most appropriate to address this issue.

3. DELINEATION OF MENTAL MODELS AS A MULTILEVEL PHENOMENON

Kozlowski and Klein (2000) stated that "the first and foremost task in crafting a multilevel theory or study is to define, justify, and explain the level of each focal construct that constitutes the theoretical system" (p. 27). Because the objective of this study is to investigate the relationship between mental models and performance using a multilevel framework, the goal of the following sections is to review the extant literature on both individual and team mental models, and delineate a multilevel theory to characterize the relationship between mental models and performance across levels of analysis by integrating previous conceptual work in this area (e.g., Klimoski & Mohammed, 1994; Mohammed & Dumville, 2001). Specifically, a proportional theory of homology (Chen, Bliese, & Mathieu, 2005) is offered as a framework to test the multilevel validity of mental models as a predictor of complex skill acquisition.

3.1. Mental Models

In the context of Kraiger, Ford, and Salas' (1993) taxonomy, a mental model measures an individual's knowledge organization. Mental models reflect the structural relationships that exist between a set of concepts within a given knowledge domain (Johnson-Laird, 1983; Kraiger et al., 1993; Rouse & Morris, 1986; Wilson & Rutherford, 1989). Structural knowledge assessment is the most common approach for assessing mental models (Kraiger et al., 1993). In structural assessments, individuals provide pair-wise relatedness ratings amongst a sample of relevant terms (i.e., concepts or tasks). The pair-wise relatedness ratings are then used to generate a representation of the individual's mental model. The preponderance of the mental model research has focused on the sharedness or similarity, and accuracy of mental models. Similarity is typically used in the context of team research, and refers to the extent to which team members share an understanding of the team's task (e.g., Edwards et al., 2006). In contrast, accuracy represents the extent to which an individual's (or team's) mental model approximates the "true" state of the world or "the" expert model[2] (Acton et al., 1994).

One reason for measuring mental models is to understand the relationships that exist between different components of a specified knowledge domain. Assessing mental models is important because of the informational value of characterizing the relationships between these basic knowledge components. In other words, mental model assessment techniques emphasize the importance of eliciting the *configural* property of knowledge (i.e., how knowledge is organized) rather than the *amount* of information an individual can recognize or recall, which is typically captured through traditional declarative knowledge tests as typified by multiple-choice exams.

Although the use of verbal reports (e.g., thinking aloud protocols) for the purpose of uncovering users' mental models is not uncommon, this approach is limited in that much expert knowledge consists of mental processes that are automatic and unconscious. Basically, in order to speed up performance, many of the processes and strategies that experts employ to solve problems are combined into chunks of

---

[2] Unlike the term "expert-based model" the term "expert model" is used here as analogous to the term "expert system", a computer program to emulate human decision-making ability. For all intents and purposes, an expert model is short for correct or "true" structure of the underlying domain.

productions, and thus, it is often difficult for the expert to reconstruct the original steps, which leads to difficulties in the explicit expression of knowledge (Nisbett & Wilson, 1977). Consonant with automaticity as a defining feature of expertise, researchers often ask participants to "not ponder their judgment" (Schvaneveldt et al., 1985, p. 704) while completing the pairwise relatedness task. Thus, the impetus for using the pairwise elicitation method is that it offers the possibility of capturing expert knowledge while overcoming one important limitation of traditional interview and verbal protocol techniques which is the fact that experts have difficulty in reliably reporting on their mental processes (Cooke & McDonald, 1986). Not at odds with this premise, Kraiger and Salas (1993) found that mental models were more sensitive as training criteria than an explicit measure of knowledge (i.e., a multiple-choice test)—that is, mental models discriminated between a control group and a group of trainees who underwent a naval aircrew coordination training, but multiple-choice tests did not—which suggests that mental models capture something that explicit knowledge measures cannot.

3.2. Team Mental Models

Teams are two or more individuals who work interdependently, have specific roles and assignments, and interact and coordinate to achieve a common goal (Baker & Salas, 1996; Kozlowski & Ilgen, 2006; McIntyre & Salas, 1995). Because team tasks require some degree of coordination amongst team members to successfully perform their task, a defining feature or job characteristic of team tasks is the presence of interdependence (Arthur, Edwards, Bell, Villado, & Bennett, 2005; Arthur, Glaze, Bhupatkar, Villado, Bennett, & Rowe, 2012).

The implicit coordination of teams operating in complex environments has been linked to the cognitive underpinnings of team tasks. Because teams perform cognitive tasks (e.g., decision making, problem solving, planning), an understanding of team cognition seems critical to understanding team performance (Cooke, Kiekel, Salas, Stout, Bowers, & Cannon-Bowers, 2003; Hinsz, Tindale, & Vollrath, 1997). Consonant with this reasoning, several reviews of teams have identified cognitive emergent states as a relevant antecedent of team functioning (Ilgen, Hollenbeck, Johnson, & Jundt, 2005; Kozlowski & Bell, 2003; Kozlowski & Ilgen, 2006; Mathieu, Maynard, Rapp, & Gilson, 2008). These reviews have focused on two constructs related to team cognition—team mental models (e.g., Edwards et al., 2006) and transactive memory systems[3] (Moreland, Argote, & Krishnan, 1996). For instance, research has shown that mental model sharedness is positively associated with team performance. Such findings are typically explained in terms of team members' increased ability to anticipate the needs and actions of other team members (e.g., Mohammed & Dumville, 2001; Mohammed, Ferzandi, & Hamilton, 2010) which facilitates team coordination and team performance.

In the team literature, the analogue of an "individual" mental model is a "team" mental model. Team mental models have been defined as organized mental representations of relevant knowledge that enable team members to coordinate their efforts, facilitate information processing, provide mutual support, and, in general, adapt

[3] Transactive memory systems refer to knowledge that is distributed among team members. Although transactive memory systems have been shown to outperform shared mental models as an antecedent of both team processes and team performance (see Tables 5 and 6 in DeChurch & Mesmer-Magnus, 2010b), there is no individual level analogue of this construct. Consequently, it is inappropriate to postulate a multilevel homology theory for this construct.

to cognitively demanding tasks (e.g., Edwards et al., 2006). The term team mental model—instead of, for instance, *shared* mental model—refers to a broad category that encompasses both shared mental models and accurate mental models (or *quality* mental models; Mathieu et al., 2005). In the team training literature, mental models have been recognized as a process variable that mediates the relationship between team interventions and team outcomes (Mohammed, Ferzandi, & Hamilton, 2010). DeChurch and Mesmer-Magnus' (2010b) meta-analysis demonstrated that team mental models are valid predictors of behavioral processes, motivational-affective states, and team performance, and that team mental models explain additional variance in team performance after accounting for behavioral processes and motivational-affective states.

As previously noted, an important issue in the team mental model literature is the distinction between similarity and accuracy. Team mental model sharedness or similarity involves knowledge that is common among team members (Cannon-Bowers, Salas, & Converse, 1993; Cannon-Bowers & Salas, 2001), whereas team mental model accuracy indexes the extent to which team mental models match an expert mental model (e.g., Lim & Klein, 2006).

Whereas team mental model accuracy and similarity have been shown to be highly related (e.g., *r*s between .61 and .67; Edwards et al., 2006) accuracy tends to be more strongly associated with performance than similarity (ρ = 34 and .30, respectively; DeChurch & Mesmer-Magnus, 2010b). However, there are many tasks for which it may be difficult to derive one "correct" model and thus it is not uncommon to find discrepancies between experts' mental models (e.g., Acton et al., 1994). Mathieu et al.

(2005) argued that such discrepancies may very well exist in the presence of equifinality—that is, when there are multiple equally good, but yet different ways to successfully complete a task. Although one could use heterogeneous expert models to yield normative comparisons, sharedness or similarity is more appropiate than accuracy when tasks exhibit equifinality. Nonetheless, when there is a "true score" against which mental models can be compared to, similarity among team members is expected to increase with accuracy because teammates' mental models will coalesce as they approximate the true state of the world.

Consequently, because complex team tasks are often characterized by equifinality, similarity or sharedness will be useful in most, if not all situations, whereas accuracy requires a known "true state of the world" (Edwards et al., 2006). In addition, the formation of shared mental models is concomitant with enhanced team processes (e.g., communication and coordination) which have been emphasized in team research due to their critical role as antecedents of team performance (Mathieu et al., 2000). Accordingly, the relevance of mental model sharedness or similarity for team coordination and effective team performance has been acknowledged in models of teamwork. Salas, Sims, and Burke (2005), for instance, identified shared mental models as one of the three coordinating mechanisms—along with close-loop communication and mutual trust—that support what they called the "Big Five" of teamwork. Despite the ubiquity of sharedness compared to accuracy, meta-analytic results also support the validity of team mental model accuracy (DeChurch & Mesmer-Magnus, 2010b).

Although a team can develop mental models of a number of facets of their job or task, most researchers in this domain collapse the content of team mental models into two categories—taskwork and teamwork (Mohammed et al., 2010). Mathieu et al. (2000) described taskwork mental models as knowledge of the technology or equipment with which the user is interacting, as well as knowledge regarding the task procedures, task strategies, likely contingencies or problems, and environmental/task conditions. In contrast, teamwork models include an understanding of the team interaction requirements (e.g., roles and responsibilities, information flow) as well as specific information about the team members (e.g., knowledge, skills, attitudes, preferences, strengths and weaknesses). Researchers agree that effective teams have individuals who not only perform task-related functions well but can also work together as a team.

Whereas team cognitions are often conceptualized as antecedents of performance, of equal or perhaps greater importance is the issue of explaining the mechanisms whereby individual cognitions develop into a team mental model or "group mind" (Klimoski & Mohammed, 1994). Team training interventions have been shown to have a positive impact on team-level cognitive outcomes (ρ = .42; Salas et al., 2008) which demonstrates the usefulness of team cognitions as training criteria. Team processes, such as planning (Stout, Cannon-Bowers, Salas, & Milanovich, 1999), leader pre-briefs and team interaction training (Marks, Zaccaro & Mathieu, 2000), and cross-training (Marks, Sabella, Burke, & Zaccaro, 2002) have also been shown to enhance mental model sharedness.

3.3. Mental Model Measurement

There are different measurement strategies consistent with the definition of mental models as a structural knowledge construct. Structural assessment is a general approach for assessing mental models or knowledge structures (Kraiger et al., 1993) that involves three steps: (1) knowledge elicitation, (2) knowledge representation, and (3) evaluation of an individual's knowledge representation (Goldsmith, Johnson, & Acton, 1991).

Knowledge elicitation refers to the technique used to capture the content of knowledge, such as similarity ratings, concept maps, and card sorting tasks.[4] Similarity ratings can be obtained using a pairwise relatedness task in which participants are asked to report on their perceptions about the relation between concepts or how often events co-occur. Card sorting is a well-known technique in psychology that involves writing concepts on cards and then sorting or placing them as to what is closest to what. Marks et al. (2000), for instance, used team-interaction concept maps to operationalize team mental model similarity. For this measure, participants selected 8 cards with concepts written on them and placed them on a concept map. The concept map rows represented action sequences of team members and the concept map columns represented a cross-section of what all team members should be doing concurrently. The overlap between team members' concept maps was used to operationalize team mental model similarity.

[4] Although Likert scale questionnaires can be used to elicit the content of mental models, these are not customarily used in conjunction with a representation technique to capture the cognitive structure underlying individual or team cognition. So, consonant with Mohammed et al. (2000) they are not discussed here as a mental model measurement technique.

Knowledge representation refers to the technique or scaling procedure used to derive an individual's cognitive structure. Alternative methods for representing the structure of knowledge exist such as UCINET, multidimensional scaling (MDS), and Pathfinder.

UCINET (Borgatti, Everett, & Freeman, 2002) is a social network analysis program that also uses nodes (or actors if the network is social in nature) and links between nodes to reflect the structural organization of knowledge. MDS generates a representation in geometric space for understanding similarities between a set of objects and reveal underlying dimensions respondents use when evaluating those objects. A single $n \times n$ square symmetric matrix of similarities serves as the input data which are then analyzed via Euclidean distances (e.g., Carroll & Chang, 1970). Concept mapping characterizes the causal linkage among concepts derived from observations, interviews, and questionnaire data. Concept mapping has been used for the study of organizational research topics such as decision making, negotiation, and organizational cognition (Mohammed et al., 2000). As an example, individuals may be interviewed about the causal relationships occurring between a set of concepts related to medical diagnosis. The content of the interview can then be coded to determine the presence of causal links between symptoms (e.g., coughing) and plausible causes (e.g., viral illnesses, infections, smoking, etc.). Pathfinder networks are based on similarity ratings that reflect the psychological proximity or distance between concepts. The Pathfinder algorithm (Schvaneveldt et al., 1988) generates a structure consisting of nodes with links connecting some pairs of nodes.

The final step of structural assessment is knowledge evaluation which consists of evaluating the extent to which mental models are similar to a referent structure (i.e., accuracy) or the extent to which team members' mental models are similar (i.e., sharedness or similarity).

Although the usefulness of these techniques may depend on the research context (Mohammedet al., 2000), empirical evidence within the team literature has shown that mental model measurement strategies differ in terms of their associations with team outcomes (DeChurch & Mesmer-Magnus, 2010a). Specifically, stronger relationships with team processes is evident when similarity ratings are used as the elicitation method and the Pathfinder network algorithm is used to represent structure. In contrast to team processes, measurement strategy does not seem to affect the relationship between mental model similarity and performance. Empirical data on the effectiveness of measurement strategy as a moderator of the validity of *individual* mental models is scarce but consistent with the team literature in that Pathfinder-based scores yield stronger predictive validities than MDS (Goldsmith & Johnson, 1990). Thus, based on the available literature on mental models, the decision was made to operationalize individual and team cognition using Pathfinder instead of other measurement strategies.

3.3.1. Pathfinder Networks

Pathfinder networks (Schvaneveldt et al., 1988) can be used to identify structural aspects of knowledge, such as memory organization or category structure. A computer program, namely Pathfinder (Schvaneveldt, 2009a), can be used to draw a spatial representation of a mental model in which each node represents a component of the model (e.g., a concept), and each link between nodes reflects the relatedness that exists between the model components. Networks are derived using two parameters, *r* and *q*, to determine how network node distance is calculated (*r* is typically set to infinity and *q* is set to equal the number of concepts minus one).

The resultant networks can then be evaluated in terms of similarity, and accuracy scores. Whereas similarity refers to the extent to which team members share an understanding of the team's environment, accuracy scores represent the extent to which an individual's (or team's) mental model approximates an expert model, or a referent structure that best reflects the true knowledge structure of the domain (Acton et al., 1994). However, the metric used to represent mental model similarity and accuracy is based on the same index, namely, closeness (*C*; Goldsmith & Davenport, 1990). Goldsmith and Johnson (1990) advocate the use of *C* as the best way to operationalize the extent to which mental models resemble each other. *C* is roughly equal to the ratio of the number of common links between two networks divided by the total number of links in both; it varies from 0 to 1 with 1 representing a perfect match between two mental models (Day et al., 2001; Kraiger et al., 1995).

