Abstract

SURFACE, ERIC ALAN. An Integration of the Training Evaluation and Job Performance Modeling Literatures: Confirming BE KNOW DO with United States Army Special Forces Training Data. (Under the Direction of Dr. Mark Alan Wilson.)

Training data from 1441 graduates of the U.S. Army Special Forces Qualifications Course (SFQC) offered an opportunity to test a multidimensional model of training performance. A three-factor model based partially on the Kraiger, Ford, & Salas (1993) framework was operationalized with level-of-performance training criteria (Sackett & Mullen, 1993) and successfully confirmed using confirmatory factor analysis (CFA). The model utilized the BE KNOW DO terminology from the U.S. Army's leadership model to describe the cognitive, skill-based, and affective training outcome factors. Several alternative models were tested and found not to be identified, suggesting the Kraiger et al. (1993) version provided the best description of performance. Additionally, the BE KNOW DO model was successfully confirmed for two individual phases of the SFQC training separated in time. Therefore, structural equation modeling (SEM) analyses were conducted to determine the relationship of similar constructs over time. Each construct from the initial training phase was found to predict its counterpart in the later phase. The degree of relationship varied for the constructs, suggesting some were more influenced by time and situation. Two performance modeling issues—the specificity of performance constructs and the impact of overfitting a model to the idiosyncratic characteristics of the initial sample on cross-validation—were investigated as well. Results related to the specificity of modeling performance content were inconclusive. Both the one- and three-factor construct models failed to provide adequate fit. The over-modified model provided a worse fit upon cross-validation in 11 out of 12 cases, demonstrating the importance of cross-validating modified

models. An integration of the training evaluation and job performance literatures is presented and serves as the rationale for proposing a general three-factor performance model. The idea that all performance can be described in terms of three factors regardless of the context, content, situation, measurement method, or performance level should be investigated. Future directions for practice and research are discussed.

An Integration of the Training Evaluation and Job Performance Modeling Literatures: Confirming BE KNOW DO with United States Army Special Forces Training Data

By

Eric A. Surface

A dissertation submitted to the Graduate Faculty of North Carolina State University in partial fulfillment of the requirements for the Degree of

Doctor of Philosophy

December 6, 2002

Industrial/Organizational & Vocational Psychology

Approved By:

_____
Dr. Mark A. Wilson
Chair of Advisory Committee

_____ _____
Dr. Don W. Drewes Dr. Bert W. Westbrook

_____ _____
Dr. Don L. Martin Dr. Mike G. Sanders

Date filed with the Department: _____

Dedication

I would like to dedicate this research to the members of the United States Army Special Forces, whose diligent and often silent work protects and promotes the interests of the United States and its citizens around the world, and to the men and women at the John F. Kennedy Special Warfare Center and School (JFKSWCS), who ensure that Special Forces soldiers are selected and trained to be the best. I would like to recognize the contribution and sacrifice of 1st Lt. Tallas Tomeny, a SF candidate who died in a training accident during the Robin Sage field training exercise in 2002. In the very next iteration of the course, I participated in Robin Sage as a guerilla soldier at the base camp where Tomeny was stationed. On an outcropping of rock near Robbins, NC, a memorial is carved to him. His sacrifice for our country will not be forgotten. After reading the memorial, participating in Robin Sage, and working with members of this community over the past five years, it has become apparent to me that SF soldiers must possess the highest levels of commitment, self-sacrifice, and professionalism even to complete the training. I have tremendous respect for soldiers who have chosen to walk this path.

Additionally, I would like to dedicate this dissertation to my family, especially my parents, Daniel and Sallie Surface, who have been unwavering in their support, and my maternal grandparents, Ernest and Dorothy Pait, who were both advocates of education and who passed away during this research project.

Biography

Eric Alan Surface was born on September 8, 1970, in Ahoskie, NC. Eric is the oldest son of Daniel and Sallie Surface, and he grew up in Murfreesboro, NC, and Floyd, VA, with a younger sister and brother, Julianne and Brian.

After graduating from Murfreesboro High School as the Salutatorian of its last class, he attended Wake Forest University in Winston-Salem, NC. At Wake Forest, Eric participated in numerous student organizations and leadership activities, including helping to restart the Sigma Nu chapter, to start a monthly publication that covered fraternities and sororities, and to establish a leadership organization. Upon his graduation in 1992 with a BA degree in psychology and a minor in computer science, he took a position as interim director of college relations at Chowan College in his hometown. While at Chowan, he helped the college restore its public image that had been tarnished by a turbulent transition to four-year status. In 1994, Eric entered graduate school in industrial/organizational psychology at East Carolina University (ECU) in Greenville, NC. At ECU, he taught undergraduate psychology courses, worked in the graduate admissions office, and interned at Weyerhaeuser's New Bern Pulp site. At Weyerhaeuser, Eric was given the responsibility of conducting the employee survey and its follow-up activities for the site. In 1996, he was accepted into the Ph.D. program in industrial/organizational and vocational psychology at North Carolina State University. That summer, Eric began working as an intern at Caterpillar's Building Construction Products Division in Clayton, NC, where he worked primarily on employee surveys and organizational development (OD) activities.

1997 was an eventful year. Eric accepted a position as a consortium research fellow with the Army Research Institute's (ARI) field office at Fort Bragg, NC, conducting applied

research with United States Army Special Forces, a position that he has held for over five years. He successfully defended his thesis at ECU and received his MA in psychology. Additionally, in 1997, Caterpillar contacted Eric and asked him to do an employee survey project for them as a consultant. Unable to do the work alone, Eric invited his friend and fellow graduate student, Stephen Ward, to help with the project. They used the project as an opportunity to learn about consulting and business. They formed a consulting partnership, Surface, Ward, & Associates, to handle the business. Both readily admit to making mistakes and learning a tremendous amount. Over the following five years, the partnership did a number of projects on a very part-time basis (i.e., one to two small projects a year) for companies like IBM. Their business was generated totally by referrals. In January 2002, both Stephen and Eric took a hiatus from consulting in order to make progress on personal goals; Eric focused solely on finishing his dissertation research and writing several conference proposals and article submissions.

Since 1997, the fellowship with ARI has provided Eric with a number of interesting opportunities, including collecting data in Germany, participating in Robin Sage as a guerilla soldier, and observing SF training. In 1999, Eric was made an honorary member of Army Special Operations Forces for his work with the 1998 USASOC command climate survey and its follow-up activities. While working with ARI, Eric has worked on a variety of projects, including climate surveys, training and performance research, electronic data collection implementation and research, the development of a competency model and 360-degree assessment instrument for USASOC supervisors, and various survey follow-up interventions. Eric has valued the eclectic and flexible nature of the fellowship and the highly talented and competent people with whom he has had the opportunity to work. He indicated

that working with Special Forces and ARI has been his most rewarding professional experience to date.

During the fall of 2000 and the spring 2001, Eric taught classes as an adjunct professor in the human resources management program at Peace College in Raleigh, NC. This afforded him the opportunity to teach I/O psychology and organizational development to undergraduates. He also supervised a couple of research projects. After leaving Peace, he collaborated on a conference paper with a former student who is interested in attending graduate school in I/O.

Late in his graduate career, Eric realized the value of publishing research work. Although he had always excelled at doing projects, he had never published a research article. Eric believes that the transfer of knowledge advances the science and practice of I/O and that I/O psychologists have a responsibility to publish regardless of their career path. Writing and publishing articles helps demonstrate your intellectual capital regardless of whether you engage in teaching, research, consulting, or organizational practice. In 2002, Eric was a co-author on two papers presented at conferences and the second author of an article accepted by *Personnel Psychology*. His previous publications include two book reviews published in *Personnel Psychology*. At the time of final submission, Eric had several co-authored conference posters accepted for SIOP 2003 and two articles under review, with several submissions planned for the near future. In 2003, Eric plans to continue his research with ARI while preparing for a career in research or academia.