In addition to *C* scores, Pathfinder yields three additional indices, mental model correlations, number of links, and coherence scores. The index of correlation between an individual's relatedness-ratings matrix and a referent structure accomplishes the same purpose as *C* scores. However, in contrast to *C* scores which can be conceptualized as an index of agreement, Pathfinder's index of correlation indexes consistency.

A mental model can also be evaluated in terms of its parsimony and coherence. The number of links reflects a knowledge structure's parsimony, where fewer links indicate more parsimony and vice versa. More parsimonious models are considered superior to relatively less parsimonious models. In contrast, coherence scores are usually considered a measure of internal consistency. Coherence scores can be used as an indication of the individual's ability to provide a consistent pattern of responses (Schvaneveldt, 2009a). Whereas previous research has shown that *C* scores, correlations, and coherence scores yield comparable validities in the prediction of skill-based performance, the number of links has not been demonstrated to be a valid predictor of skill-based performance (Schuelke et al., 2009).

3.4. Summary

Parallel to research on individual expertise (Chi, Glaser & Farr, 1988), cognition has been shown to be also important for team performance. In essence, team and individual mental models are homologous constructs (Chen, Bliese, & Mathieu, 2005). That is, the *function* of team cognition is analogous to individual cognition in that mental models serve as a basis for selecting actions that are consistent with current task demands, thereby accounting for behavioral differences observable at both individual and team levels.

However, the *processes* whereby team and individual cognitions translate into performance are evidently different. At the individual level, the effective application of knowledge requires processes such as retrieval and pattern recognition, whereas at the team level processes such as communication and coordination are critical for transforming a collection of individuals' knowledge into effective team knowledge (Cooke et al., 2004). Previous research has demonstrated that team processes—strategy formation and coordination, cooperation, and communication—are associated with mental model sharedness, such that teams with convergent taskwork and teamwork mental models tend to display higher quality team processes (Mathieu et al., 2000).

Interestingly, past literature reviews have focused solely on *team* mental models and have paid limited attention to *individual* mental models. However, a recent meta-analysis by McDonald and Muñoz (2013), demonstrated that the effect size of individual mental models is almost twice as large as the effect size of team mental models for predicting task performance (r = .35 and .18, respectively).[5] Thus, although previous meta-analytic evidence suggests that team mental models are valid predictors of team performance (DeChurch & Memer-Magnus, 2010b), difficulties in the process of integrating team members' knowledge for optimal task execution may attenuate, mask, or understate the importance of team cognitions for team performance. For instance, Steiner's (1966, 1972) notion of process loss indicates that a team's potential performance may be reduced due to factors related to poor coordination or other group processes (e.g., low motivation) which may reduce the influence of team cognitions on performance.

---

[5] The meta-analytic estimate of r = .18 for the association between team mental models and performance from McDonald and Muñoz (2013) is considerably lower than DeChurch and Mesmer-Magnus' (2010a) estimate of r = .28. Estimates from both meta-analyses come from studies that use Pathfinder to represent structure. However, McDonald and Muñoz employed a relatively larger number of studies (k = 16, N = 1,050) than De DeChurch and Mesmer-Magnus (k = 5, N = 287), which may account for the noted estimate differences.

4. MENTAL MODELS AS A MULTILEVEL CONSTRUCT

In the previous section *individual* mental models were distinguished from *team* mental models. In the following section, issues pertaining to the conceptual basis and justification for scaling individual mental models to the team level are discussed within the context of a multilevel framework for conceptualizing mental models as a multilevel phenomenon. Next, a proportional theory of homology is offered as a multilevel framework to test the validity of mental models as a predictor of complex skill acquisition across levels of analysis. Finally, the statistical procedures to test such model are outlined and described.

4.1. Nature of Emergence

Kozlowski and Klein (2000) distinguish between two forms of emergence—composition and compilation. Whereas composition models of emergence describe "phenomena that are essentially the same as they emerge upward across models . . . compilation describes phenomena that comprise a common domain but are distinctively different as they emerge across levels" (p. 16). Composition and compilation are also distinguished by their underlying theoretical models. Whereas composition is based on isomorphism—where the type and amount of elemental content is similar for all individuals in the collective—compilation is based on a model of discontinuity—where either the amount or type of elemental content is different, or both the amount and type are different.

Although the measurement techniques used to index team mental models may be at the individual level, a team mental model is an emergent property of the collective

(Klimoski & Mohammed, 1994). Thus, team mental models can be conceptualized as a bottom-up emergent construct that originates in the cognitions of individuals. Some scholars have suggested that at the root of the distinction between similarity and accuracy resides an important conceptual distinction with relevant implications for understanding the multilevel nature of this construct. Mathieu et al. (2005), for instance, posited that whereas similarity emerges through compilational processes to manifest as higher-level phenomena, accuracy is akin to a compositional construct. Specifically, they state that:

> It is important to note that the notion of shared mental models is a *configural* type of team construct. It derives from the consistency of individuals' models, yet there is no "team model" per se . . . rather, the convergence index itself represents the extent to which individuals share a common knowledge structure. . . . However, team QMM [quality of mental models] represents a *summary index* of the quality of members' models relative to some standard(s). (Mathieu et al., 2005, pp. 38-39)

Interestingly, some disparity exists amongst scholars regarding the nature of the emergence of team mental models. For instance, in contrast to Mathieu et al.'s (2005) perspective, DeChurch and Mesmer-Magnus (2010b) consider both sharedness and accuracy as compositional constructs. Kozlowski and Klein (2000) posited that differences in the conceptualization of emergence for a given construct may very well exist. For instance, most researchers use mean accuracy scores to operationalize team mental model accuracy (e.g., Cooke et al., 2003; Lim & Klein, 2006). After mental model accuracy is calculated for each individual team member, these indices are averaged to yield a team accuracy score. Alternatively, one could use the minimum, the maximum, the variance, or the profile or pattern of team members' accuracy scores to operationalize team mental model accuracy. Each of these would correspond to a different form of emergence and, thus, engender different assumptions regarding the cognitive underpinnings of team performance (i.e., how team members integrate their cognitions during performance). Stout, Cannon-Bowers, and Salas (1996) suggested that the effective operationalization of team mental models depends on the task demands. For instance, under conditions in which communication between team members is difficult, shared mental models become more crucial to team functioning because in these situations teams are precluded from strategizing "on the fly". Consistent with this reasoning, mental model accuracy may also become more important when communication is difficult. On the contrary, if team members can communicate freely, then they may help each other by sharing their knowledge during the performance episode (e.g., helping hypothesis; Lepine, Hollenbeck, Ilgen, & Hedlung, 1997). In the latter case, the highest within-team mental model accuracy score will be more predictive of team performance than the team mean accuracy.

The point is that, in the absence of a universal form of emergence for a given construct, it seems important to justify (either theoretically or empirically) the specific form of emergence and correspondent procedures employed to aggregate data to the team level. Although researchers rarely provide a conceptual basis for using mean accuracy scores as the aggregation method, this procedure seems to be aligned with the cognitive underpinnings of team performance. Because team performance depends on the successful execution of all individual roles, it has been suggested that an overall estimate of team members' ability is the most effective operationalization of team ability for predicting team performance (Day, Arthur, Miyashiro, Edwards, Tubré, & Tubré, 2004). Also, using the mean in this context is consistent with the premise that mental models serve to enhance team members' ability to anticipate the needs and actions of other team members (Mohammed & Dumville, 2001; Mohammed, Ferzandi, & Hamilton, 2010). Although mental model assessment could target team members' accurate knowledge of their *own* role, the content of mental models may comprise team members' understanding of their own role *and* of the task as a whole, including other team members' tasks (i.e., interpositional knowledge). In this case, two individuals with non-overlapping mental models may have similarly low accuracy scores due to, for instance, their accurate taskwork knowledge about their own roles and their *lack* of knowledge of the other team members' role (Cooke et al., 2003, pp. 194-195). In contrast, when team mean accuracy is high, team average scores would rightly suggest that team members are knowledgeable about the different facets of the job comprising other team members' tasks as well as their own. Yet if the content of mental models focuses solely on role knowledge, then the minimum individual score may be more predictive of team performance because unknowledgeable members could have a substantial adverse effect on team performance (e.g., Chen, Thomas, & Wallace, 2005)

Although it is possible that nonredundant mental representations may fit together in a complementary way to create a whole (Kozlowski, Gully, Salas, & Cannon-Bowers, 1996), having an understanding of the whole task seems germane for the team cognition construct to be meaningful in the sense of explaining the successful integration of team members' actions during task performance. Although this idea is reminiscent of Lepine et al.'s (1997) *helping hypothesis*, anticipating the needs of other team members is not the same as "helping" because coordination among team members does not imply differential team member ability. In other words, team members coordinate their efforts because their roles are interdependent, not to compensate for each other's lack of ability.

Thus, using Kozlowski and Klein's (2000) typology, team mental model accuracy is best represented by a pooled unconstrained emergence type,[6] in which the type of content remains the same across levels but variation in the *amount* of elemental content is possible (i.e., team members accuracy scores may vary considerably). Because in pooled unconstrained models one does not need to meet the assumption of isomorphism—as in pure composition models—within-group agreement is not necessary to justify the aggregation of lower level variables to reflect the higher level construct. At issue here is that the team members do not have to have equally accurate mental models to permit the aggregation of individuals' models to the team level to reflect the team mental model. Having said that, researchers recommend that the group-mean reliability (ICC(2)) should be estimated in pool unconstrained models.[7] Even in instances where there are no theoretical reasons to expect clustering to occur, the estimation of non-

---

[6] Chan (1998) referred to them as additive models.

[7] Bliese (2000) uses the term *fuzzy composition* for a model that is fundamentally equivalent to the pooled unconstrained model identify by Kozlowski and Klein (2000).

independence (i.e., the degree to which lower-level responses are affected by higher-level factors) is important if the data are nested or have a hierarchical nature (Bliese, 2000) because even a small degree of non-independence can substantially bias statistical models (Bryk & Raudenbush, 1992). Thus, "estimates of non-independence . . . provide a way of determining the degree to which lower-level relationships will be biased if higher-level effects are not adequately controlled for . . . regardless of whether or not one is explicitly interested in modeling contextual effects" (Bliese, 2000, p. 365).

Whereas mental model similarity has been described sometimes as a compositional construct (e.g., DeChurch & Mesmer-Magnus, 2010b), in the present study similarity is conceptualized as a compilational construct, a position that is in accordance with that espoused by Mathieu et al. (2005). Similarity is typically operationalized by comparing team members' mental models to each other and then averaging the similarity between team member pairs to obtain a team-level similarity index. To reiterate Mathieu et al.'s position, the similarity index itself represents the extent to which individuals share a common knowledge structure, and it is used directly as a predictor of team performance, which is analogous to using climate strength as opposed to the content of the climate for predicting organizational-level outcomes.

It is also important to mention that whereas the individual contributions to the higher-level phenomenon can be constrained to be similar by imposing a threshold to justify aggregation to the higher level, this approach would reduce or restrict *between-team variance in similarity* (for instance, teams with low rw_gs would be discarded). Thus, yet another reason for conceptualizing similarity as a compilational construct (Mathieu et al., 2005) is that this approach allows for *between-team variance in similarity* which is the variance purportedly associated with changes in team performance. To the extent that team member mental models are similar, team performance is likely to improve (provided that mental models are accurate).

4.2. Testing a Proportional Theory of Homology

Rousseau (1985) and Kozlowski and Klein (2000) defined *homologous* multilevel models as models in which the relationships linking a set of constructs are generalizable across organizational levels. Chen, Bliese, and Mathieu (2005) further refined this framework by developing a theoretical and methodological approach to test homologous models.

Chen, Bliese, and Mathieu (2005) posited that homology theories evolve from exploratory to confirmatory phases. Historically, mental model research has made substantial progress, to the point that this line of research is now viewed as "one of the more developed collective cognition literature streams" (Mathieu, Maynard, Rapp, & Gilson, 2008, p. 429). One can literally see how this literature has advanced from the exploratory to confirmatory phase by looking at the titles of two widely cited papers in this research domain—from "Team Mental Model: Construct or Metaphor?" (Klimoski & Mohammed, 1994) to "Metaphor No More: A 15-Year Review of the Team Mental Model Construct" (Mohammed et al., 2010). Thus, it is not unreasonable to specify a theory of homology that states more explicitly the nature and strength of the relationships between team mental models and performance across levels of analysis. Specifically, consonant with previous research showing that mental model-performance relationships are stronger at the individual (e.g., Acton et al., 1994; Davis, Curtis, & Tschetter, 2003; Day et al., 2001; Goldsmith et al., 1991; Rowe et al., 1996) than the team level (e.g., Cooke et al., 2003; Lim & Klein, 2006; Marks, Sabella, Burke, & Zaccaro, 2002), the present study proposes a proportional theory of homology to describe the mental model-performance relationship such that proportionately stronger associations will be demonstrated at the individual level relative to the team level. This proposition is also consistent with Chen, Thomas, and Wallace's (2005) finding that task-related knowledge was a stronger predictor of performance at the individual level than the team level. Consequently, the proposed study is an extension of previous multilevel research on the knowledge-performance relationship using a structural assessment technique to measure individual and team knowledge—instead of a multiple-choice test as in Chen, Thomas, and Wallace’s study.

So, on the basis of the above noted distinction between levels of similarity, and in view of previous studies showing stronger associations between mental models and performance at the individual level of analysis, it was hypothesized that:

Hypothesis 1: The relationship between mental model accuracy and performance will be stronger at the individual level than the team level.

It is well-established that the acquisition of job knowledge is one of the best predictors of job performance (e.g., Hunter, 1986). Knowledge acquired during training has also been shown to predict performance on complex tasks (e.g., flying pilots work samples; Ree, Carretta, & Teachout, 1995). Finally, evidence from multilevel studies have shown that declarative knowledge is more strongly related to performance at the individual then the team level of analysis (Chen, Thomas, & Wallace, 2005).[8] Assessing

---

[8] Chen, Thomas and Wallace (2005) used items that covered both declarative, and procedural and strategic knowledge. However, for the sake of simplicity and given the nature of the method used to assess knowledge (i.e., a multiple-choice test) these aspects of knowledge are referred henceforth as declarative knowledge.

the validity of declarative knowledge in addition to mental models was important to replicate previous findings concerning the validity of these constructs within as well as across level of analysis. Also, comparing how declarative knowledge and mental models function across levels of analysis will contribute to further understanding the role of cognitive measures in complex training performance.