Acknowledgements

I would like to acknowledge the help and support of my advisor, my committee and the personnel at the John F. Kennedy Special Warfare Center and School (JFKSWCS). Additionally, I would like to acknowledge the contributions of personnel from the United States Army Special Operations Command (USASOC), the USASOC Psychological Applications Directorate (PAD), the Army Research Institute for the Behavioral and Social Sciences (ARI), the Consortium of Colleges and Universities of the Greater Metropolitan DC Area, and the Psychology Department at North Carolina State University. A special note of recognition should be given to all those who helped in the Herculean data collection process. I owe a special debt of gratitude to all the individuals who helped me complete this process.

Although there is always a danger of forgetting someone, I would like to recognize the contributions of some specific individuals not only to this dissertation but also to my graduate career. First, I would like to thank all the members of my committee for their service. I would like to thank my advisor and chair, Dr. Mark A. Wilson, for his guidance and patience over the course of my long graduate career. I would like to thank Dr. Mike Sanders for his guidance and mentoring as my ARI supervisor for over five years. I would like to thank Dr. Don Martin for his mentoring and support, especially in the areas of technology and collaboration, and for our many philosophical conversations. I would like to thank Dr. Don Drewes for introducing me to SEM and a new way of thinking about data analysis and for his curiosity and excitement about discovering the “story” the data is telling. I would like to thank Dr. Bert Westbrook, who taught me two classes my first semester at NCSU, for his service on the committee although he was not involved in any Army projects,

for being an example of a genuine southern gentleman, and for overlooking my long breaks during the three-hour vocational psychology seminar class.

Second, I would like to thank the numerous military and civilian personnel from JFKSWCS, the Special Forces Medical Training Battalion (SOMTB), USASOC, and USASOC PAD who helped to make this research possible and a success. Specifically from JFKSWCS, I would like to thank MAJ General Boykin, COL Joe Kilgore, COL Charles King, CSM Ronnie McCan, LTC Nagata, LTC Robert Marrs, MAJ Clifton Poole, MAJ Everheart, MAJ Joyce, MAJ William Banker, CPT Mark Homan, Mrs. Flowers, and Harriet Craven. I would like to extend a special thanks to CPT Showalter, MSG Garner, "MAJ Vincent" and "Cowboy" for my Robin Sage experience. Specifically from SOMTB, I would like to thank Sandra Strickland, Jim Rorke, and Roy Hutchinson. Specifically from USASOC, I would like to thank LT General Doug Brown, LT General Tangeny, MAJ General Toney, COL Stan Florer, Luke Taylor, Gary Barrett, MAJ Gary Kolb, and Sue Rose. Specifically from USASOC PAD, I would like to thank SFC Michael Perkins, LTC Morgan Banks, MAJ Gary Hazlett, COL Larry Lewis, COL Gary Greenfield, LTC Fred "Doc" Brown, Joyce Melvin, Kristin Richmond, Gigi Gill, and Renee.

Third, I would like to recognize some individuals from ARI, the Consortium, and the NCSU Psychology Department for their support and assistance. Specifically from ARI, in addition to Dr. Mike Sanders, I would like to acknowledge Drs. Bob Kilcullen, Mike Rumsey, and Michelle Wisecraver. Specifically from the Consortium, I would like to thank Dr. Robert Ruskin, Julie Waller, and Amy Schaub for their support and assistance over the past 6 years. Dr. Ruskin deserves an additional nod for the guidance he has provided over the years. From the NCSU Psychology department, I would to thank Dr. Don Mershon and

Darnell Johnson for their assistance in the completion of this process. Additionally, I would to thank Drs. Bob Pond, Jim Kalat, and Bill Cunningham, who taught me classes at NCSU.

Fourth, I want to acknowledge all those who helped with the data collection process. I would like to thank Jat Thompson for his help with form development, data checking, and data scanning as well as his friendship and our many philosophical conversations while in transit to Fort Bragg. Additionally, I would like to thank Greg Lemmond, Kemp Ellington, LT Kevin Hosier, Sidonia Shumann, Ursula Mannix, Diane Perkins, SFC Michael Perkins, Dr. Michelle Wisecarver, CPT Rick Banks, Justin Whitener, and Dr. John Viehe for their assistance in the data collection and checking process. Dr. John Viehe gets a special nod for handling the nightmare known as payroll.

Fifth, I would like to acknowledge the contributions of various friends and other individuals. I would like to thank Dr. Heather Lee for covering my classes while I was working on my initial proposal when I was teaching as an adjunct at Peace. I would like to thank the following people for their support and friendship during this two-year-plus process: Stephen Ward, Dr. Lori Foster Thompson, Dr. Erich Dierdorff, Dr. Barbara Grimes, Stephanie Asbeck, Kari Yoshimura, Penny Koommoo, Kemp Ellington, Jat Thompson, Julie Hoffman, Jackson McQuigg, Marieke Pieterman, Matt Shriner, Mike Richty, Scott Bublitz (and Joanna), Tyler Jones, John Head, Scott Hagaman, Greg Lemmond, Mary Schroder, Jennifer Lindberg, Zeke Creech, Kristen Brewer, Krista DeBose, and Dr. Bill Grossnickle. Mike at Global Village deserves a special thank you for letting me sit and work all day, often on one cup of coffee. The majority of this dissertation was written and edited there, although I must admit to doing work at other coffeehouses as well (Helios, Starbucks, Port City, Caribou, Borders, and Barnes and Noble).

Finally, I would like to recognize the contribution of numerous musical artists whose music I played constantly as I created and refined this document. I thank the following artists for making great music for working: the barenaked ladies, The Calling, Sister Hazel, the Dave Mathews Band, Five For Fighting, U2, Norah Jones, the Gin Blossoms, Enya, Stevie Nicks, Fleetwood Mac, the Talking Heads, Queen, the Violent Femmes, the Counting Crows, Tom Petty and the Heartbreakers, emilia dahlin, Lifehouse, Jack Johnson, James Taylor, Mozart, and John Mayer. I must acknowledge that Fleetwood Mac's *The Dance* and the barenaked ladies's *stunt* received the most play with Sister Hazel's …*somewhere more familiar* and the greatest hits albums of James Taylor and Tom Petty and the Heartbreakers following closely.

To all those that I failed to acknowledge, I apologize for the oversight. In the course of a two-year-plus ordeal, it is easier than it should be to forget an important contribution. It does not mean your help and support was not appreciated.

Table of Contents

Table of Tables

Table of Tables (continued)

Table of Figures

Table of Figures (continued)

List of Appendices

List of Appendices (continued)

CHAPTER ONE: INTRODUCTION

This chapter introduces and frames the current research project. First, this research study is framed in terms of its general importance and of the two streams of research—training evaluation and job performance modeling—to be thoroughly reviewed in chapter two. The general research goals and questions are presented. The introduction provides a guide to the nature and structure of the argument being constructed in this dissertation.

Framing the Research

Increased globalization, advances in technology, changes in the nature of work, a shrinking pool of highly-skilled workers, and increasing competition have made organizations more dependent on their human capital for success than ever before in history (Davenport, 1999). As Kozlowski, Brown, Weissbein, Cannon-Bowers, and Salas (2000) expressed it, "organizations are increasingly pressured by technology, political, economic, societal, and cultural changes that are global in scope and impact" (p.157). This necessitates selecting, training, and retaining employees who have the attributes to perform successfully under current requirements and to adapt successfully to dynamic work situations in the future. To do this, a model of job performance is required. Current thinking in psychology recognizes that job performance is a multidimensional construct (Borman, Hanson, & Hedge, 1997; Avery & Murphy, 1998; Schmitt & Chan, 1998; Campbell, 1999; Hough & Oswald, 2000; Viswesvaran & Ones, 2000). Therefore, when selecting or training employees, it is important to understand the job performance dimensions that are most appropriate for the work situation in question. A model of job performance must guide the recruitment, selection, training, and management processes. The model of job performance utilized has great implications for selection and training design (Campbell, 1999). Without a clear

understanding of job performance, selection and training systems may not achieve their desired results. In certain situations, where failure to perform can have life-and-death and/or national security consequences, having selection and training systems based on an accurate and relevant model of job performance becomes an imperative. Occupations like police officers, firefighters, emergency medical technicians, soldiers, intelligence analysts, and hazardous materials handlers would be examples highlighted by the tragedy of September 11, 2001.

Examining previous training evaluation and job performance research can facilitate the objective of aligning selection, training, and performance management processes with an accurate model of job performance. By integrating the two streams of research, our understanding of both training and job performance can be improved, potentially leading to improvements in practice and to a macro-level performance model.

Unlike job performance research, where there is a research literature devoted to modeling the criterion space, a literature devoted solely to modeling training performance for the sake of understanding training performance does not exist. In the research literature, the "training performance" terminology typically refers to measuring simulation or work sample performance during or at the end of training (e.g., Kozlowski, Gully, Brown, Salas, Smith & Nason, 2001). There is no separate training performance modeling literature per se—training research seems to have been more process and outcome oriented. The training criteria space has been specified and examined as a necessary part of training evaluation and training effectiveness research. According to Kraiger, Ford, and Salas (1993), evaluation is conducted to determine "whether training objectives were achieved and whether accomplishment of those objectives results in enhanced performance on the job" (p. 311).

Training effectiveness seeks to discover "why training did or did not achieve its intended outcomes" (p. 311). Training effectiveness is a broader concept and encompasses training evaluation and its criteria. Sackett and Mullen (1993) suggest that evaluation is about answering two different categories of questions—one refers to how much change has occurred as a result of the training, and the other relates to the attainment of a specified level of achievement or performance by each trainee. For the purposes of this research, the training evaluation research literature, with its focus on criteria, provides a more appropriate underpinning.

For years, the Kirkpatrick model provided the predominant model of training criteria in the training evaluation literature (Alliger & Janak, 1989; Kraiger et al., 1993; Salas & Cannon-Bowers, 2001). Kirkpatrick's (1959a, 1959b, 1960a, 1960b, 1967, 1979, 1996) four levels of criteria—reactions, learning, behavior, and results—have been used to guide training evaluation and the measurement of training performance for over 40 years. Current thinking in the training evaluation literature expands Kirkpatrick's framework (Kraiger et al., 1993; Alliger, Tannenbaum, Bennett, Traver, & Shotland, 1997). The measurement of learning criteria—typically defined in terms of a change in declarative knowledge or skill—has evolved beyond the conceptualization in level two of Kirkpatrick's model. Kraiger and his colleagues (1993) indicate learning in training can be classified into three categories of criteria—cognitive, skill-based and affective learning. Kirkpatrick's model and these other frameworks, used primarily to guide training evaluation, suggest that training performance is multidimensional. However, to this researcher's knowledge, no studies have modeled the training performance criterion space comprehensively using confirmatory factor analysis (CFA) or structural equation modeling (SEM) techniques to develop or to test empirically a

multidimensional model of training performance, unlike the modeling conducted with the job performance criterion space. This state of affairs is changing as more researchers adopt the Kraiger et al. (1993) model in their training research. Two recent studies (Colquitt, LaPine, & Noe, 2000; Kozlowski et al., 2001) utilized path analytic techniques to explore the Kraiger et al. (1993) model. Tracey, Hinkin, Tannenbaum, and Mathieu (2001) actually used CFA to test a model that included a partial operationalization of the Kraiger et al. (1993) learning criteria.

Job performance is a multidimensional construct according to the prevailing view in psychology (Borman et al., 1997; Schmitt & Chan, 1998; Campbell, 1999). Several models of performance (e.g., Campbell, McCloy, Oppler, & Sager, 1993) include performance determinants (e.g., declarative knowledge) as well as performance content dimensions (e.g., non-job-specific task proficiency). Performance determinants are theorized to mediate fully the relationship between human attributes and job performance content dimensions. In the Campbell model, declarative knowledge (DK), procedural knowledge and skill (PKS), and motivation (M) are considered the direct determinants of performance—they fully mediate the relationship between the indirect determinants (e.g., abilities, personality, and experience) and job performance (e.g., written and oral communication). Campbell does not provide the only framework. There are other models of job performance (e.g., Motowidlo, Borman & Schmit, 1997) that specify content factors and direct determinants. The performance determinants posited by these models are similar to the three categories of learning criteria proposed by Kraiger and his colleagues (1993). Additionally, three-factor job performance models by Grant (1996) and Wilson and Grant (1997) posit performance content factors that resemble the three Kraiger et al. (1993) categories of learning criteria. Because of these

similarities, this study of training performance draws heavily on the job performance modeling literature.

This research provides an opportunity to integrate training evaluation and job performance research to improve the understanding of both training and job performance and to explore several relevant questions. Given the similarity of the learning criteria categories (Kraiger et al., 1993), the performance determinants in some models (e.g., Campbell et al., 1993), and the performance factors in other models (e.g., Wilson & Grant, 1997), are training performance and job performance separate construct models, are they different cases of a general performance model, or is training performance a component of the job performance model? Can training criteria can be considered the direct determinants of job performance as well as training performance from the *level-of-performance* training evaluation perspective expressed by Sackett and Mullen (1993)?

In development or promotion decisions, Kraiger (1999) points out that job performance is often used as a surrogate for whether the individual has the relevant knowledge, skill, and motivation to accomplish current or future work responsibilities. Can all performance be described in terms of three factors—knowledge, skill, and a third factor that can be thought of as motivation (McCloy, Campbell, & Cudeck, 1994), work habits (Motowidlo et al., 1997), affective criteria (Kraiger et al., 1993), and/or citizenship (Wilson & Grant, 1997) depending on your favorite model— regardless of the performance level or situation? Are these factors content and context independent (e.g., general factors applicable across content) or dependent (e.g., specific factors applicable to a content domain)? Do these factors change over time? Do training performance factors fully mediate (as the performance modeling literature suggests) or partially mediate (as training evaluation research suggests)

the relationship between human attributes and job performance? Although many of these questions, like the mediation question, cannot be addressed by this study or definitively by any one study, this research expands what is known and offers the best potential answers to some of these questions based on the findings. For the questions that are beyond the scope of this study, hopefully, this research will serve as a foundation for future exploration.

It should be noted that the U.S. Army might have specified the underlying model of performance with its BE KNOW DO model of leadership (U.S. Army, 1999). Additionally, the BE terminology may more accurately reflect the nature of the third factor than other alternatives (e.g., motivation)—there is a specific way of being to demonstrate you are a successful performer in a specific organization or context (i.e., perceived fit between the individual and the performance context). Therefore, the performance model proposed in this study—a three-factor model described in detail later—recognizes this fact and uses the Army terminology to label the three latent constructs.

Research Goals

This research project has six basic research goals: (a) to provide a review of the relevant research literatures—training evaluation and job performance modeling; (b) to provide a theoretically sound argument for integrating models of training performance and the direct determinants of job performance, allowing for training criteria to serve as both; (c) to explore the latent structure of training performance and issues related to the modeling performance; (d) to expand on previous Special Forces Qualifications Course (SFQC) research by moving beyond pass-fail performance measures (e.g., Zazanis, Zaccaro, & Kilcullen, 2001) and confirming a multidimensional criterion model; (e) to suggest future

research directions for the field and for SF training; and (f) to suggest possible answers to some of the questions related to the nature of performance posed in the previous section.