Chen, Thomas, and Wallace's (2005) study used a 10-item multiple choice test to assess specific *role* knowledge. Due to the highly interdependent nature of the task, Chen, Thomas, and Wallace used the minimum individual score to aggregate knowledge at the team level because one team member with low role knowledge could disrupt team functioning. Whereas the performance task of the present study was also highly interdependent, the mean was used to operationalize team declarative knowledge because the declarative knowledge measure used in the present study assessed team knowledge, which combines both role *and* interpositional knowledge. Therefore, consistent with the expectation of Hypothesis 1, it was hypothesized that:

> *Hypothesis 2: The relationship between declarative knowledge and performance will be stronger at the individual level than the team level.*

5. RELATIONSHIP BETWEEN MENTAL MODELS AND BEHAVIORS

Team performance indexes the degree to which the team was able to successfully meet the demands of the situation and can be operationalized using both objective (e.g., number of targets destroyed, completion time) and subjective (e.g., supervisor ratings) measures. Within the team literature, the preponderance of mental model research has focused on team performance (e.g., points on a computer simulation) rather than behavioral processes, which comprises the actions and interactions of team members during team performance. The distinction between results and processes is commonly acknowledged in models of training criteria (i.e., Kirkpatrick, 1994) and criterion development for personnel decision making (e.g., Campbell, Dunnette, Lawler, & Weick, 1970). The risk of relying exclusively on performance results or output measures is that it limits the understanding of the psychological and/or behavioral processes involved in performance (e.g., Cascio & Aguinis, 2011). Consonant with this view, team actions and behavioral processes (what teams *do*) have also been included in models of team performance (e.g., Kozlowski & Ilgen, 2006).

Evidence in support of a mental model-behaviors association seems germane for advancing mental model research. Such consistencies would demonstrate that mental model assessments reflect the processes utilized during task performance. For instance, the procedural knowledge contained in a mental model should map on to an observable pattern of behaviors. In turn, the set of enacted behaviors is the mechanism that "translates" a mental model into effective performance. Thus, an additional contribution of the present study is to expand the criterion domain of mental model research by including the behavioral component of team performance.

5.1. Team Performance as a Compilational Construct

Whereas the focus of the present study is on mental models, it is essential to also describe the nature of the emergence of team performance, the study's focal dependent variable. Observing and analyzing these interrelated phenomena (i.e., team mental models and team performance) yields a clearer understanding of the mechanisms that underlie the association between team cognition and team performance.

As previously noted, Kozlowski and Klein (2000) describe two types of emergence—composition (shared unit properties) and compilation (configural unit properties). As a higher-level phenomenon, team performance emerges from the pattern of team member's execution of their individual roles. Consequently, the behavioral processes that result in team performance are akin to a compilation model where the configuration of different lower-level behaviors emerge, bottom-up, to characterize the performance of the team as a whole. For instance, consider a case where one member of a team serves as the gunner and a second one serves as the driver of a tank. Task interdependency exists at the team level such that the tank cannot be operated successfully without the combined effort of the gunner and the driver. Team performance emerges as the gunner and driver enact their roles and execute their interdependent yet qualitatively distinct tasks.

At the individual level, the processes and strategies that experts employ are thought to be combined into chunks of productions. Mental models capture this knowledge (e.g., if-then rules) and, thus, uncover what guides experts' actions during a performance episode. Unlike individual performance, team performance requires that team members coordinate their actions to successfully perform as a team, and coordination demands increase with task interdependence. Consequently, if all the tasks that comprise a job are interdependent, then team members must work with each other constantly to perform effectively. However, complex tasks are usually comprised of individual and team tasks (e.g., Arthur et al., 2012). In the context of teams, disparities between team members' mental models may result in weaker associations between mental models and behaviors compared to individual tasks. In addition, the magnitude of the relationship between mental model accuracy and behaviors may be attenuated by less than perfect coordination amongst team members. Thus, stronger mental-model-behavior associations are expected at the individual level than the team level.

Because not all the similarity ratings obtained during the mental model assessment translate into observable behaviors exhibited during team performance, the relationship between mental models and behaviors was limited to an examination of a subset of similarity ratings as they relate to their correspondent behaviors (see Method section for details). Thus, it was hypothesized that:

Hypothesis 3: The relationship between individual similarity ratings and individual behaviors will be stronger than the relationship between team similarity ratings and team behaviors.

7. METHOD

7.1. Participants

The study sample consisted of 243 individuals nested in 81 3-person teams. Participants were recruited from the subject pool of the psychology department of Texas A&M University to participate in the study in partial fulfillment of a course requirement. In addition to course credit, participants were eligible to earn a monetary reward of $80, $40, or $20 (per person) for teams with the three highest average performance scores during the study.

A power analysis was conducted using G*Power 3.1 (Faul, Erdfelder, Buchnar, & Lang, 2009) to estimate the optimal number of participants. Because the results of the power analysis will differ depending on the type of homologous model tested, the following power estimates are based on the more restrictive homology models—scalar or metric similarity. For instance, to test for metric similarity, Chen, Bliese, and Mathieu (2005) recommended the use of an *F*-test to evaluate the fit of two nested models (see Statistical Analyses section for details). If the resultant *F* value is not significantly different from zero (i.e., observed R^2 increase is not statistically significant) then metric similarity is demonstrated.

A power analysis was conducted to determine the power level associated with the final sample size. Consonant with Chen, Bliese, and Mathieu's (2005) recommendation, a liberal *p* value of .20 was adopted to evaluate the statistical significance of the *F* test. Using an alpha of .20, the sample size of 81 teams resulted in a 99% chance of detecting a medium effect size ($F^2 = 0.15$) and a 76% chance of detecting a small to medium effect size ($F^2 = 0.05$).

Configural homology is demonstrated when observed correlations are statistically significant across individual and team levels. So, based on the sample size of 81 teams, the power to reject the null hypothesis that correlations across levels are different from zero is higher than the power to reject scalar or metric similarity. Consequently, the power to test configural similarity was higher than the power to test scalar and metric similarity. At the team level, the achieved power for detecting a medium effect size ($\rho = .30$, two-tailed, $\alpha = .20$, $N = 81$), was 84% whereas at the individual level, the achieved power for detecting a medium effect size ($\rho = .30$, two-tailed, $\alpha = .05$, N = 243) was 99%.

7.2. Measures

7.2.1. Crisis in the Kodiak: Oilrig Search and Rescue

The performance task was a disaster response simulation that was developed to include task, goal, and feedback interdependencies (Arthur, Naber, Jarrett, Glaze, Shurig, McDonald, & Muñoz, 2011). Missions entail the roles of oilrig workers, helicopter aviators, and boat captains tasked with responding to an off-shore oilrig explosion. Each team operates nine platforms (three for each role) to achieve two goals of shutting off oil valves and rescuing survivors (see Figure 2). Platforms have unique capabilities that can be used individually and interactively to accomplish mission objectives. So, team members must coordinate the three platforms that comprise their assigned roles to achieve the mission goals. Each mission lasted 10 minutes. Participants also completed the task individually by controlling all nine platforms simultaneously; otherwise, the individual and team missions were identical. Team members communicated with each other via voice-activated microphones and headphones. Team members performed the individual and team missions in the same room at their own computer stations facing the wall.

Figure 2. Example screenshot of the *Crisis in the Kodiak: Oilrig Search and Rescue* simulation.

7.2.1.1. Performance and Behaviors

Participants have two equally important goals—shutting off valves to stop the flow of oil fueling the oilrig fire, and healing and picking up survivors. Points are earned for survivors healed, survivors rescued, and oil valves shut off. There were 20 survivors. Participants receive 10 points for each survivor healed and another 10 points for each healed survivor picked up. There are 4 shutoff valves and each valve successfully turned off is worth 50 points. The maximum performance score is 600 points per mission—regardless of whether the mission is performed individually or as a team. Individual and team performance total scores across the five trials (baseline performance, and performance missions 1 through 4) served to operationalize individual and team performance.

Behaviors were assessed by counting the number of times participants engaged in specific types of actions during the missions. For instance, an oilrig worker has capabilities that allow this platform to heal an injured survivor by itself or in conjunction with the helicopters or the boats. So, as an example, if an individual chose to use only the helicopter to heal survivors over the boats, then a stronger mental model link should exist between the concepts of “healing survivors” and “helicopter” than between the concepts of “healing survivors” and “boats”. Consequently, the frequency with which specific actions were carried out during the performance episode were used to operationalize behaviors at the individual and team level.

7.2.2. Mental Models

Target, a graphical user interface developed by Schvaneveldt (2009b) was used to assess participants' mental models. Taking each concept in turn, the program presents the concept at the center of a bull's eye-type diagram. Participants were instructed to drag-and-drop the remaining concepts around the focal concept or target to reflect the closeness or similarity of the concepts to the focal concept or target. Hence, relatedness ratings were based on the distance between a given concept and the target (0 = *less related or unrelated*, 4 = *synonyms*). The same procedure is repeated until every concept is positioned as the focal concept (see Figure 3 and the Appendix).

Fifteen concepts were used to measure mental models in the present study. These concepts were developed through a task analysis (Table 2). The concepts were reviewed, revised, and finalized via consensus by three SMEs who were senior Ph.D. Industrial and Organizational Psychology graduate students involved in the design and development of Crisis In The Kodiak: Oilrig Search And Rescue and with more than a year of experience with the research tool. All SMEs have more than 50 hours of experience playing the simulation as individuals and teams. As a team, the SMEs consistently succeeded at shutting off the four oilrig valves in the scenario while healing and rescuing over 15 of the 20 survivors (approximately 500 out of 600 possible points). Individually, SMEs were able to shut off 3 of the 4 oilrig valves and heal and rescue about 10 survivors (approximately 350 points). As a comparison, mean individual performance scores among pilot study participants was 157.45 (SD = 93.83). In contrast, mean team performance for the pilot study was 222.67 (SD = 96.89). Thus, individually and as a team, SMEs outperformed pilot study participants by approximately 2 SDs. The SMEs included the performance task simulator lead programmer and the author of the present study.

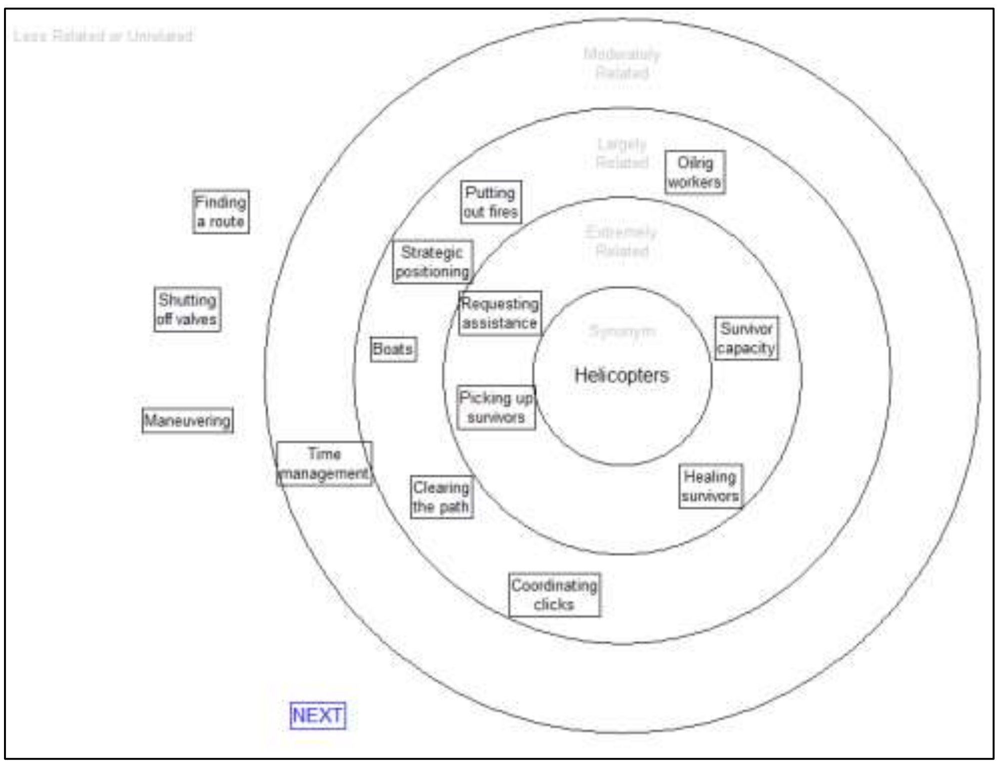

Figure 3. Screen capture of a participant's view of the bull's eye diagram for presenting mental model terms. The focal term ("Helicopter") is at the center of the bull's eye while the remaining terms are being dragged and dropped around the focal concept to indicate their degree of relatedness (synonyms, extremely related, largely related, moderately related, and less related or unrelated [outside the bull's eye]).

Table 2

Concepts Utilized for Mental Model Assessment

Concepts	Description	Final mental model terms
Platforms		
Oilrig Worker	Rescue platform moving on the surface of the rig with the capacity of autonomously healing burn injuries, putting out all types of fire, and shutting off valves. It can also stop hemorrhage and hypothermia but only in conjunction with other platforms.	Oilrig Worker
Helicopter	Rescue platform flying above the rig with the capacity of autonomously stopping hemorrhages, putting out electrical fires, and evacuating survivors by air. It can also heal burns and hypothermia and putting out structural and oil fires but only in conjunction with other platforms.	Helicopter
Boat	Rescue platform circling around the rig with the capacity of autonomously stopping hypothermia, putting out electrical fires, and evacuating survivors by water. It can also heal burns and hypothermia and putting off structural and oil fires but only in conjunction with other platforms. Boats have a larger view range compared to other platforms. Thus, although boats are limited to circle around the rig, their view range permits their involvement in actions occurring inside the rig.	Boat

Table 2 (cont'd)

Concepts Utilized for Mental Model Assessment

Concepts	Description	Final mental model terms
Capability tasks		
Matching vulnerabilities and capabilities		
Healing survivors	Finding relevant information for diagnosing and treating worker injuries.	Healing survivors
Putting out fires	Finding relevant information for diagnosing and extinguishing different types of fire.	Putting out fires
Shutting off valves	Using shut-off valve capability to shut down valves.	Shutting off valves
Picking up survivors	Rescuing healed survivors by picking them up.	Picking up survivors
In-game planning		
Finding a route	Finding the best path to get to a specified destination quickly.	Finding a route
Coordinating clicks	Communicating with other decision makers to use combined capabilities in a timely manner.	Coordinating clicks
Time management	Completing tasks and goals within the specified time limit.	Time management
Maneuvering around obstacles	Moving around scenario quickly.	Maneuvering
Managing survivor capacity	Being aware of the remaining pick-up capacity.	Survivor capacity
Non-capability tasks		
Strategic positioning	Positioning platforms strategically for searching for valves and survivors, and helping other decision makers.	Strategic positioning
Clearing the path for oilrig workers	Putting out fires along a path for oilrig workers to reach valves quickly.	Clearing the path
Requesting assistance	Communicating needs to other platforms.	Requesting assistance

7.2.2.1. Mental Model Accuracy and Similarity

Accuracy scores represent the extent to which an individual's (or team's) mental model approximates an expert model. Mental model accuracy was obtained using C scores to index the closeness of each individual mental model to the expert-based model. Team mental model accuracy was operationalized as the mean accuracy (mean C scores) of each team. The normative or referent structure used for the present study was the leading programmer's mental model. As expected, the normative mental model displayed acceptable psychometric characteristics. First, the referent structure coherence score of .24 was higher than the cutoff point of .20 recommended by Schvaneveldt (2009a). Second, as indicated by the ratio of links (20 links) to the total number of components (15 components) the network was considered fairly parsimonious (Figure 4).