General Research Questions

There are four research questions addressed by this study.[1] What is the dimensionality of training performance? Does the latent structure of training performance change over time (i.e., across the phases of training)? Does one general factor or several specific factors describe BE? (i.e., Does one general factor or several specific factors describe the latent structure of the affective or citizenship-like training performance?) What is the impact of overfitting on cross-validation? (i.e., What is the impact of overfitting a CFA model to fit the data of one sample on the cross-validation of the modified model on another sample?) The first two questions focus on modeling the latent structure of training performance, employing CFA and SEM to test the viability of the three-factor model. For the second question, the latent structure of performance for each phase must be confirmed prior to modeling the relationship across the phases. Although important, the last two questions focus on issues related to modeling performance in general. Chapter 2 provides a rationale and detailed description for each question, chapter 3 presents the research models to be tested and the methodology to test them, and chapter 4 presents the results of testing these models. Chapter 5 discusses the results and their implications.

Chapter One Summary

Basically, this research explores the latent structure of training performance and related issues, utilizing the integration of the training evaluation and job performance modeling literatures as an organizing structure. Chapter 1 has introduced the research topic and its context. The increasingly dynamic nature of work requires a thorough understanding

of performance—whether it is training or job performance—and of the issues related to performance measurement and modeling. This knowledge is critical to the development and implementation of effective selection, training, and performance management processes to achieve organizational strategy and goals. The training evaluation and job performance modeling literatures provide the underpinning for this study. Both lines of research have evolved to hold a multidimensional view of performance. However, job performance research has been more adapt to apply CFA and SEM techniques to investigate the structure of performance. This may be an artifact of the difference in how training evaluation research and job performance research have been approached—training performance is typically addressed in the context of training evaluation and training effectiveness research, and job performance is addressed as an independent research topic. Additionally, this chapter has posed some general questions about the nature of performance. Although answers to most of these questions are beyond the scope of this study, the results of this study will provide a foundation for future research to address these questions more definitively. Chapter 1 concluded by introducing the basic research goals and questions addressed by this dissertation. The research literature reviewed is presented in chapter 2. Additionally, the next section introduces the integration framework and the research questions.

CHAPTER TWO: RESEARCH LITERATURE REVIEWED

This chapter reviews the relevant psychological research and issues related to training evaluation and job performance modeling. First, a basic review of performance and criterion measurement issues is presented. Then, the relevant job performance and training evaluation research is examined. Next, an argument for integrating the two research areas is presented, and the four research questions are introduced. Confirmatory factor analysis (CFA) and structural equation modeling (SEM) issues related to the impact of overfitting a model on one sample to the cross-validation of the model on another sample are discussed in relation to the fourth research question. The chapter is divided into seven sections: (a) what is performance? (b) selecting "good" criteria; (c) job performance research; (d) training evaluation research; (e) integrating training evaluation and job performance research; (f) integration inspired research questions; and (g) chapter 2 summary. The research questions are presented in the sixth section of chapter 2.

What Is Performance?

This section explores the basic definition of performance and reviews a set of criteria for determining whether a construct is performance. The section concludes with a brief discussion of whether training criteria meet the definition of performance. The purpose of this section is to define the basic construct and to create a shared understanding and common point of reference that will facilitate the integration of the job performance and training evaluation literatures and the comparison of various research studies.

Before a meaningful discussion can be had, terms must be defined to create a shared understanding. The definition of performance has received much less attention than other variables in psychology (Campbell, 1990; Schmitt & Chan, 1998), but this has changed over

the last decade (Borman et al., 1997; Avery & Murphy, 1998). To a layperson, performance is what you do on the job. For an I/O psychologist, this is *necessary but not sufficient* for the definition. Unfortunately, many researchers have failed to define performance more specifically than a layperson—instead accepting the client's definition of performance (i.e., performance is what our trait-based performance appraisal form measures)—and ignoring the potential construct and measurement problems (Gatewood & Field, 1998). The definition of performance is important because it not only guides the measurement of performance but also has tremendous impact on human resources (HR) processes, like selection and training (Campbell, 1999). As Schmitt and Chan (1998) indicate, job performance is "a variable of considerable societal concern and central to selection research" (p.70). A generally accepted definition of performance is needed.

Many "specific" definitions of performance have been used over the years—including "what is measured by the performance appraisal"—but only one framework will be discussed here. Over the past decade, Campbell and his associates have educated the world about the nature of job performance (Campbell, 1990; Campbell et al., 1993; McCloy et al., 1994; Campbell, Gasser & Oswald, 1996; Campbell, 1999) and given us a general definition of performance and a specific theory. Although there are several frameworks of performance content, most agree with Campbell's basic definition of performance (Borman et al., 1997; Motowidlo et al., 1997; Schmitt & Chan, 1998), so these points of agreement will be utilized here. The specifics of the Campbell model (i.e., the elements and their relationships) will be described later in this chapter.

According to Campbell (1999), "performance is defined as behavior or action that is relevant for the organization's goals and that can be scaled (measured) in terms of the level

of proficiency (or contribution to goals) that is represented by a particular action or set of actions" (p. 402). Campbell's model makes specific assertions about performance that will be described later in this chapter.

First, the performance of interest to I/O psychologists takes place in the context of a work organization—specifically, in the context of a person's job or role in that organization (Campbell, 1990). Second, the performance must be relevant to the goals of the job, work unit or organization (Campbell et al., 1993). Third, performance is the behavior or action, not the outcome or consequence. Fourth, the performance must be under the control of the individual—free of contamination from other factors like resource availability or technology. Fifth, performance is not effectiveness, productivity, or utility—judgments about the results of individual or group work behavior are not performance. Sixth, performance should be representative of what is actually done on the job. Finally, performance is multidimensional. Other theories mirror these ideas.

For example, the theory of individual differences in task and contextual performance (Motowidlo et al., 1997) posits four basic assumptions, which are similar to those of Campbell: (a) performance is behavioral; (b) performance is episodic—performance is made up of many discrete events; (c) performance episodes are evaluated; and (d) performance is multidimensional. Then, performance can be defined "as the aggregated value to the organization of the discrete behavioral episodes that an individual performs during a particular period" (p. 59, Motowidlo & Schmit, 1999).

Once you have agreed on a general definition of performance, you can develop a specific definition of performance for any given job and/or work situation—usually after doing a job analysis guided by a performance model—and this specific performance

definition will guide the development of criterion measures. In this discussion, performance has been defined and illustrated in terms of job performance. Do training criteria represent performance? Do these basic definitions and assumptions about job performance hold for training performance? The basic and safe answer is that it depends on the training context and the criteria. For example, take a basic tenet of Campbell's definition—performance is a behavior that is relevant to the organization's mission and that can be measured at the individual level. Obviously, some training criteria meet this standard and others do not. Likewise, some purported measures of job performance do not meet the standards of the definition or the assumptions about performance. If a training criterion measure meets the basic requirements outlined above, then it is performance. An argument can be made that criterion measurements meeting the definitional requirements in any context are performance and that the equivalence of the latent structure of performance across situations should be the focus. Should the macro-level view of performance be conceptualized as the same general constructs across training and job performance contexts? This is one of the most important theoretical questions for the future of performance research. Regardless of whether it is a job or training performance measure, a criterion variable must have certain characteristics.

Selecting "Good" Criteria

This section provides a discussion of what constitutes "good" criteria by reviewing the characteristics established by various authors in the literature. These characteristics include relevance, measurability, reliability, lack of contamination, and cost. The application of these standards for developing good criteria to archival research is briefly discussed. The purpose of this section is to establish a shared understanding of what is meant by good

performance measures and to introduce a set of criteria by which the training performance measures in this study can be evaluated.

When measuring performance—whether training or job—we are concerned with the development and measurement of criteria that represent the "true" level of proficiency or success in the relevant performance content domains and contexts. As Grant (1996) points out, "effective Human Resource decisions and practices depend on 'good' criteria" (p. 2). Although psychologists have known about the importance of having good performance criteria for decades, the quality of the criterion has been ignored in some cases when the researcher accepts the existing organizational criteria and, therefore, the client's definition of performance (Schmitt & Chan, 1998). The historic lack of concentration on criterion development in psychology is referred to as the "criterion problem" (Austin & Villanova, 1992). Although criteria was not developed for this study, existing criterion measures had to be selected from a large archival data set, and characteristics of good criteria aided in the selection.

Researchers have been discussing criterion issues for years. For example, Nagle (1953) discussed three issues in criterion development—relevancy, reliability, and combining criteria—and provided a four-step procedure for developing criteria. The first and most important step in criterion development is defining performance. The criterion is defined—or accepted—by the researcher. "The criterion measure defines what is meant by job performance…high scores on this measure, therefore, define what is meant by 'successful' job performance" (p. 659, Gatewood & Field, 1998). Basically, the criterion is the operational definition of performance that permits for individual differences in performance levels to be measured (Astin, 1964)—it is the precise method and measurement definition of

performance in a specific situation. If the criterion is poorly chosen and does not represent the performance domain, the selection system or training that is validated against the criterion may be inappropriate or may be unnecessarily eliminated or changed. The criterion measure is of great importance because a poor criterion measure can cause estimates of validity to be inaccurately low (Schmitt & Chan, 1998)—there are several issues that must be considered in developing or selecting criteria. Again, although these issues and characteristics of good criteria are discussed in terms of job performance, they can be applied to training criteria as well.

Gatewood and Field (1998) discuss eight characteristics of appropriate selection criteria—selection research and validation studies are basically job performance research (Schmitt & Chan, 1998)—individualization, controllability, relevance, measurability, reliability, variance, practicality, and lack of contamination. These can apply to training criteria as well.

Individualization refers to the fact that the measure must represent the performance of one and only one participant—the unit of participation is most likely an individual but could be a team. Controllability refers to the fact that the measure must "allow for differences in KSAs to be reflected in performance" (p.678, Gatewood & Field, 1998). This is similar to Campbell's (1999) specification that performance measures must allow for differences in knowledge, skill, and motivation to impact measurement. Measurability refers to the fact that the performance can be scaled for an individual. Relevance relates to the overlap between the criterion measure and what is actually done on the job—this is a measure of validity (Landy & Farr, 1983) and should correspond with the job analysis information (Gatewood & Field, 1998). In training evaluation, criterion relevance refers to the overlap

between the instructional or learning objectives of the training—either identified through training needs assessment or job analysis (can be the same thing)—and the definition of criterion measure (Goldstein, 1993). Often, training objectives are related directly to what is actually done on the job. Criterion deficiency refers to the extent to which performance components identified by the job analysis or in the training objectives are not present in the measurement instrument. The criterion must be relevant to job performance or to training objectives, and it must adequately sample the major components of the content domain of either. Basically, if the KSA is determined by a needs assessment or job analysis and is present in the criterion, then you have criterion relevance (Goldstein, 1993). If the KSA is determined by a needs assessment or job analysis and is not present in the criterion, then you have criterion deficiency. However, using multiple criteria to operationalize performance constructs can help address deficiencies. Reliability refers to the consistency and stability of measurement. Reliability can be defined as "the extent to which a set of measurements is free from variance due to random error or the extent to which the variance in a set of measures is due to systematic sources" (p. 9, Landy & Farr, 1983). This leads to two questions—is job performance reliable and are the defined measures of job performance reliable? Researchers usually take the answer to the first question on faith. As Schmitt and Chan (1998) point out, criterion measures are often "distorted" by measurement error or unreliability. Statistical techniques, such as confirmatory factor analysis (CFA), allow you to model and control reliability at the latent and measurement levels. The same applies for training performance. The variance issue refers to measuring an aspect of performance on which people actually vary—if the threshold of a performance component is so low that everyone does it, then there is no need to measure it because the "amount of KSAs of

workers is apparently irrelevant" (p. 680, Gatewood & Field, 1998). In training, if everyone passes a job knowledge test with a score of 100%, then there is no variance and the job knowledge test is not useful as a measure of training performance to discriminate between trainees in terms of proficiency level or to predict future performance. However, in certain situations, such measures might be reasonable for basic employability certification or legal defensibility. Practicality refers to the cost effectiveness of collecting the metric. Contamination refers to the measurement of something other than the "true" score. Contamination pertains to other sources of variation present in the criteria that result in the "measure not appropriately representing" the construct (Goldstein, 1993). Criterion contamination can lead to incorrect conclusions about selection systems and training program success. Goldstein identified several sources of contamination, including opportunity bias—which is an increasingly important issue as work becomes more dependent on technology and people in organizations have differential access to technological resources (Schmitt & Chan, 1998). The criterion must be under the control of the individual—the differences in performance must be attributable to differences in individuals not other factors. As Campbell (1999) says, "measurement contamination puts the psychologist out of business, or at least at considerable risk" (p. 403).

Landy and Farr (1983) add accuracy—which is often confused with reliability and validity—to the list. "Accurate measurement implies the concepts of reliable and valid measurement, but the reverse is not necessarily true" (p. 22). Accuracy is concerned with the "absolute level of performance" and the extent to which the individual's score reflects the individual's "true" performance level at the time of measurement. Accuracy is a difficult issue and one that has no easy answer.

According to Gottfredson (1991), any performance measurement system should be evaluated using the following factors: (a) validity, (b) reliability, (c) susceptibility to compromise, (d) financial cost, and (e) acceptability to interested parties. Some other criteria for performance measurement that have been identified include: exhaustiveness, uniqueness, comprehensibility, flexibility, maintainability, and acceptability (Tesoro & Tootson, 2000). Exhaustiveness and uniqueness are related to definition of performance—the performance measures should cover the content domain with as little overlap between measures as possible. The other characteristics relate to the practical use of the findings by the client or other stakeholders. Acceptability to the client is of paramount importance because if the client does not agree that the criterion represents useful performance in their organization, the battle is lost before it is begun.

Many methods have been used to measure performance, including ratings, rankings, nominations, work samples, simulations, job knowledge tests, outcome measures (e.g., sales numbers), withdrawal behavior (e.g., absenteeism), training proficiency, and counterproductive behaviors (e.g., employee theft) (Gatewood & Field, 1998; Schmitt & Chan, 1998). The important issue to remember is that the performance measure is representative of performance only to the extent that it meets the requirements of good criteria outlined in this section. These characteristics can be applied to training criteria as easily as they are to job performance criteria.

When working with archival data, the researcher does not have the opportunity to address these issues at the time of criterion development or measurement, but research criteria can be selected from the existing criterion measures based on the extent to which they exhibit the characteristics of good criteria. In field research, especially with archival data, it

is often necessary to accept and utilize measures with less than perfect psychometric characteristics. Unfortunately, this is a major limitation of conducting research in the naturalistic organizational setting.

Job Performance Research

Since most of what is known about the nature of performance comes from modeling job performance, this section provides a review of job performance models, relevant empirical evaluation of the models, and other performance modeling issues. The two predominate models—the Campbell model and the task and contextual performance model—and several other models developed from empirical work, including Wilson and Grant's (1997) three-factor model, are discussed. This is followed by a discussion of the current empirical evidence for each model. Because of multiple models, the primary modeling issue is determining the nature and structure of performance, and this central question leads to an endless number of other questions—for example, what is the number of performance constructs? Are these constructs universal across performance context and content? How does choice of measures and model structure impact construct interpretation? These are some of the many questions raised by reviewing the job performance literature. Although the last question is briefly discussed, they all cannot be addressed here. This section concludes with a summary of the current state of job performance research—what is known and not known. The purpose of this section is to review what is known about job performance research to prepare for the integration of job performance and training evaluation research and to guide the application of methods and approaches from job performance research to training performance.

Overview of Job Performance Models

Current thinking in psychology views performance as behavior, not the outcomes of that behavior (Schmitt & Chan, 1998). However, both are important. As Surface (2000) points out, few psychologists are going to say outcomes should not be measured—they are of practical importance to the organization—but performance outcomes offer a limited view of performance and have limited usefulness from the perspective of understanding and developing employee performance. For performance research, behavioral measures are preferred if available (Campbell, 1999).