Team mental model similarity is typically operationalized using the average similarity between team members. Within each team, individual members' mental models were compared to each other. Then, the mean similarity between team member pairs was used to index team mental model similarity—that is, average *C* scores within the team.

7.2.3. Declarative Knowledge

A 3-alternative 42-item test was developed to assess participants' declarative knowledge of *Crisis in the Kodiak*. This test was administered at baseline prior to task exposure and at three additional time-points throughout the study protocol. Items on this test covered basic knowledge concerning the team mission (e.g., "How many points does a team earn for each shutoff valve it successfully turns off ?") and the simulation interface (e.g., "In which tab does Crisis in the Kodiak display information about what first aid items can be used to heal a survivor?"). It also included items to assess team knowledge (i.e., knowledge about team members' capabilities). Example items of team knowledge were ""Which platforms cannot pick up healed survivors?" and "Which platforms can independently extinguish an oil fire without help from another platform?". Consequently, the team knowledge items assessed role knowledge (i.e., knowledge about team members own capabilities) as well as interpositional knowledge (i.e., knowledge about other team members capabilities.)

At each administration, individual total scores were calculated as the percentage of correct answers. Based on the premise that team knowledge is best represented by an additive composition model in which the team-level construct is the mean of the lower-level scores, team knowledge was operationalized as the average of individual members' declarative knowledge test scores within each team. Excluding the baseline measure, the reliability of the declarative knowledge scores were acceptable. At the individual level, the test-retest reliabilities for the declarative knowledge scores were .78, .72, and .78 for Time 2/Time 3, Time 2/Time 4, and Time 3/Time 4, respectively. At the team level, the test-retest reliabilities for the declarative knowledge scores were .85, .81, and .82 for Time 2/Time 3, Time 2/Time 4, and Time 3/Time 4, respectively.

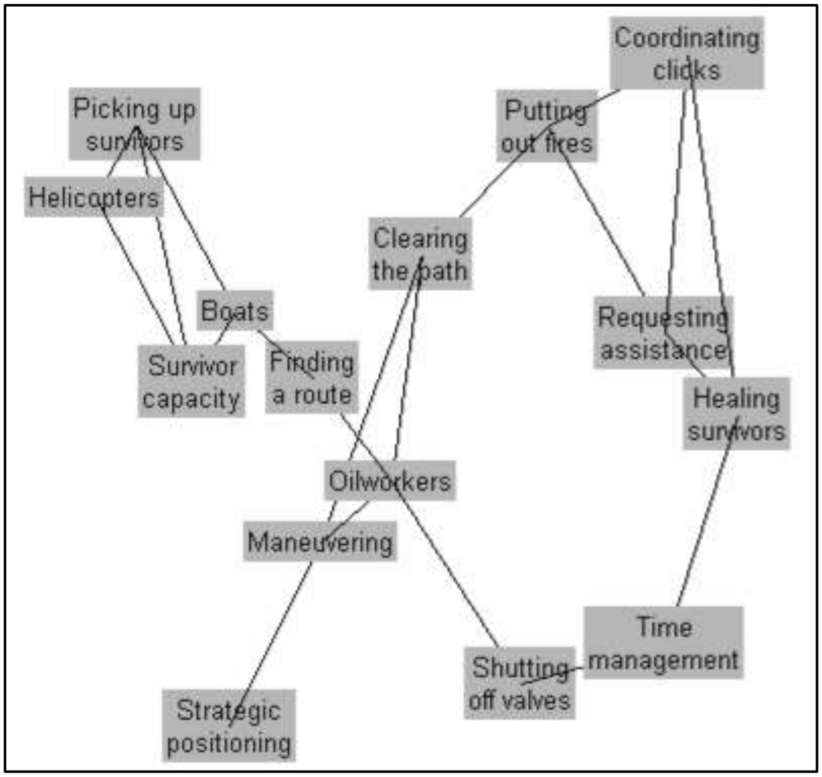

Figure 4. Pathfinder network of the referent mental model.
Coherence = .24; Number of links = 20.

7.3. Design and Procedure

Because the present study is concerned with the relationship between measured variables (i.e., mental models and performance), the study is a multilevel correlational design in which individuals are nested in teams. Individuals and teams were trained to perform the computer-based simulation over the course of a 2-day 48-hour interval protocol (Table 3). At the beginning of the study, participants were randomly assigned to one of three roles (oilrig worker, helicopter aviator, or boat captain) and performed this role during team missions throughout the duration of the study.

Participants participated in a dynamic, networked computer-based simulation. Three-person teams operated the simulator collectively (through specialized roles) and as individuals (performing all roles simultaneously). Thus, participants performed the same complex task as individuals and teams.

During the first phase of the study (Day 1), participants were trained to operate the simulator and perform a series of individual and team missions. Next, after a 48-hour interval, participants returned to complete the second and final session of the study (Day 2). The Day 1 and Day 2 protocols were 3 and 2 hours long, respectively. Participants received pre-recorded in-role and interpositional tutorials, which were self-paced and interactive. A task aid was available onscreen during training and performance.

In addition to the baseline team and individual performance missions, participants completed eight additional missions (four as individuals and four as a team). Team and individual missions were counterbalanced—individual missions followed by team missions, and team missions followed by individual missions. Prior to all team missions, participants were given 2-minute planning periods with their teammates where they were encouraged to formulate a mission strategy. Mental models were measured at the end of Day 2.

Table 3

Study Protocol

Day	Activity[A]
1	Spoken Training
	Briefing/Planning
	Team Mission 0
	Briefing/Planning
	Individual Mission 0
	In-Role Training
	Interpositional Training
	Briefing/Planning
	Team Practice Mission
	Individual Practice Mission
	Briefing/Planning
	Team Mission 1
	Briefing/Planning
	Individual Mission 1
	Briefing/Planning
	Team Mission 2
	Briefing/Planning
	Individual Mission 2
48-hour interval	
2	Briefing/Planning
	Team Mission 3
	Briefing/Planning
	Individual Mission 3
	Briefing/Planning
	Team Mission 4
	Briefing/Planning
	Individual Mission 4
	Mental Model Assessment

Note. The order of team and individual missions (practice and test missions) was counterbalanced.

7.4. Statistical Analyses

In addition to descriptive statistics and intercorrelations amongst study variables at individual and team levels, Chen, Bliese, and Mathieu's (2005) statistical procedures for testing similarity across levels was used to test the study hypotheses. Although the study hypotheses focus on differences between relationships across levels of analysis, a first step is to determine if these associations are statistically significant within each level. If the relationships between mental models and performance are statistically significant within each level, then the higher-level estimates can be compared to the lower-level estimates to determine how much those estimates differ and, more importantly, if the observed difference is consistent with the directionality of the study hypotheses. Specifically, the following steps are necessary to test for homologous theories:

1. Estimate the lower level model (M_1) by regressing individual performance on the individual level predictor using RCM.
2. Calculate the expected team performance for each team by multiplying each team-level predictor by the regression weights obtained in step 1. Then, estimate the proportionally constrained model by regressing the actual team performance scores on the predicted team performance scores (M_2).
3. Estimate a higher level model in which the estimates of the higher level regression weights are freely estimated (M_3).
4. Compare the baseline (M_3) and proportionally constrained model (M_2) using Equation 1:

$$F = \frac{[SS_{error(M2)} - SS_{error(M3)}]/(k_{M3} - k_{M2})}{SS_{error(M3)}/\ (N - k_{M3})} \tag{1}$$

Where $SS_{\text{error}(M2)}$ and $SS_{\text{error}(M3)}$ are the sum of squares error for M_2 and M_3, k_{M3} is the number of parameters estimated in M_3, k_{M2} is the number of parameters estimated in M_2, and N is the team-level sample size.

5. Scalar similarity is supported if the F test is nonsignificant. Failure to reject scalar similarity indicates that the pattern of the magnitude of effects does not differ within versus between units. That is, a single scaling factor applies equally well to the set of predictors.
6. To test for metric similarity, the fit of M_2 is compared to a new model (M_4) in which the rescaling factor is set to 1. Specifically, the following equation was used to test for metric similarity:

$$F = \frac{[SS_{error(M4)} - SS_{error(M2)}]/(k_{M2} - k_{M4})}{SS_{error(M2)}/\ (N - k_{M2})} \tag{2}$$

Where $SS_{\text{error}(M4)}$ and $SS_{\text{error}(M2)}$ are the error terms for M_4 and M_2, k_{M2} is the number of parameters estimated in M_2, k_{M4} is the number of parameters estimated in M_4, and N is the team-level sample size.

8. RESULTS

8.1. Mental Model Accuracy and Declarative Knowledge as Predictors of Performance Across Levels of Analysis

Individual- and team-level descriptive statistics are presented in Tables 4 and 5, respectively. The pattern of performance scores suggest that individuals and teams improved their performance—from nearly zero points for baseline performance up to 160.08 (Individual Mission 4) and 268.77 points (Team Mission 4). In contrast, participants' scores on the declarative knowledge test improved substantially following the baseline assessment (48%) but then remained fairly stable around 76-80% accuracy. In addition, intercorrelations between performance scores are consistent with a simplex pattern, with adjacent trials been more strongly correlated compared to trials further apart. Specifically, correlations amongst performance scores for adjacent trials ranged from .27 to .78 at the team level, and from .22 to .79 at the individual level. As expected, correlations between the baseline performance and performance scores from other missions are lower than correlations between missions 1-4 due to a floor effect for initial task performance. Extremely low scores (and variance) during initial task performance is not surprising considering that the task is deemed very difficult for someone with no previous experience with the task.

Table 4

Individual-Level Descriptive Statistics and Intercorrelations Amongst the Study Variables

Variable	*N*	*M*	*SD*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
1. Age	243	18.98	1.78											
2. Sex	243	-	-	.15*										
3. DK Time 1	243	48.25	9.00	.04	.07									
4. DK Time 2	237	76.43	11.76	-.04	.22*	.19*								
5. DK Time 3	243	76.53	12.29	-.04	.21*	.25*	.78*							
6. DK Time 4	237	79.88	11.22	-.07	.18*	.23*	.72*	.78*						
7. Mental Model Accuracy	243	0.26	0.06	-.02	.01	-.01	.11	.23*	.23*					
8. Individual Mission 0	240	4.38	15.27	.04	.14*	.01	.05	.09	.06	-.03				
9. Individual Mission 1	241	78.42	62.90	.00	.45*	.00	.23*	.31*	.24*	.18*	.22*			
10. Individual Mission 2	242	111.69	74.23	-.12	.37*	-.01	.23*	.26*	.25*	.21*	.21*	.70*		
11. Individual Mission 3	240	118.08	78.94	-.07	.42*	.07	.28*	.32*	.31*	.25*	.25*	.70*	.76*	
12. Individual Mission 4	243	160.08	89.58	-.08	.44*	.06	.30*	.33*	.32*	.24*	.18*	.61*	.69*	.79*

Note. DK = declarative knowledge. DK scores indicate percentage of correct answers. Dummy codes for sex are female = 0 (N = 140) and male = 1 (N = 103).For each mission, performance scores could range from 0-600. $^{*}p < .05$ (two-tailed).

Table 5

Team-Level Descriptive Statistics and Intercorrelations Amongst the Study Variables

Variable	*N*	*M*	*SD*	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
1. Age	81	18.98	1.19												
2. Sex	81	-	-	.28*											
3. DK Time 1	81	48.25	5.37	-.07	.09										
4. DK Time 2	79	76.43	7.90	-.03	.25*	.34*									
5. DK Time 3	81	76.53	8.34	-.05	.19	.42*	.85*								
6. DK Time 4	80	79.90	7.40	-.04	.14	.37*	.81*	.82*							
7. Mental Model Accuracy	81	0.26	0.03	.08	.04	.14	.20	.18	.14						
8. Mental Model Similarity	81	0.36	0.09	.05	-.07	.07	.07	.20	.20	.27*					
9. Team Mission 0	78	2.82	10.80	.01	-.06	-.11	.09	.09	.12	.16	.15				
10. Team Mission 1	80	150.75	84.28	.05	.41*	.06	.41*	.43*	.34*	.19	.30*	.27*			
11. Team Mission 2	79	204.18	91.49	-.04	.37*	.14	.41*	.43*	.32*	.33*	.22	.15	.78*		
12. Team Mission 3	78	230.13	99.39	.09	.34*	.09	.34*	.39*	.31*	.27*	.34*	.17	.72*	.71*	
13. Team Mission 4	81	268.77	98.53	-.18	.19	-.05	.19	.24*	.18	.28*	.29*	.13	.58*	.62*	.73*

Note. DK = declarative knowledge. DK scores indicate percentage of correct answers. Dummy codes for sex are female = 0 (N = 140) and male = 1 (N = 103). For each mission, performance scores could range from 0-600. $^*p < .05$ (two-tailed).

Because the study used a longitudinal design, Tables 4 and 5 include data collected at different time points (e.g., declarative knowledge measured at Time 1, 2, 3, and 4). Nevertheless, because mental models were assessed only at the end of Day 2, performance scores from team mission 4 and individual mission 4 were used as the focal dependent variables to test the study hypotheses.

Although the objective of the present study was to investigate the relationship between mental model *accuracy* and performance, the relationship between mental model *similarity* and performance was also examined. As can be seen in Table 5, the relationship between mental model similarity and team performance was positive and statistically significant for team mission 4 (r = .29, p < .05). Also, mental model similarity was moderately positively correlated to mental model accuracy (r = .27, p < .05). These finding suggest that teams with accurate mental models tend to have similar mental models because their mental models converge as they approximate the expert mental model which, in turn, is thought to be a reasonably good approximation of the true state of the world.