Current theory (e.g., Motowidlo et al., 1997; Schmitt & Chan, 1998; Campbell, 1999; Viswesvaran & Ones, 2000) and empirical research (e.g., Motowidlo & Van Scotter, 1994; Wilson & Grant, 1997) support the notion that performance is a multidimensional phenomenon, not a unitary construct as the "classic" model assumed (Campbell et al., 1993). Since the Campbell model was initially introduced, many attempts (e.g., Motowidlo et al., 1997) have been made to define the latent structure of job performance and its determinants. Campbell (1999) believes most of these theories are compatible and hierarchical—for example, the two-factor task and contextual performance model (Borman & Motowidlo, 1993) is a higher-order model than Campbell's eight factors. Borman and colleagues (1997) indicate that two forms of "job performance models are beginning to foster more scientific understanding of criteria" (p. 302)—one explores the latent structure of performance in all jobs (e.g., Campbell et al., 1993), and the other investigates the relationships between the elements of performance (e.g., Hunter, 1983; Borman, White & Dorsey, 1995). Viswesvaran and Ones (2000) suggest that job performance models and research can be categorized along two dimensions—occupational focus (i.e., limited to a specific occupation or family of jobs

versus applicable across all jobs) and developmental context (i.e., created as a stand alone dimension versus set of dimensions). Borman and associates (1997) encourage continuing research efforts aimed at developing a comprehensive taxonomy of performance.

Predominant Job Performance Models

According to Borman and his colleagues (1997) in their review of selection research, the predominant models of job performance are the Campbell model (e.g., Campbell, 1999) and the individual differences model of task and contextual performance (e.g., Motowidlo et al., 1997). Each model specifies a number of performance content factors, performance determinants, and relationships between the model components. The basics of both models, including some empirical research, are discussed before surveying other models of performance.

The Campbell Model

For those unfamiliar with the Campbell model (Campbell, 1990; Campbell, McHenry, & Wise, 1990; Campbell et al., 1993; McCloy et al., 1994; Campbell et al., 1996; Campbell, 1999; Campbell & Knapp, 2001), a brief overview will be provided. The model was developed and tested as part of the U.S. Army's "Project A", a large-scale selection and classification research project sponsored by the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI). Campbell's model presents a behavioral and multidimensional model of job performance. Campbell's model contains performance components, performance determinants, and antecedents (i.e., predictors) of performance determinants—these antecedents become indirect determinants and the determinants become direct determinants in later versions of the model (Campbell et al., 1996; Campbell, 1999).

The Campbell model's performance components (i.e., factors or dimensions) are the latent constructs of performance that capture "the distinguishable categories of things people are expected to do in a job" (p. 42, Campbell et al., 1993). Campbell posits eight performance components, with three factors—core technical proficiency, demonstrating effort, and maintaining personal discipline—being relevant for every job. Core technical proficiency includes both job- and non-job-specific technical proficiency. The eight performance dimensions are: (a) job-specific task proficiency; (b) non-job-specific task proficiency; (c) written and oral communication task proficiency; (d) demonstrating effort; (e) maintaining personal discipline; (f) facilitating peer and team performance; (g) supervision and leadership; and (h) management and administration. The eight factors are meant to describe the latent structure of all jobs in the Dictionary of Occupational Titles (DOT). However, all eight may not apply to every job, and the factors may not be in the same form across all jobs. A definition and discussion of each performance component can be found in Campbell et al. (1993) or Campbell (1999).

Performance factors are "viewed as a function of three, and only three…determinants" (p. 45, Campbell et al., 1993). These direct determinants of performance—declarative knowledge (DK), procedural knowledge and skill (PKS), and motivation (M)—are in turn a function of the indirect determinants of performance—constructs such as ability, personality, interests, and interventions (Campbell, 1999). According to Campbell, the precise nature of these functions will probably never be known, but the performance component function can be expressed generically as PC = f (DK, PKS, M) with each determinant being a function of indirect determinants. DK refers to "the ability to state the facts, rules, principles, or procedures that are a prerequisite of successful task

performance"; PKS refers to "the capability attained when DK has been successfully combined with knowing how and being able to perform a task"; and M refers to "the combined effects of three choice behaviors: (a) the choice to expend effort, (b) the choice of what level of effort to expend, and (c) the choice to persist in the expending of the chosen level of effort" (p. 494, McCloy et al., 1994).

These direct determinants are functions of indirect determinants. DK is a function of ability, personality, interests, education, training, experiences, and aptitude-treatment interactions (Cronbach's ATI framework, as cited in Campbell, 1999). PKS adds practice to the list of indirect determinants. For M, Campbell suggests to review the "independent variables stipulated by research and theory in motivation" (p. 409, exhibit 12.1, Campbell, 1999). Campbell's model states that job performance—or any behavior—cannot occur without motivation, making it a necessary condition. Some minimal levels of PKS and DK are necessary but not sufficient for performance. Motowidlo and colleagues (1997) sum it up well, "individual differences in personality, ability, and interest are presumed to combine and interact with education, training, and experience to shape DK, and PKS" (p. 73).

Before moving on to discuss empirical support for the Campbell model—especially its direct determinants—it should be noted that the Campbell model makes many assertions about the nature of performance: (a) interventions, like training, influence performance only through impacting the three direct determinants of performance (Campbell, 1999); (b) the three direct determinants—DK, PKS, and M—are not performance components themselves because performance components are a joint function of knowledge, skill, and motivation (Campbell, 1999); (c) the relationship between performance components and indirect determinants is fully mediated by direct determinants (Campbell, 1999); (d) a performance

measure can be “determinant deficient” when individual differences in one of the direct determinants is held constant and/or not allowed to influence the individual’s performance (Campbell, 1999); (e) structural constraints that impact performance and do not impact the direct determinants are considered a characteristic of the job and criterion contamination (McCloy et al., 1994); (f) any variation not attributed to the direct determinants is contamination or measurement error (McCloy et al., 1994); (g) measures of the determinants should be isomorphic in performance content—otherwise, “a given criterion (e.g., job knowledge test) could contain variance that was unique to that measure and would erroneously be considered irrelevant” (p. 465, McCloy et al., 1994); and (h) direct determinants are not performance because “the skill and knowledge must used to accomplish the goal-relevant tasks” (p. 495) that are performance, and people are not typically rewarded for their knowledge and skill (McCloy et al., 1994). Although these assertions about the nature of performance have a basis in theory or previous research, some have not been tested, and others are points of contention. Much research is needed into the nuances of the Campbell model.

For the most part, all of the empirical work testing and supporting the Campbell model has been related to Project A. However, although she failed to support the Campbell model, Grant (1996) tested the performance content factors from the Campbell model using CFA on data from incumbent state troopers and found that the model did not achieve acceptable fit. Most of Campbell’s recent articles (e.g., Campbell, 1999) have been based on past Project A research and tend to be more conceptual and descriptive in nature. However, one exception is the McCloy, Campbell, and Cudeck (1994) article that discussed and confirmed Campbell’s direct determinants of performance—DK, PKS, and M—using data

from Project A. This is the only published empirical test of the Campbell et al. (1993) direct determinants found by this researcher. McCloy and associates (1994) had two goals—to confirm the model of determinants and explore the relationship between several predictors and the determinants. The authors, using a SEM technique, found confirmation for the three determinant model using Project A performance data. Additionally, regression analysis was used to estimate the relationship between several predictor constructs—cognitive, computer, temperament, and interest—and the determinants. A detailed description of the study by McCloy and his associates follows.

McCloy and his colleagues (1994) utilized existing Project A data to operationalize the performance determinants. The manifest measures used consisted of job knowledge tests, work sample tests (i.e., hands-on tests of maximal performance), supervisor ratings, and personnel file data (i.e., awards and letters and Articles 15 and flag actions). The authors posit that job knowledge tests are a function of DK, work sample tests are a function of DK and PKS, and ratings and personnel data are functions of DK, PKS, and M. They argue that tests and work samples, as measures of maximum performance, prevent individual differences in motivation from operating and that personnel file data and ratings are measures of typical job performance. However, this point can be argued. The model further stipulates that "performance measures are a function of different determinants and these are independent of content" (p. 495, McCloy et al., 1994).