8.2. Testing a Homology Theory of Mental Models

Chen, Bliese, and Mathieu (2005) posited that a necessary condition to conduct scalar similarity analyses is the presence of consistent statistically significant correlations across levels of analysis. An examination of Tables 4 and 5 indicates that the relationship between mental model accuracy was statistically significant at both the individual (r = .24, p < .05) and team level of analysis (r = .28, p < .05). In contrast, declarative knowledge's association with performance was statistically significant at the individual level ($r = .32$, $p < .05$) but not at the team level ($r = .18$, $p > .05$). Additional analyses conducted at the individual level demonstrated that mental model accuracy predicted additional variance in performance after accounting for the effect of declarative knowledge, $\Delta R^2 = .03$, $F\ (1{,}234) = 7.23$, $p < .05$. Thus, although declarative knowledge and mental model accuracy were correlated at the individual level ($r = .23$, $p < .05$), they accounted for different portions of the criterion space.

Because only mental model accuracy was a statistically significant predictor of performance across levels, tests of scalar and metric similarity were conducted using mental model accuracy as the sole predictor. Consonant with Chen, Bliese, and Mathieu's (2005) framework for testing similarity across levels, a RCM was estimated—using the PROC MIXED procedure in SAS (version 9.3)—with mental model accuracy as the predictor and individual performance as the dependent variable (Model 1). As recommended by Chen, Bliese, and Mathieu (2005), all variables in the model were previously standardized in order to make the resulting estimates comparable across levels. Results from Model 1 indicated that mental model accuracy was a statistically significant predictor of performance ($\beta = .23$, $t = 3.80$, $p < .05$) after accounting for team nesting.

Next, a new predicted score was created by multiplying the higher-level predictor (i.e., team mental model accuracy) by its respective lower-level estimate obtained in the previous step. The specific values used to create the new performance scores were $\beta_0 =$

.00 and β_1 = .23. The next step was to test for scalar similarity by regressing team performance on the newly created predictor (Model 2). Because only one predictor was used to estimate the new variable (i.e., mental model accuracy), the parameter estimate resulting from this step is interpreted directly as the rescaling factor to equate the individual- and team-level estimates for the mental model accuracy-performance relationship. According to Model 2 the team-level parameter was slightly stronger than the individual-level parameter. Specifically, the team-level parameter was 1.21 the size of the individual level parameter.

To evaluate scalar similarity, Model 2 is typically compared to a third model (Model 3) in which the association between team mental model accuracy and performance is freely estimated. However, a comparison between Model 2 and Model 3 is meaningless if the models have only one predictor. When there is more than one predictor, the comparison is meaningful in the sense that one could evaluate whether the rescaling factor (i.e., the parameter estimate of 1.21 from Model 2) applies equally well to a *set* of predictors. So, because the comparison between Models 2 and 3 was unnecessary, the next step was to examine whether the observed value of 1.21 was statistically significantly different from 1—that is, to test for metric similarity. To accomplish this, the fit of Model 2 was compared to a model where the team-level and individual-level parameters were set to be identical (Model 4). Again as per Chen, Bliese, and Mathieu's (2005) recommendation, to test for metric similarity the fit of

Model 4 was compared to the fit of Model 2 by comparing the error terms of both models using Equation 2.

The *SS* error for the full metric similarity model was 73.80, whereas the *SS* error for Model 2 was 73.61. Applying Equation 2 yielded an F value of 0.20 ($p > .05$) which indicates a failure to reject metric similarity. In other words, the observed difference between the magnitudes of the individual and team-level coefficients were not statistically significant—the rescaling factor of 1.21 was not statistically significantly different from 1. Hypothesis 1 stated that the relationship between mental model accuracy and performance would be stronger at the individual than the team level of analysis. Contrary to Hypothesis 1, the comparison between the constrained and unconstrained models demonstrates that the magnitude of the relationship between mental model accuracy and performance was similar at the team level and the individual level.

Hypothesis 2 stated that declarative knowledge's relationship with performance would be stronger at the individual level compared to the team level of analysis. Because declarative knowledge was not a statistically significant predictor of performance at the team level, scalar or metric invariance tests were precluded. Notwithstanding, the present results support Hypothesis 2.

The previous analyses used the mean to operationalize team declarative knowledge and team mental model accuracy. However, to compare the efficacy and appropriateness of alternative composition models, additional analyses were conducted using the minimum and the maximum team member mental model accuracy and declarative knowledge as predictors of team performance. Results indicated stronger associations between team mental models and performance when using the minimum (r = .29, p < .05) rather than the maximum (r = .12, p > .05) to operationalize team mental model accuracy. In contrast, the pattern of results for declarative knowledge were in the opposite direction with weaker associations between declarative knowledge and performance when using the minimum (r = .10, p > .05) instead of the maximum (r = .21, p > .05). The predictive validities of the alternative operationalizations of mental model accuracy and declarative knowledge—the minimum and the maximum, respectively—were not statistically significantly different from the predictive validities of mental model accuracy and declarative knowledge using the *mean*. However, the differences between the minimum and the maximum for predicting performance were statistically significant. Specifically, the differences of .17 (r_{min} - r_{max} = .29 - .12 = .17) for mental model accuracy and -.11 (r_{min} - r_{max} = .10 - .21 = -.11) for declarative knowledge were statistically significantly different from zero (p < .10; see Meng, Rosenthal, & Rubin, 1992). Thus, although similar to the mean, the specified alternative operationalizations (i.e., minimum and maximum) of mental models and declarative knowledge were statistically significantly different from each other in predicting team performance.

Concordant with the previous analyses, homology tests were conducted only if statistically significant effects were found at the individual *and* team levels. Thus, given that the maximum declarative knowledge was not statistically significantly associated with performance ($r = .21$, $p > .05$), homology tests were not conducted for this measure. Homology tests conducted for mental models using the minimum indicated that the scaling factor for team mental model accuracy was 1.23. However, this value of 1.23 (very close to the 1.21 obtained for the mean mental model accuracy) was not statistically significantly different from 1, $F = 0.27$, $p > .05$.

Together, these results suggest that individual and team mental model accuracy are valid predictors of individual and team performance and that the magnitude of this relationship is similar across the individual and team levels. Further analyses suggested that the method for aggregating declarative knowledge and mental models has substantial implications for predicting team performance. For mental model accuracy, the minimum was superior to the maximum for predicting team performance, whereas for declarative knowledge the maximum was superior to the minimum for predicting team performance.

8.3. Relationships Between Similarity Ratings and Behaviors Across Levels of Analysis

A set of 15 observed behaviors—obtained from an analysis of the text output that is automatically generated during each mission—were used as criteria for the analyses of the relationships between similarity rating and behaviors (Table 6). Figure 5 displays an example of the text output.

```
1       Attack Initiated:
2       13 attacked by:
3               Boat 1 using Blankets
4
5
6       Attack Joined:
7       13 attacked by:
8               Boat 1 using Blankets
9       Helo 2 using Blankets
10
11
12      Attack Expired:
13      13 attacked by:
14              Boat 1 using Blankets
15              Helo 2 using Blankets
16
17
18      Attack Initiated:
19      13 attacked by:
20              Boat 1 using Oxygen
21
22
23      Attack Expired:
24      13 attacked by:
25              Boat 1 using Oxygen
26
27
28      Attack Initiated:
29      13 attacked by:
30              Boat 1 using Oxygen
31
32
33      Attack Joined:
34      13 attacked by:
35              Boat 1 using Oxygen
36              Helo 2 using Blankets
37
38
39      Attack Expired:
40      13 attacked by:
41              Boat 1 using Oxygen
42              Helo 2 using Blankets
```

Figure 5. Example of Text Output with Behavioral Data

In the example presented in Figure 5, lines 2-3 indicate that one of the boats used blankets to heal survivor #13. Subsequently, one of the helicopters joined the boat by providing blankets (lines 6-9). Lines 12-15 indicate that the event was terminated as no other platform joined the action. Although a single target was engaged, the events from lines 2 to 15 were coded as a three separate behaviors—boat engaging a survivor, helicopter engaging a survivor, and boat interacting with helicopter. Lines 18-42 suggest that the previous engagement was unsuccessful as survivor #13 was again engaged by the boat and the helicopter now using a different set of capabilities (i.e., oxygen and blankets). Events from lines 18-42 constitute three additional events—again, boat engaging a survivor, helicopter engaging a survivor, and boat interacting with helicopter. The 15 behaviors coded for the present study are presented in Table 6.

Table 6

Coded Behaviors and Corresponding Similarity Rating Pairs

Type of Behaviors	Similarity rating pairs
Helicopter healing survivor	Helicopter - Healing survivors
Helicopter picking up survivor	Helicopter - Picking up survivors
Helicopter putting out fire	Helicopter - Putting out fire
Helicopter shutting off valves[A]	Helicopter - Shutting off valves
Oilrig Worker healing survivor	Oilrig Worker - Healing survivor
Oilrig Worker picking up survivor[B]	Oilrig Worker - Picking up survivor
Oilrig Worker putting out fire	Oilrig Worker - Putting out fire
Oilrig Worker shutting off valves	Oilrig Worker - Shutting off valves
Boat healing survivor	Boat - Healing survivor
Boat picking up survivor	Boat - Picking up survivor
Boat putting out fire	Boat - Putting out fire
Boat shutting off valves[A]	Boat - Shutting off valves
Helicopter working with Oilrig Worker	Helicopter - Oilrig Worker
Helicopter working with Boats	Helicopter - Boats
Oilrig Worker working with Boats	Oilrig Worker - Boats

Note. [A]Neither the helicopters nor the boats are capable of shutting off valves, but putting off a fire on top of a valve counts as a valve engagement. [B]Because oilrig workers cannot pick up survivors, the frequency count for these behaviors is effectively zero.

Table 7

Frequency of Individual and Team Behaviors per Type of Behavior

	Individual level $N = 159$				Team level $N = 53$				
Type of Behaviors	*M*	*SD*	*Min*	*Max*	*M*	*SD*	*Min*	*Max*	*d*
Helicopter healing survivor	8.56	5.94	0	29	12.42	5.43	0	22	0.68*
Helicopter picking up survivor	6.08	4.69	0	22	8.42	5.06	0	23	0.48*
Helicopter putting out fire	2.89	2.61	0	10	7.47	14.78	0	108	0.43*
Helicopter shutting off valves[A]	2.75	3.27	0	25	2.79	2.82	0	10	0.01
Oilrig Worker healing survivor	3.87	3.31	0	12	5.17	4.21	0	20	0.34*
Oilrig Worker picking up survivor[B]	0.00	0.00	0	0	0.00	0.00	0	0	0.00
Oilrig Worker putting out fire	8.28	5.11	0	26	15.55	21.35	4	162	0.47*
Oilrig Worker shutting off valves	4.86	4.38	0	28	7.58	6.28	0	43	0.50*
Boat healing survivor	3.89	4.35	0	18	10.06	5.22	0	23	1.28*
Boat picking up survivor	1.74	2.85	0	18	4.79	3.62	0	18	0.94*
Boat putting out fire	1.54	2.41	0	18	8.32	19.22	0	108	0.49*
Boat shutting off valves[A]	0.94	1.65	0	11	2.34	3.14	0	14	0.56*
Helicopter working with Boats	5.20	4.34	0	17	4.19	4.18	0	21	-0.24*
Oilrig Worker working with Boats	3.23	3.61	0	19	5.19	5.30	0	23	0.43*
Oilrig Worker working with Helicopter	1.15	1.67	0	7	1.30	1.81	0	8	0.09
Total	54.97	18.21	8	122	95.58	45.17	34	378	1.18*

Note. [A]Neither the helicopters nor the boats are capable of shutting off valves, but putting off a fire on top of a valve counts as a valve engagement. [B]Because oilrig workers cannot pick up survivors, the frequency count for these behaviors is effectively zero.

Table 7 presents the mean frequency of each type of behavior at the individual and team level of analysis. Overall teams performed more behaviors (M = 95.58, SD = 45.17) than individuals (M = 54.97, SD = 18.21), $t(158) = 11.10$, $p < .05$, $d = 1.18$. Interestingly, the number of times that oilrig workers and helicopters worked together to perform a task was similar between individual and team missions. Whereas oilrig workers interacted more with boats during team missions, helicopters interacted more with boats during individual missions.

Because the mental model assessment was based on 15 concepts, each individual mental model was comprised of $15 \times (15 - 1) = 210$ similarity ratings. A subset of 15 similarity ratings (out of the 210) that matched the 15 specified behaviors was used to test Hypothesis 3. For instance, the rating for the pair *helicopter-healing survivors* was used to estimate the number of times a helicopter healed a survivor. Because each pair was rated twice during the mental model assessment, the two ratings were combined into a single rating by calculating the median between the two. The decision to use the median (instead of the mean, for instance) is consistent with the procedure for combining mental models from different individuals in Pathfinder (Schvaneveldt, 2009a).

.

Table 8

Mean Individual and Team Similarity Ratings by Concept Pair

	Individual level $N = 159$				Team level $N = 53$				
Pair	*M*	*SD*	*Min*	*Max*	*M*	*SD*	*Min*	*Max*	*d*
Helicopter healing survivor	2.64	0.75	0.00	4.00	2.81	0.31	2.00	3.50	0.30*
Helicopter picking up survivor	2.97	0.54	1.00	4.00	2.92	0.28	2.00	3.50	-0.12
Helicopter putting out fire	2.56	0.66	0.00	4.00	2.63	0.41	1.50	3.00	0.13
Helicopter shutting off valves	1.22	1.09	0.00	4.00	1.14	0.90	0.00	3.00	-0.08
Oilrig Worker healing survivor	2.32	0.78	0.00	4.00	2.33	0.54	1.00	3.00	0.01
Oilrig Worker picking up survivor	0.69	0.95	0.00	3.00	0.50	0.74	0.00	3.00	-0.22
Oilrig Worker putting out fire	2.64	0.63	0.00	4.00	2.75	0.40	1. 50	3.50	0.21
Oilrig Worker shutting off valves	3.37	0.52	2.00	4.00	3.35	0.47	2.50	4.00	-0.04
Boat healing survivor	2.60	0.72	0.00	4.00	2.75	0.33	2.00	3.50	0.27*
Boat picking up survivor	2.70	0.80	0.00	4.00	2.78	0.54	0.00	4.00	0.12
Boat putting out fire	2.50	0.73	0.00	4.00	2.58	0.42	2.00	3.50	0.13
Boat shutting off valves	1.08	1.06	0.00	4.00	0.95	0.85	0.00	3.00	-0.14
Helicopter working with Boats	1.38	1.25	0.00	4.00	1.36	1.08	0.00	4.00	-0.02
Oilrig Worker working with Boats	1.11	1.18	0.00	4.00	1.00	0.97	0.00	3.00	-0.10
Oilrig Worker working with Helicopter	1.41	1.28	0.00	4.00	1.37	1.09	0.00	4.00	-0.03
Total	2.08	0.49	0.87	3.20	2.08	0.31	1.53	2.93	0.00

Note. Similarity ratings range from 0 (*unrelated*) to 4 (*synonyms*). Team-level similarity was operationalized as the median similarity across individual team members.

The individual ratings were aggregated at the team level using the median which, again, is consistent with the procedure that Pathfinder (Schvaneveldt, 2009a) uses to aggregate similarity ratings from multiple raters. Table 8 presents the individual- and team-level similarity ratings for the 15 concept pairs utilized during the mental model assessment. As can be seen in Table 8 differences between team and individual similarity ratings are generally small and non-significant.

Hypothesis 3 stated that the relationship between similarity ratings and behaviors would be stronger at the individual level compared to the team level. To test this hypothesis, the correlation between similarity ratings and behavior frequencies was computed for each individual and team (with stronger associations between the 15 similarity ratings-behavior pairs indicating greater consistency between mental models and behaviors). Because individuals were nested in teams, the mean correlation difference between individuals and teams was tested using a paired-samples *t*-test. Results from this analysis indicated that the mean correlation between similarity ratings and behaviors at the individual level ($M = .29$, $SD = .22$) was substantially lower than the team-level mean correlation ($M = .46$, $SD = .19$), $t(158) = -8.45$, $p < .05$, $d = -.83$. Contrary to Hypothesis 3, these results suggest that teams were more effective at implementing their mental models than individuals.

9. DISCUSSION

The purpose of the present study was to examine the relationship between mental models and the acquisition of a complex skill *across* levels of analysis. It was hypothesized that the relationship between mental models and performance would be stronger at the individual level than the team level due to inherent challenges in the process of integrating team member's knowledge during the execution of team tasks. Contrary to this expectation, the magnitude of this relationship was similar across levels of analysis. In addition, it was hypothesized that declarative knowledge would be a stronger predictor of performance at the individual level compared to the team level of analysis. The present results successfully replicated Chen, Thomas, and Wallace's (2005) findings on the relationship between declarative knowledge and performance. Specifically, results demonstrated stronger associations between declarative knowledge and performance at the individual level compared to the team level of analysis. Finally, the association between similarity ratings and behaviors was substantially stronger among teams than individuals.

Multilevel studies have potential implications for team selection and team composition. Stronger associations at the individual level compared to the team level, would indicate the utility of the selection system will lessen when team performance is the criterion. On the other hand, stronger associations at the team level suggest the presence of synergistic performance effects occurring at the team level (Larson, 2010). The results of the present study indicate *similar* relationships between mental model and performance across levels which suggests that individual mental model assessments serve equally well to predict individual and team performance, a finding that has direct implications for team selection and composition.

Mental model accuracy and declarative knowledge were valid predictors of *individual* performance. Subsequent analyses conducted at the individual level demonstrated that mental model accuracy accounted for variance beyond the effect of declarative knowledge, which is also consistent with previous findings in this domain (e.g., Cooke, Kiekel, & Helm, 2001; Kraiger & Salas, 1993). However, as previously noted, declarative knowledge was not a valid predictor of *team* performance. There are two plausible explanations for this finding. First, the relationship between team declarative knowledge and team performance was lower at Time 4 compared to Time 2 ($r = .41$, $p < .05$) and Time 3 ($r = .39$, $p < .05$) which may reflect a shift in the relationships between determinants of team performance across trials, which is reminiscent of individual skill acquisition models (e.g., Ackerman, 1988; Anderson, 1982). Specifically, by the end of training, teams may have been transitioning from controlled to automatic processing strategies, which would explain the observed decrease in declarative knowledge's predictive validity at the end of training. However, because this pattern of result was observed only at the team level, this change may be attributed to the acquisition of (automatized) *teamwork* skills rather than taskwork knowledge. Thus, the ability of team members to work together effectively may be more important at this stage than in previous stages in terms of differentiating between high- and low-performing teams.

The second explanation for the lack of a relationship between declarative knowledge and team performance is that the method for aggregating declarative knowledge at the team level was inadequate. Consequently, alternative operationalizations of declarative knowledge as a team level construct were assessed. Although not anticipated, the maximum yielded the highest declarative knowledge-team performance association (compared to the minimum). Thus, just as the effective operationalization of team mental models depends on task demands (Stout et al., 1996), the best operationalization of declarative knowledge depends on the nature of the task and the manner in which team members interact during task performance. Because teams were able to freely communicate during the performance episodes, it is plausible that more knowledgeable team members passed on information and subsequently help developed their teammates' knowledge during training. This would not be the case with mental models because the content of mental models cannot be communicated, or at least it would be more difficult to articulate and communicate this knowledge explicitly (e.g., Nisbett & Wilson, 1977). Thus, in hindsight, using the maximum rather than the mean for operationalizing declarative knowledge at the team level appears to be the appropriate composition model for declarative knowledge in situations in which team members can communicate freely.

Recent research examining the processes whereby team mental models affect performance has examined the role of implicit coordination for team performance (Fisher, Bell, Dierdorff, & Belohlav, 2012). Specifically, Fisher et al. demonstrated that implicit coordination mediated the relationship between team mental model similarity and team performance. If mental models and declarative knowledge correspond to implicit and explicit forms of team cognition, then these forms of cognition should exert their influence on team performance by means of explicit and implicit coordination mechanisms, respectively. So, for instance, an examination of the pattern of communication between team members during a performance episode would show that team members with the highest declarative knowledge scores offer explicit guidance to their fellow teammates more often.

Evidence and theory supporting the usefulness of mental models for predicting skill acquisition appears to be closely related to the use of similarity ratings for the elicitation of knowledge. Compared to declarative knowledge measures which are typically assessed via multiple-choice tests, similarity judgments capture individuals' *intuitive* understanding of system function, and thus should be more sensitive to differences between experts and novices. This view is consistent with empirical findings showing that referent structures obtained via consensus are less predictive of team performance than mechanical aggregations of knowledge structures (e.g., Cooke et al., 2001; Day et al., 2001). As an example, team mental models could have been obtained by having team members rate concept pairs as a group to arrive at a consensus rating (see Cooke et al., 2001). However, both declarative knowledge measures and consensus measures are subject to the same limitation; they tap *explicit* knowledge rather than *implicit* knowledge which makes those approaches less appropriate for measuring expertise. Thus, if high-performing teams require expert team members, then the use of measures of expertise, such as mental models, seem preferable to declarative knowledge measures which focus more on surface features of the task.

Studies comparing declarative knowledge and mental models as predictors of team performance recognize that the declarative knowledge measures may have been too easy and therefore not very sensitive (e.g., Cooke et al., 2001; Kraiger & Salas, 1993). Evidence from the present study may mitigate this concern. As with previous studies, the declarative knowledge measure used in the present study was relatively easy (77 to 80 percent accuracy, not including the baseline measure). In spite of this, declarative knowledge remained a valid predictor of individual performance and a valid predictor of team performance (at least during team missions 1 to 3). Notwithstanding, further evidence is needed to support the use of alternative operationalizations of declarative knowledge as a predictor of team performance (e.g., using the maximum).

The relationship between similarity ratings—obtained during the mental model assessment—and specific behaviors observed during the performance episode was positive at the individual level and the team level. Essentially, stronger associations between concepts obtained via similarity ratings translated into more behaviors consistent with those associations. However, contrary to Hypothesis 3, these associations were lower at the individual level compared to the team level (d = -0.83). This result is important because it demonstrates that the behaviors of teams reflect the cognitive structures that support team functioning, even when these cognitive structures may not necessarily translate into performance *results* which are influenced by other factors that fall outside the cognitive domain.

In summary, the results of the present study further support the usefulness of mental models as training criteria for individual performance as well as team performance. In addition, stronger associations between similarity ratings and behaviors at the team level suggest that teams were more effective than individuals in terms of "translating" their mental models into action.

9.1. Implications for Training

Kozlowski and Salas (1997) considered organizational system factors that influence the effectiveness of training. Specifically, they discussed the importance of vertical transfer, an issue which has been largely neglected by researchers (Salas & Cannon-Bowers, 2001). Gagne (1965) distinguished lateral from vertical transfer. Lateral or horizontal transfer refers to the application of trained skills over a set of situations at the same level of complexity or difficulty. In contrast, vertical transfer is defined as the propagation of individual-level training outcomes to team- and organizational-levels. This distinction implies that the skills acquired as an individual do not necessarily generalize to other situations, such as a team context in which additional interrole behaviors need to be acquired.

Although vertical transfer is defined as an "upward" propagation of skills, vertical transfer may also occur downwardly. For instance, in the active interlocked modeling (AIM) protocol (e.g., Arthur, Day, Bennett, McNelly, & Jordan, 1997), trainees are trained in a team context to perform later as individuals. Specifically, in the AIM protocol each member of a team controls only one aspect of the task (e.g., piloting or shooting), alternating roles with their partner(s) between sessions (e.g., Arthur et al., 1997; Shebilske, Regian, Arthur, & Jordan, 1992). Previous studies comparing the AIM protocol with standard individual training have repeatedly demonstrated the effectiveness of the former in terms of effective use of time and resources. In addition to variables such as group facilitation, is has been posited that the AIM protocol engenders cognitive characteristics that facilitate individual learning. Specifically, consonant with Kanfer and Ackerman (1989), Arthur et al. (1997) proposed that “requiring trainees to be responsible for only [a portion] of the task . . . may free up trainees cognitive resources” (p. 784) that are very important during initial acquisition. Thus, although evidence of upward vertical transfer is scarce, there is some evidence of downward transfer—that is, cooperative learning—in that teams can serve as an effective context to acquire individual skills.

Kozlowski et al. (2000) proposed that the contribution of training to organizational outcomes “will be enhanced to the extent that the training system is aligned with the form of vertical transfer as a composition or compilation process” (p. 179). For example, compilation-based outcomes depend on the successful integration of the unique knowledge and skills that each individual member brings to the team. Thus, training design for compilation-based outcomes should focus on the team level (e.g., teamwork behaviors) and pay special attention to the sequencing of training content. The sequencing of training is critical for compilation-based outcomes because team functioning requires that all team members are proficient in their specific roles.

The complexity and team interrelatedness of the tasks typically encountered within highly specialized teams (e.g., medical emergency teams, disaster response teams, law enforcement special units) necessitates training that recognizes the interdependent nature of the tasks performed by those teams. For these tasks, it is critical that individual members master their specific roles as well as the teamwork behaviors for optimal execution of their tasks as a team. In contrast, composition-based outcomes require a somewhat different training approach. Sales representatives, for instance, can be trained individually and the order in which individuals are trained is not critical for optimal team performance because failure of one team member to meet the sales performance goals should not influence the performance of other team members.

In the present study, team mental model accuracy was operationalized using the mean of individual team members' mental model accuracy. This determination was based on the premise that team members need to master their own role and also understand the task as a whole in order to anticipate the needs of other team members and solicit their assistance when needed. Consistent with this reasoning, mean team mental model accuracy was a valid predictor of team performance.

Understanding the roles, responsibilities, and information needs of other team members is critical to the facilitation of team interaction. This type of knowledge is referred to as interpositional knowledge (Cannon-Bowers, Salas, Blickensderfer, & Bowers, 1998; Volpe, Cannon-Bowers, Salas, & Spector, 1996).[9] Volpe et al. posited that interpositional knowledge is critical to team functioning because "it allows team members to anticipate the task needs of fellow team members, thus allowing enhanced

---

[9] This type of knowledge has been sometimes subsumed as an aspect of teamwork knowledge (e.g., Mathieu et al., 2000).

coordination with minimal communication requirement” (p. 88). This explanation is nearly identical to the mechanism invoked for explaining the influence of mental models on team performance. In fact, it is not unreasonable to posit that the generalizability of mental models across levels of analysis may be due to the inclusion of role *and* interpositional knowledge. However, the inclusion of both types of knowledge may have resulted in a *hybrid* model of emergence, an issue that is discussed below.

The minimum mental model accuracy was also predictive of team performance. The fact that the mean *and* the minimum were predictive of team performance is not surprising considering that the two operationalizations were highly correlated ($r = .76$, $p < .05$). Paradoxically, this additional finding suggests a hybrid model of emergence which encompasses both compositional and compilational elements. In hindsight, describing the performance task used in the present study as *highly* interdependent, rather than *purely* interdependent, is accurate in the sense that although several tasks require collaboration between team members, some tasks can be accomplished individually, without the assistance of other team members. Furthermore, it is possible to complete almost every task using only two members of the team without necessitating the help of the third member of the team. (The only exception to this is shutting off valves which can only be accomplished by the oilrig worker; thus, the helicopter and the boat working together would be able to complete every task by themselves except shutting off valves.)

From a training standpoint, if the team task is comprised of both compositional and compilational elements, then team performance will likely benefit from using a combination of individual and team training. If team functioning is not entirely dependent on its least capable member (as in conjunctive tasks; Steiner, 1972), then novice members can be safely incorporated to the team even if they have not yet fully developed the necessary knowledge and skills. In addition, novice members may continue their training while gaining experience as members of their new team. As an example, having a medical intern assisting in an operating room may have an impact on compilational outcomes such as length of the surgery, but it should not have an impact on the success of the operation as a whole. However, faced with a medical emergency, the surgical team may decide not to have an intern in the operating room.

Thus, the validity of the mean and the minimum to operationalize team mental model accuracy is consistent with the coexistence of individual- and team-based outcomes in the present study's performance task. Although a single individual could not successfully accomplish all the required tasks, a low ability team member would not severely disrupt team functioning either. Thus, the present study illustrates the importance of matching the cognitive processes assessed with the cognitive processes underlying performance. Specifically, the inclusion of role knowledge *and* interpositional knowledge in the mental model assessment is consistent with the task demands of the simulation which included both compilational- and compositional-based outcomes.

9.2. Limitations and Future Directions

From a multilevel perspective, it is important to reiterate that inferences from individual-level data cannot be assumed to hold at team- or organizational-levels. For instance, it is questionable to assess the effectiveness of selection systems at the individual level and then generalize these results to higher-level systems (Schneider et al., 2000). One of the objectives of the present study was to more precisely estimate the magnitude of the difference between individual and team mental models as predictors of performance. Although the present results successfully replicated previous findings within each level of analysis (i.e., individual and team levels), the observed difference between team and individual level coefficients were inconsistent with similar estimates obtained from single-level studies. This discrepancy can be attributed to the use of a multilevel design to perform a *direct* comparison between estimates across levels which was deemed a better approach than comparing coefficients from single-level studies. Nonetheless, other methodological characteristics of the study may also account for these discrepancies. Specifically, whereas the present study used a command-and-control task which is often used as performance criteria for studying mental models in the *team* training literature, research on *individual* mental models have been conducted using other tasks and criteria (e.g., exam performance [e.g., Acton et al., 1994, Goldsmith, et al., 1991]; computer programming [Davis & Yi, 2004]) which do not include the behavioral component typically encountered in command-and-control simulations. The mechanism whereby mental models predict performance of command-and-control tasks may differ from the mechanism whereby mental models predict performance of decision making tasks, which in turn may influence the extent to which individual and team mental models predict performance. Consequently, an obvious extension of the present work would be to replicate these findings comparing the validity of individual and team mental models using diverse tasks.