McCloy and his colleagues (1994) required all the measures to be isomorphic in content—with the exception of the personnel file data and that is pointed out as a weakness in their discussion section—in other words, all measures should cover the same performance content domain. "Failure to achieve isomorphism in content confounds content with the

properties of the observed measures, there by disrupting the structure of the variance decomposition required for a confirmation analysis of the model" (p. 495). According to the authors, "the model provides an empirical description of the degree to which various criterion measures are contaminated or deficient, given the isomorphic content" (p.502). In their discussion, McCloy and associates (1994) discuss the need for measures to be isomorphic in content and how the lack of isomorphism of the motivation measures in their study "most probably acts to reduce the amount of shared variance for the indicators of M" (p. 502).

From a CFA or SEM perspective, isomorphic content is clearly not a requirement—the model can be structured to account for homogenous or heterogeneous content factors if the model is identified enough to estimate all the necessary parameters. Additionally, by modeling all the measures as a function of DK, the common factor variance for DK would most likely include the variation from the isomorphic content as well as the DK. Remember, any measure can be expressed as a combination of the common factor variances (i.e., shared variance with other measures of the construct) of the factors upon which the measure loads and its uniqueness term (Long, 1983). Therefore, if the content is shared by all measures of a construct and the latent factor is meant to be modeling DK only—not any content—then the variation attributed to DK is contaminated with the variation from the isomorphic content—in other words, is it declarative knowledge or declarative knowledge and shared content? The DK factor would contain the content variation shared across all the measures in addition to the shared DK variation, while the PKS and M would not because all the measures do not load on each of those factors. This suggests isomorphism of content is an issue best handled through model structure—in organizational settings, having content-isomorphic criteria measured with multiple methods is the exception, not the norm.

The study used an unusual methodology to test the model of determinants. Determinants were defined as factors but no factor loadings were estimated, nor were the correlations between latent factors. They used a "design matrix" of 1s and 0s that specify whether a variable does or does not load on a factor, respectively. "The relevant variance in performance criteria is thus described by the estimated factor variances; and the irrelevant variance in each estimate is relegated to that measure's uniqueness term" (p. 499). The authors reported the factor variance estimates for eight military occupational specialties in the study (e.g., infantryman: DK = 0.481, PKS = 0.728, M = 1.804). Their model defines the latent determinants as independent, so no factor covariance was modeled.

Why did McCloy and associates (1994) not use a "standard" confirmatory factor analysis (CFA) that estimated loadings and covariances for the model? If you do the math for a CFA model based on their structure (i.e., the counting rule for identification; Hatcher, 1994), the model is identified enough to estimate all of the parameters. They estimated three factor variances, eight errors, and seven "nuisance" parameters, but no loadings or correlations between the factors. So, why not do a more standard CFA? Is the model not identified for another reason? If the model was not identified, does that invalidate the variance estimates? This methodology, as well as splitting manifest measures to have enough indicators, makes this a less compelling confirmation. Plus, no alternative models were tested—a technique recommended to strengthen confidence in the model proposed (Loehlin, 1998; MacCallum & Austin, 2000)—but the cross-validation of the model across samples from different military occupational specialties is impressive. Although many questions remain, it should be restated that McCloy et al. (1994) is the only empirical test of the direct determinants in the Campbell model and an important piece of research. More research is

needed into the nuances of the Campbell model and its hypothesized relationships between constructs.

The Task and Contextual Performance Model

The other predominant model of job performance is the two-factor model of task and contextual performance (Borman & Motowidlo, 1993; Motowidlo et al., 1997). Task performance (Borman & Motowidlo, 1993) refers to activities related to production—transforming raw materials into the products and services of the organization—and to servicing and maintaining the *technical core*. "Thus, task performance bears a direct relation to the organization's technical core either by executing its technical processes or by maintaining and servicing its technical requirements" (p. 75, Motowidlo et al., 1997). Contextual performance behaviors "do not support the technical core itself as much as they support the organizational, social and psychology environment in which the technical core must function" (p. 73, Borman & Motowidlo, 1993). Borman and Motowidlo (1993) suggested that contextual performance is divided into five categories: (a) volunteering to carry out task activities that are not formally part of the job; (b) persisting with extra enthusiasm when necessary to complete own task activities successfully; (c) helping and cooperating with others; (d) following organizational rules and procedures even when it is personally inconvenient; and (e) endorsing, supporting, and defending organizational objectives. Contextual performance behaviors are very similar to organizational citizenship behaviors (Organ, 1988, 1997) because they overlap in definition and share the defining quality of being non-task related. However, Organ (1997) objects to the term *contextual performance* because he believes the name is "cold, gray, and bloodless" (p.91).

Motowidlo and colleagues (1997) theorize direct determinants of performance—knowledge, skill and work habits—that are specific to each performance domain (i.e., task or contextual performance). According to the theory, there are six direct determinants of performance—task knowledge, task skill, task habits, contextual knowledge, contextual skill, and contextual habits. Work habits are responses to the situation that either facilitate or interfere with performance. Habits include characteristic responses that are motivational in nature (i.e., the choice to exert effort in certain type of situation) and responses that are not motivational in nature. "Our theory predicts that individual differences in personality and cognitive ability variables, in combination with learning experiences, lead to variability in knowledge, skill, and work habits that mediate effects of personality and cognitive ability on job performance" (p. 73, Motowidlo et al., 1997). Work habits were developed from one of the constructs—characteristic adaptations—in McCrae and Costa's (1996) framework based on personality theory and research. According to McCrae and Costa (1996), basic tendencies (indirect determinants) have their impact on objective biography (performance) only through characteristic adaptations (direct determinants). At the core of characteristic adaptations, "individuals react to their environments by evolving patterns of thoughts, feelings, and behaviors that are consistent with their personality traits and earlier adaptations" (p. 74). This is similar to the process of how individual differences and situational factors influence performance through shaping DK, PKS, and M in the Campbell model. Motowidlo and his colleagues (1997) posit that the determinants directly mediate the relationship between cognitive ability and personality on one side of the model and performance on the other side. According to their theory, cognitive ability impacts task performance through the three task determinants and personality impacts contextual performance through the three contextual

determinants. Two exceptions are posited—the cognitive ability-contextual knowledge link and the personality-task habits link.

Although no research has operationalized the complete model (i.e., the entire model including performance determinants), much research has supported the existence of task and contextual performance constructs (e.g., Motowidlo & Van Scotter, 1994; Loviscky, Rosenberg, Mathieu & Mohammed, 1998; Conway, 1999). Motowidlo and Van Scotter (1994) and Van Scotter and Motowidlo (1996) found support for this two-factor conceptualization of performance. Using a sample of 715 Air Force mechanics, Motowidlo and Van Scotter (1994) found that task performance explained from 17% to 44% of the unique variance in overall performance beyond contextual performance and that contextual performance explained from 12% to 34% of the unique variance in overall performance beyond task performance. In addition, the study found that experience explains more variance in task performance and personality explains more variance in contextual performance. The authors conclude that performance—as judged and measured by supervisor ratings—is not unidimensional and that the dimensions of task and contextual performance are worth distinguishing. Again using Air Force mechanics, Van Scotter and Motowidlo (1996) provided further empirical support for the task and contextual performance model by demonstrating that ratings of contextual performance and ratings of task performance explain significant and unique variance in overall performance ratings made by a different rater. Additionally, the authors looked at two facets of contextual performance—interpersonal facilitation and job dedication—and concluded from the data that job dedication may not be distinguishable from task performance. Werner (1994) proposed a two-factor performance model using the terminology in-role and extra-role that

are basically task and contextual performance factors. He used a 3 x 2 within subjects design to demonstrate that the ratings of supervisors evaluating the job performance of secretaries were highly influenced by both in-role and extra-role behaviors. He also examined the rater accuracy and bias and found that all ratings of ratees with high levels of extra-role performance showed significantly more halo. Caligiuri and Day (2000) conducted a study utilizing a three-factor model for global assignees—technical, contextual and expatriate-specific performance dimensions—derived from the task-contextual framework to investigate the relationship between self-monitoring and performance. They suggest that certain job assignments, like global assignments, require a context-specific performance dimension. Using exploratory factor analysis (EFA), the authors found three factors consistent with their a priori performance dimensions. The results of their analyses of variance suggest that the self-monitoring construct has differential impacts across performance constructs and rater nationalities.