Differences in the operationalization of team performance may also contribute to further understanding team mental models and the processes whereby they influence team performance. Larson (2010) posited that teams offer the potential for *synergistic* performance which is a different way to conceptualize (and operationalize) team performance. Team synergy is demonstrated if a team is capable of performing better than the sum of its individual team members working independently. Specifically, if the team achieves a score greater than the sum of its individual team members performing the same task alone, then this would constitute synergistic performance. It would be informative to establish a criterion for team performance that is aligned with the synergistic aspect of team performance. Establishing a relationship between team mental models and synergistic performance—a more stringent criteria for team performance—would suggest that team mental models *facilitate* team performance by enabling teams to do more than their individual team members working independently. Furthermore, evidence from the present study suggests that teams engage in considerably more behaviors than individuals, and that the behaviors of the team are more aligned with team mental models than individual behaviors with individual mental models.

Whereas the meaning of the link between concepts in a knowledge structure is typically ambiguous or uninterpretable, the present results indicate that the links between the specified concepts may be interpreted, at least in part, as event sequences—who does what and with whom. At the same time, it is important to acknowledge that the target behaviors of the present study are only one of several possible manifestations of mental models during team performance. Consequently, one limitation of the present study is that the behaviors coded represent only a subset of all the possible links that could be evaluated using the information obtained from the mental model assessment. The decision to examine the specified set of 15 behaviors was necessary to limit the study to those behaviors that could be objectively coded, thus eliminating the need for using subjective ratings. In other words, there was a trade-off between objectively scoring individual and team behaviors and representing the criterion space thoroughly.

Results from the present study showed that the operationalization of declarative knowledge is critical for predicting team performance—using the maximum declarative knowledge seems more predictive of team performance than the mean or the minimum declarative knowledge. To explain these results it was argued that team members with relatively high declarative knowledge may communicate with their teammates and provide them critical information to perform their tasks. There are at least two ways in which future research could investigate this issue further. First, communication data could be analyze to identify the flow and content of information between team members. The expectation would be that more knowledgeable team members would pass on more work-related task information compared to other team members. However, this could also be tested experimentally by manipulating the implicit coordination requirements of the task. Specifically, teams could be assigned to a free communication condition and another condition in which team members are not allowed to communicate freely (no-communication condition). Team declarative knowledge is expected to influence team performance in the free communication condition but not in the no-communication condition. Comparing the validity of declarative knowledge and mental models across conditions would clarify the role and impact of explicit versus implicit cognition on team performance.

Previous work has found that taskwork knowledge assessed immediately after *individual* training predicts team performance, such that individuals with quality mental models tend to be in high-performing teams (Cooke et al., 2001). Results from Cooke et al. are not surprising given that the training delivered in her study included both role and interpositional knowledge. Thus, in line with the previous discussion, the relevance of including both role and interpositional knowledge as part of the mental model assessment for predicting team outcomes cannot be overstated. To further demonstrate the utility of interpositional knowledge for measuring team mental models, it would be useful to partial out the effects of role and interpositional knowledge and evaluate the extent to which each of them predicts team performance independently. For instance, even if team members understand the roles and tasks of their teammates—and they do so accurately—differences in specific role knowledge may lead to dissimilar mental models. To put it in a nutshell, it is critical to focus on *what needs to be shared* among team members to maximize the utility of team mental models.

10. CONCLUSION

A review of previous research suggested that mental models tend to display higher validities for predicting individual-level outcomes than team-level outcomes. However, prior work in this domain has been conducted using single-level studies which precludes a direct comparison between estimates across levels of analysis. Differences between the previous findings and the present study's results illustrate the need to conduct multilevel studies to obtain more precise estimates of these relationships.

The findings of the present study have direct implications for selection and training. For the design of selection systems, because composition-based outcomes are based on additive models, selection decisions depend on establishing a specified level of performance for the group. Once the performance level for the group is established, then one could literally calculate how many individuals would be needed to produce the desired level of output. However, compositional-based outcomes require the integration of team members' unique knowledge, skills, and abilities (KSAs). In such situations, selection systems need to be designed in accordance with the compilational nature of the higher level outcome. However, in addition to selecting individuals with complementary KSAs, the full potential of teams will also depend on the implementation of appropriate team behaviors which highlights the criticality of team training for compilational-based outcomes. From a training perspective, for compilational-based outcomes, the development of individual- and team-level skills will benefit team performance.

Studies have shown that the acquisition of teamwork knowledge is facilitated by previous taskwork knowledge (Cooke et al., 2003). Results from the present study showed stronger associations between declarative knowledge and performance at the individual level compared to the team level. Concurrently, the relationship between mental models and performance was similar across levels. In conjunction, these results clarify the role of team cognitions for team performance. Whereas declarative knowledge measures are critical for initial team performance, team mental models are critical for differentiating between low- and high-performing teams at later developmental stages. In other words, the adequacy of team cognition measures is dependent on the developmental phase of the team. Related to this, yet another contribution of this study is to provide a better understanding of the sequencing of taskwork and teamwork knowledge acquisition.

Accurate individual mental models can serve equally well as a training criteria for individual- *and* team-based outcomes provided that they capture both role and interpositional knowledge. Because the validity of mental models for team performance is based on the premise that mental models allow team members to anticipate the needs and actions of other team members (e.g., Mohammed & Dumville, 2001), broadening the scope of mental model assessment is critical for using individual mental models (or *aggregates* of individual mental models) for predicting team outcomes. This may be particularly important in situations when the opportunity to perform the task in conjunction with other team members is limited, or when the training team is different from the performing team. In such situations, mental models may be used appropriately as an individual-level training outcome and as a predictor of subsequent team performance.

REFERENCES

Ackerman, P. L. (1988). Determinants of individual differences during skill acquisition: Cognitive abilities and information processing. *Journal of Experimental Psychology*, *117*, 288-318.

Acton, W. H., Johnson, P. J., & Goldsmith, T. E. (1994). Structural knowledge assessment: Comparison of referent structures. *Journal of Educational Psychology*, *86*, 303-311. doi: 10.1037/0022-0663.86.2.303.

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review, 89*, 369-406. doi: 10.1037/0033-295X.89.4.369

Arthur, W. Jr., Edwards, B., Bell, S., Villado, A., & Bennett, W. Jr. (2005). Team task analysis: Identifying tasks and jobs that are team based. *Human Factors*, *47*, 654–669. doi: 10.1518/001872005774860087

Arthur, W. Jr., Glaze, R. M., Bhupatkar, A., Villado, A., Bennett, W. Jr., & Rowe, L. J. (2012). Team task analysis: Differentiating between tasks using team relatedness and team workflow as metrics of team task interdependence. *Human Factors*, *54*, 277-295. doi: 10.1177/0018720811435234

Arthur, W. Jr., Naber, A. M., Jarrett, S. M., Glaze, R. M., Schurig, I., McDonald, J. N., & Muñoz, G. J. (2011). *Game-based assessment for team training and empirical research: The development of a DDD-based research tool.* Contract No. FA8650-05-D-6502, L-3 Communications Corporation, Link Training and Simulations Systems, Mesa, AZ; and Department of Defense, USAF Air Force

Research Laboratory (AFRL), Warfighter Training Research Division, Mesa, AZ.

Baker, D. P., & Salas, E. (1996) Analyzing team performance: In the eye of the beholder? *Military Psychology*, *49*, 909–936. doi: 10.1207/s15327876mp0803_7

Beus, J. M., Muñoz, G. J., Arthur, W. Jr., & Payne, S. C. (2013). *A multilevel construct validation of safety climate*. Paper presented at the 2013 Academy of Management Meeting, Orlando, FL.

Bliese (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 349-381). San Francisco: Jossey-Bass.

Borgatti, S. P., Everett, M. G., & Freeman, L. C. (2002). *Ucinet for Windows: Software for social network analysis*. Harvard, MA: Analytic Technologies.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models*. Thousand Oaks, CA: Sage.

Campbell, J. P., Dunnette, M. D., Lawler, E. E., & Weick, K. E. (1970). *Managerial behavior, performance, and effectiveness*. New York: McGraw-Hill.

Cascio, W. F., & Aguinis, H. (2011). *Applied psychology in human resource management*. New York: Prentice Hall.

Cannon-Bowers, J. A., Salas, E., & Converse, S. (1993). Shared mental models in expert team decision making. In J. A. Cannon-Bowers, E. Salas, & S. Converse (Eds.), *Individual and group decision making: current issues* (pp. 221-246). Hillsdale, NJ, England: Lawrence Erlbaum Associates.

Cannon-Bowers, J. A., & Salas, E. (2001). Reflections on shared cognitions. *Journal of Organizational Behavior*, *22*, 195-202. doi: 10.1002/job.82

Cannon-Bowers, J. A., Salas, E., Blickensderfer, E., & Bowers, C. A. (1998). The impact of cross-training and workload on team functioning: A replication and extension of initial findings. *Human Factors*, *40*, 92-101. doi: 10.1518/001872098779480550

Carroll, J. D., & Chang, J.-J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. *Psychometrica*, *35*, 283-319. doi: 10.1007-BF02310791

Chan, D. (1998). Functional relations among constructs in the same content domain at different levels of analysis: A typology of composition models. *Journal of Applied Psychology*, *83*, 234-246. doi: 10.1037/0021-9010.83.2.234

Chen, G., Bliese, P. D., & Mathieu, J. E. (2005). Conceptual framework and statistical procedures for delineating and testing multilevel theories of homology. *Organizational Research Methods*, *8*, 375-409. doi: 10.1177/1094428105280056

Chen, G., Mathieu, J. E., & Bliese, P. D. (2005). A framework for conducting a multilevel construct validation. In F. J. Yammarino & F. Dansereau (Eds.), *Research in multi- level issues: Multilevel issues in organizational behavior and processes* (Vol. 3, pp. 273-303). Oxford, UK: Elsevier. doi: 10.1016/S1475-9144(04)03013-9

Chen, G., Thomas, B., & Wallace, J. C. (2005). A multilevel examination of the relationships among training outcomes, mediating regulatory processes, and adaptive performance. *Journal of Applied Psychology*, *90*, 827-841. doi: 10.1037/0021-9010.90.5.827

Chi, M. T. H., Glaser, R., & Farr, M. J. (1988). *The nature of expertise*. Hillsdale, NJ, England: Lawrence Erlbaum Associates.

Cooke, N. J., Kiekel, P. A, & Helm, E. (2001). Measuring team knowledge during skill acquisition of a complex task. *International Journal of Cognitive Ergonomics*, *5*, 297-315. doi: 10.1207/S15327566IJCE0503_10

Cooke, N. J., Kiekel, P. A., Salas, E., Stout, R., Bowers, C., & Cannon-Bowers, J. (2003). Measuring team knowledge: A window to the cognitive underpinnings of team performance. *Group Dynamics: Theory, Research, and Practice*, *7*, 179-199. doi: 10.1037/1089-2699.7.3.179

Cooke, N. M, & McDonald, J. E. (1986). A formal methodology for acquiring and representing experrt knowledge. *Proceedings of the IEEE*, *74*, 1422-1430. doi: 10.1109/PROC.1986.13643

Cooke, N. J., & Rowe, A. L. (1993). *An approach to on-line assessment and diagnosis of student troubleshooting knowledge* (Air Force Technical Paper, AL-TP-1993-0004). Retrieved from Human Resources Directorate, Manpower and Personnel Research Division website: http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA262289&Location=U2&doc=GetTRDoc.pdf.

Cooke, N. J., Salas, E., Kiekel, P. A., & Bell, B. (2004). Advances in measuring team cognition. In E. Salas & S. M. Fiore (Eds.), *Team cognition* (pp. 83-106). Washington, D. C.: American Psychological Association.

Dansereau, F., Alutto, J., & Yammarino, F. J. (1984). *Theory testing in organizational behavior: The variant approach*. Englewoods Cliffs, NJ: Prentice Hall.

Dansereau, F., & Yammarino, F. J. (2000). Within and between analysis: The variant paradigm as an underlying approach to theory building and testing. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 3-90). San Francisco: Jossey-Bass.

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*, 191-206.

Davis, F. D., & Yi, M. Y. (2004) Improving computer skill training: Behavior modeling, symbolic mental rehearsal, and the role of knowledge structures. *Journal of Applied Psychology*, *89*, 509-523. doi: 10.1037/0021-9010.89.3.509

Day, E. A., Arthur, W. Jr., & Gettman, D. (2001). Knowledge structures and the acquisition of a complex skill. *Journal of Applied Psychology*, *86*, 1022-1033. doi: 10.1037//0021-9010.86.5.1022

Day, E. A., Arthur, W. Jr., Miyashiro, B., Edwards, B., Tubré, T., & Tubré, A. H. (2004). Criterion-related validity of statistical operationalizations of group general cognitive ability as a function of task type: Comparing the mean, maximum, and minimum. *Journal of Applied Social Psychology*, *34*, 1521-1549. doi: 10.1111/j.1559-1816.2004.tb02017.x

DeChurch, L. A., & Mesmer-Magnus, J. R. (2010a). Measuring shared team mental models: A meta-analyses. *Group dynamics: Theory, research, and practice*, *14*, 1-14. doi: 1037/a0017455

DeChurch, L. A., & Mesmer-Magnus, J. R. (2010b). The cognitive underpinnings of effective teamwork: A meta-analysis. *Journal of Applied Psychology*, *95*, 32-53. doi: 10.1037/a0017328

Devine, D. J., & Philips, J. L. (2001). Do smarter teams do better: A meta-analysis of cognitive ability and team performance. *Small Group Research*, *32*, 507-532. doi: 10.1177/104649640103200501

Edwards, B. E., Day, E. A., Arthur, W. Jr., & Bell, S. T. (2006). Relationships among team ability composition, team mental models, and team performance. *Journal of Applied Psychology*, *91*, 727-736. doi: 10.1037//0021-9010.91.3.727

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavioral Research Methods*, *41*, 1149-1160. doi: 10.3758/BRM.41.4.1149

Fisher, D. M., Bell, S. T., Dierdorff, E. C., & Belohlav, J. A. (2012). Face personality and surface-level diversity as team mental model antecedents: Implications for implicit coordination. *Journal of Applied Psychology*, *97*, 825-841. doi: 10.1037/a0027851

Glaser, R. & Chi, M. T. H. (1988). Overview. In M. T. H. Chi, R. Glase, & M. J. Farr (Eds.), *The nature of expertise* (pp. XV-XXVIII). Hillsdale, NJ, England: Lawrence Erlbaum Associates.

Goldsmith, T. E., & Davenport, D. M. (1990). Assessing structural similarity of graphs. In R. W. Schvaneveldt (Ed.), *Pathfinder associative networks: Studies in knowledge organization* (pp. 75-87). New Jersey: Ablex.

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

Goldsmith, T. E., Johnson, P. J., & Acton, W. H. (1991). Assessing structural knowledge. *Journal of Educational Psychology*, *83*, 88-96. doi: 10.1037/0022-0663.83.1.88

Gully, S. M., Incalcaterra, K. A., Joshi, A., & Beaubien, J. M. (2002). A meta-analysis of team-efficacy, potency, and performance: Interdependence and level of analysis as moderators of observed relationships. *Journal of Applied Psychology*, *87*, 819-832. doi: 10.1037//0021-9010.87.5.819

Hinsz, V. B., Tindale, R. S., & Vollrath, D. A. (1997). The emerging conceptualization of groups as information processors. *Psychological Bulletin*, *121*, 43-64. doi: 10.1037/0033-2909.121.1.43

House, R., Rousseau, D. M., & Thomas-Hunt, M. (1995). The meso paradigm: A framework for integration of micro and macro organizational behavior. In L. L.

Cummings & B. Staw (Eds.), *Research in organizational behavior* (Vol. 17, pp. 71-114). Greenwich, CT: JAI Press.

Hunter, J. E. (1986). Cognitive ability, cognitive demands, job knowledge, and job performance. *Journal of Vocational Behavior, 29*, 340-362. doi: 2048/10.1016/0001-8791(86)90013-8

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage.

Ilgen, D. R., Hollenbeck, J. R., Johnson, M., & Dustin, J. (2005). Teams in organizations: From input-process-output models to IMOI models. *Annual Review*, *56*, 517-543. doi: 10.1146/annurev.psych.56.091103.070250

Johnson-Laird, P. N. (1983). *Mental models* (6th ed.). Cambridge, MA: Harvard University Press.

Kirkpatrick, D. L. (1994). *Evaluating training programs: The four levels*. San Francisco: Berrett-Koehler Publishers, Inc.

Klein, K. J., Dansereau, F., & Hall, R. J. (1994). Level issues in theory development, data collection, and analysis. *Academy of Management Review*, *19*, 195-229.

Klimoski, R., & Mohammed, S. (1994). Team mental model: Construct or metaphor? *Journal of Management*, *20*, 403-437. doi: 10.1177/014920639402000206

Kozlowski, S. W. J., & Bell, B. S. (2003). Work groups and teams in organizations. In W. C. Borman, D. R. Ilgen, & R. J. Klimoski (Eds.), *Handbook of psychology: industrial and organizational psychology* (Vol. 12, pp. 333-375). New York: Wiley.

Kozlowski, S. W. J., Gully, S. M., Salas, E., & Cannon-Bowers, J. A. (1996). Team leadership and development: Theory, principles, and guidelines for training leaders and teams. In M. Beyerlein, D. Johnson, & S. Beyerlein (Eds.). *Advances in interdisciplinary studies of work teams: Team leadership* (Vol. 3, pp. 251-289). Greenwich, CT: JAI Press.

Kozlowski, S. W., & Ilgen, D. R. (2006). Enhancing the effectiveness of work groups and teams. *Psychological Science in the Public Interest*, *7*, 77-124. doi: 10.1111/j.1529-1006.2006.00030.x

Kozlowski, S. W. J., Brown, K. G., Weissbein, D. A., Cannon-Bowers, J. A., & Salas, E. (2000). A multilevel approach to training effectiveness: Enhancing horizontal and vertical transfer. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 3-90). San Francisco: Jossey-Bass.

Kozlowski, S. W. J., & Klein, K. J. (2000). A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 3-90). San Francisco: Jossey-Bass.

Kozlowski, S. W. J., & Salas, E. (1997). An organizational systems approach for the implementation and transfer of training. In J. K. Ford, S. W. J. Kozlowski, K. Kraiger, E. Salas, & M. Teachout (Eds.), *Improving training effectiveness in work organizations* (pp. 247-287). Mahwah, NJ: Lawrence Erlbaum Associates.

Kraiger, K., Ford, J. K., & Salas, E. (1993). Application of cognitive, skill-based, and affective theories of learning outcomes to new methods of training evaluation. *Journal of Applied Psychology, 78*, 311-328. doi: 10.1037/0021-9010.78.2.311

Kraiger, K., & Salas, E. (1993, April). *An empirical test of two cognitively based measures of learning during training*. Paper presented at the 23rd Annual Conference of the Society for Industrial and Organizational Psychology, San Francisco, CA.

Kraiger, K., Salas, E., & Cannon-Bowers, J. A. (1995). Measuring knowledge organization as a method for assessing learning during training. *Human Factors*, *37*, 804-816. doi: 10.1518/001872095778995535

Larson, J. R. Jr. (2010). *In search of synergy in small group performance.* New York, NY: Psychology Press.

Lepine, J. A., Hollenbeck, J. R., Ilgen, D., & Hedlung, J. (1997). Effects of individual differences on the performance of hierarchical decision-making teams: Much more than g. *Journal of Applied Psychology*, *82*, 803-811. doi: 10.1037/0021-9010.82.5.803

Lim, B.-C., & Klein, K. J. (2006). Team mental models and team performance: A field study of the effects of team mental model similarity and accuracy. *Journal of Organizational Behavior*, *27*, 403-418. doi: 10.1002/job.387

Lindsley, D. H., Brass, D. J., & Thomas, J. B. (1995). Efficacy-performance spirals: A multilevel perspective. *Academy of Management Review*, *20*, 645-678. doi: 10.5465/AMR.1995.9508080333

Loevinger, J. (1957). Objective tests as instruments of psychological theory [Monograph]. *Psychological Reports*, *3*, 635-694 (Pt. 9).

Mathieu, J. E., Heffner, T. S., Goodwin, G. F., Cannon-Bowers, J. A., & Salas, E. (2005). Scaling the quality of teammates' mental models: Equifinality and normative comparisons. *Journal of Organizational Behavior*, *26*, 37-56. doi: 10.1002/job.296

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*, 273-283. doi: 10.1037//0021-9010.85.2.273

Mathieu, J., Maynard, M. T., Rapp, T. L., & Gilson, L. (2008). Team effectiveness 1997-2007: A review of recent advancements and a glimpse into the future. *Journal of Management*, *34*, 410-476. doi: 10.1177/0149206308316061

Marks, M. A., Sabella, M. J., Burke, C. S., & Zaccaro, S. (2002). The impact of cross-training on team effectiveness. *Journal of Applied Psychology*, *87*, 3-13.

Marks, M. A., Zaccaro, S. J., & Mathieu, J. E. (2000). Performance implications of leader briefings and team-interaction training for team adaptation to novel environments. *Journal of Applied Psychology*, *85*, 971-986. doi: 10.1037//0021-9010.85.6.971

McDonald, J., & Muñoz, G. J. (2013). *Mental model complexity moderates the relationship between mental models and learning outcomes: A meta-analysis.* Manuscript in preparation.

McIntyre, R. M., & Salas, E. (1995). Measuring and managing for team performance: Lessons from complex environments. In R. A. Guzzo & E. Salas (Eds.), *Team effectiveness and decision making in organizations* (pp. 9-45). San Francisco: Jossey-Bass Publishers.

Meng, X.-L., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin*, *111*, 172-175. doi: 10.1037/0033-2909.111.1.172

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741-749.

Moreland, R. L., Argote, L., & Krishnan, R. (1996). Socially shared cognition at work: Transactive memory and group performance. In J. L. Nye & A. M. Brower (Eds.). *What's social about social cognition? Research on socially shared cognition in small groups* (pp. 57-84). Thousand Oaks, CA: Sage.

Mohammed, S., & Dumville, B. C. (2001). Team mental models in a team knowledge framework: Expanding theory and measurement across disciplinary boundaries. *Journal of Organizational Behavior*, *22*, 89-106. doi: 10.1002/job.86

Mohammed, S., Ferzandi, L., & Hamilton, K. (2010). Metaphor no more: A 15-year review of the team mental model construct. *Journal of Management*, *36*, 876-910. doi: 10.1177/0149206309356804

Mohammed, S., Klimoski, R., & Rentsch, J. R. (2000). The measurement of team mental models: We have no shared schema. *Organizational Research Methods*, *3*, 123-165. doi:

Morgeson, F. P., & Hofmann, D. A. (1999). The structure and function of collective constructs: Implications for multilevel research and theory development. *Academy of Management Review*, *24*, 249-265. doi: 10.5465/AMR.1999.1893935

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, *22*, 376-398. doi: 10.1177/0049124194022003006

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*, 231-259. 10.1037/0033-295X.84.3.231

Ree, M. J., Carretta, T. R., & Teachout, M. S. (1995). Role of ability and prior job knowledge in complex training performance. *Journal of Applied Psychology*, *80*, 721-730. doi: 10.1037/0021-9010.80.6.721

Roberts, K. H., Hulin, C. L., & Rousseau, D. M. (1978). *Developing an interdisciplinary science of organizations*. San Francisco: Josssey-Bass.

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*, 349-363. 10.1037/0033-2909.100.3.349

Rousseau, D. M. (1985). Issues of level in organizational research: Multi-level and cross-level perspectives. *Research in Organizational Behavior*, *7*, 1-37.

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*, 31-47.

Salas, E., DiazGranados, D., Klein, C., Burke, C. S., Stagl, K. C., Goodwin, G. F., & Halpin, S. M. (2008). Does team training improve team performance? A meta-analysis. *Human Factors*, *50*, 903-933. doi: 10.1518/001872008X375009

Salas, E., Sims, D. E., & Burke, C. S. (2005). Is there a "Big Five" in teamwork? *Small Group Research*, *36*, 555-599. doi: 10.1177/1046496405277134

Schneider, B., Smith, D. B., & Sipe, W. P. (2000). Personnel selection psychology: Multilevel considerations. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 3-90). San Francisco: Jossey-Bass.

Schuelke, M. J., Day, E. A., McEntire, L. E., Boatman, P. R., Espejo Boatman, J., Kowollik, V., et al. (2009). Relating indices of knowledge structure coherence and accuracy to skill-based performance: Is there utility in using a combination of indices? *Journal of Applied Psychology*, *94*, 1076-1085. doi: 10.1037/a0015113

Schvaneveldt, R. W. (2009a). *Pathfinder 6.3* [Computer software and documentation]. Las Cruces, NM: Interlink.

Schvaneveldt, R. W. (2009b). *Target* [Computer software]. Las Cruces, NM: Interlink.

Schvaneveldt, R. W. (1990). *Pathfinder associative networks: Studies in knowlege organization*. New Jersey: Ablex.

Schvaneveldt, R. W., Dearholt, D. W., & Durso, F. T. (1988). Graph theoretic foundations of Pathfinder networks. *Computers & Mathematics with Applications*, *15*, 337-345.

Schvaneveldt, R. W., Durso, F. T., Goldsmith, T. E., Breen, T. J., Cooke, N. M., Tucker, R. G., & De Maio, J. C. (1985). Measuring the structure of expertise. *International Journal of Man-Machine Studies*, *23*, 699-728. doi: 10.1016/S0020-7373(85)80064-X

Shebilske, W. L., Regian, J. W, Arthur, W. Jr., & Jordan, J. A. (1992). A dyadic protocol for training complex skills. Human Factors, 34, 369-374. doi: 10.1177/001872089203400309

Steiner, I. D. (1966). Models for inferring relationships between group size and potential group productivity. *Behavioral Science*, *11*, 273-283. doi: 10.1002/bs.3830110404

Steiner, I. D. (1972). *Group process and productivity*. New York: Academic Press.

Stout, R. J., Cannon-Bowers, J. A., Salas, E., & Milanovich, D. M. (1999). Planning, shared mental models, and coordinated performance: An empirical link is established. *Human Factors*, *41*, 61-71. doi: 10.1518/001872099779577273

Volpe, C. E., Cannon-Bowers, J. A., Salas, E., & Spector, P. E. (1996). The impact of cross-training on team functioning: An empirical investigation. *Human Factors*, *38*, 87-100. doi: 10.1518/001872096778940741

Widaman, K. F. (2000). Testing cross-group and cross-time constraints on parameters using the general linear model. In T. D. Little, K. U., Schnabel, & J. Baumert

(Eds.), *Modeling longitudinal and multilevel data: Practical issues, applied approaches and specific examples* (pp. 163-186). Mahwah, NJ: Lawrence Erlbaum.

Wilson, J. R., & Rutherford, A. (1989). Mental models: Theory and application in human factors. *Human Factors*, *31*, 617-634.

Yammarino, F. J., & Dansereau, F. (Eds.). (2002-2009). *Research in multi-level issues* (Vols. 1-8). Retrieved from http://www.emeraldinsight.com/books.htm?issn=1475-9144

Appendix

Instructions for Mental Model Assessment

Click on the icon in the Toolbar named Target. In this task, you will see several concepts listed on the left hand side of the screen. Each of these concepts will be presented as a focus concept in the bull's-eye of the target on the right hand side.

Your task is to move the concepts that are extremely, largely, moderately related, or synonyms to the target inside the appropriate gradient of the target. To move the concepts, click and drag them to the location in which you wish to drop them. Each concept that you rate must fit into one of the three related categories (extremely related, largely related, moderately related, or synonyms): There is no in between option. Concepts that are less related or unrelated should be left in place at the left side. Each concept will earn a score based on its distance from the concept at the center of the box; you do not need to think about the distances between concepts other than the one at the center of the box. You can change your mind about a concept by moving it again. Once you have moved all the concepts onto the target, click NEXT to proceed to the next target. Although we expect you to take this task seriously, do not spend too much time deliberating.

YOU SHOULD BASE YOUR RELATEDNESS JUDGMENTS ON HOW THE CONCEPTS ARE RELATED FOR SUCCESSFUL COMPLETION OF YOUR TASKS IN KODIAK ISLAND. Here are the concepts:

- Boats
- Helicopters
- Oilrig workers
- Strategic Positioning
- Clearing the path
- Healing survivors
- Putting out fires
- Maneuvering
- Survivor Capacity
- Requesting assistance
- Finding a route
- Coordinating clicks
- Time management
- Shutting off valves
- Picking up survivors

Do you have any questions before you begin? You may now begin the Target measure.