The empirical studies described above are only a sample of those conducted since the Borman and Motowidlo (1993) article. The task and contextual performance model continues to be more investigated than the Campbell model, probably because two factors are easier to operationalize than eight. However, more research is needed. The six performance determinants proposed by Motowidlo et al. (1997) have yet to be operationalized completely and empirically confirmed. The exact structure and nature of performance still remains unclear.

Other Performance Models

Empirical research has developed and confirmed several other models of performance. For example, using a sample of North Carolina Highway Patrol (NCHP)

officers (i.e., incumbents) and their importance ratings of job tasks, Grant (1996) tested several models of performance, including a two-factor model of performance (in-role versus extra-role), a version of the Campbell et al. (1993) model, and a hybrid model combining the Campbell model with the two-factor model as second-order factors. After all the a priori CFA models failed to confirm, Grant found that a post-hoc three-factor model was the best fit for the data and cross-validated the model successfully on another sample. Grant labeled the three factors: (a) know in-role, (b) do in-role, and (c) extra-role. Grant found robust correlations between all of her latent factors (e.g., r = .74 for the know in-role and extra-role relationship). Her research raises questions about the Campbell et al. (1993) model and about two-factor models. However, her model is very similar to Campbell's three direct determinants of performance (i.e., DK, PKS, & M), raising the question of whether she modeled performance, the direct determinants of performance, the cognitive structure of how incumbents view the job, or all of the above.

Wilson and Grant (1997) proposed a priori a three-factor model of performance—knowing the job, doing the job, and citizenship—as part of a selection validation study with the NCHP. The authors utilized CFA to test their model. The factors were operationalized with test scores from training for knowing the job, activity data (e.g., number of tickets) for doing the job, and ratings for citizenship. The study confirmed the a priori model. Interestingly, the authors found a small-to-moderate correlation between doing the job and knowing the job (r = .36), but citizenship had almost no relationship with knowing the job and doing the job (r = -.02 and r = -.06, respectively). The correlation found between the knowing the job and doing the job constructs is similar to the correlations found between knowledge and skill in other research (e.g., Hunter 1983). Based on Campbell's assertions

about performance, the Wilson and Grant model can be questioned for its use of training data to operationalize a job performance content factor.

Several other job performance theories are presented briefly as points of comparison. Hunt (1996) developed a nine-factor performance model for service jobs using a data set of over 18,000 performance ratings. The nine factors included schedule flexibility, attendance, and thoroughness as well as several counterproductive behaviors like theft. Viswesvaran (1993 as cited in Viswesvaran and Ones, 2000) proposed that there is an overall performance factor and a number of sub-factors, including task-specific factors. He reviewed 486 performance measures and derived 10 rational categories of performance. Then, he acquired data from over 300 studies that reported correlations across the 10 dimensions. The results of his analyses suggest a general performance factor—one factor accounted for approximately 50% of the variance shared across the 10 dimensions. Although this model has received little attention, research along these lines supports the notion that performance may be best understood in terms of content and level of abstraction (specificity versus generality). Maybe, performance can be modeled as both specific and general factors.

Other research has identified determinants of performance. Hunter (1983) conducted a meta-analysis and found that ability impacted performance (supervisory ratings) only through job knowledge and skill (work sample performance)—the ability-performance relationship was fully mediated by the determinants. In other words, there was no direct path from ability to performance. As Motowidlo and associates (1997) point out, since the measures are reasonably operationalizations of the constructs, Hunter provides support for the notion of full mediation by the direct determinants, which supports the conceptualization of direct determinants in both the Campbell et al. (1993) model and the Motowidlo et al.

(1997) model. Although a consensus exists on certain points, like the existence of multiple performance factors, the exact structure and nature of job performance is still unclear, and more research is needed to address the issue definitively.

Interpreting the Performance Constructs

One of the goals of this proposal is to integrate job performance research and training evaluation research into one stream of performance research. One obvious integration point is to apply methods used in job performance research, like confirmatory factor analysis (CFA) or structural equation modeling (SEM) techniques, to the study of training performance. These techniques require researchers to specify many aspects of the structural and measurement models in advance. When applying these techniques, researchers should understand the implications of their choices to model fit and model interpretation. The job performance research by McCloy and associates (1994), Grant (1996), and Wilson and Grant (1997) provides an opportunity to review how choices in measures, measurement methods, and model structure impact the interpretation of the latent constructs.

As a point of initial comparison, McCloy and colleagues (1994) are modeling the direct determinants of performance, and Grant (1996) and Wilson and Grant (1997) are modeling job performance constructs. Given the measures, methodologies, and confirmed factor structures of all three studies, an argument could be made for the factors as determinants, performance, or both. However, this is not the point of this section. The impact of the measures chosen on the interpretation of the latent constructs and relationships between the latent constructs can be profound. McCloy and associates (1994) used multiple measurement methods with isomorphic content and posited that each measure was a function of DK, PKS, and M—some measures were only a function of DK, others were a function of

DK and PKS, and others of DK, PKS and M. All measures were assumed to be partially a function of DK. If modeled using a standard CFA measurement model—where the factor loadings and covariances were estimated—the common variance for DK would include the variation from the isomorphic content as well as the DK. Remember, any measure can be expressed as a combination of the common factor variances (i.e., shared variance with other measures of the construct) of the factors upon which the measure loads and its uniqueness term (Long, 1983). Therefore, the DK construct would be defined by the shared content variation as well as the shared DK variation, while the PKS and M would not because all the measures do not load on each of those factors. It is DK and content. If the view is that DK is content independent—which McCloy and associates (1994) indicate—then this would confound the interpretation of DK and the other determinants. Specifying that all measures are a function of DK and having isomorphic content across measures would obscure the relationship between the determinants as well. Since their method did not estimate loadings or latent factor correlations, it is not possible to know for certain. However, since they acknowledged that one of their personal history measures was not isomorphic in content, the interpretation of the factors would be inherently difficult if estimated—most likely, all content variation in the measures not shared by the personal history measures or accounted for by other latent factors would be relegated to the error term (i.e., the term containing the measure's unique variance), possibly leading to highly correlated residuals.

Grant (1996) utilized incumbent ratings for all of her measures, and her measures loaded on only one factor each (i.e., independent clusters model; McDonald & Ho, 2002) and were not isomorphic in content. She found high correlations between her latent job performance constructs—know in-role, do in-role, and extra-role—which resemble DK,

PKS, and M. These high correlations should be of no surprise. Since all of her measures used the same measurement method in the same context, all the latent factors should be correlated because the common variation for each factor includes a component of the same method and context variance.

Wilson and Grant (1997) used different measurement methods for each factor, the task content was not isomorphic, and each measure loaded on only one factor. Wilson and Grant found a correlation between knowing the job and doing the job (r = .36) that is reasonable given the literature concerning the relationship between knowledge and skill (e.g., Hunter, 1983) and virtually no correlation between citizenship and the knowing the job and doing the job factors. Wilson and Grant's methodology more accurately assessed the relationships between the constructs because their correlations are based on having construct-relevant variation in common, not having method, task content, or context variation in common. In other words, the method variance in each of the latent factors was unique and did not overlap (i.e., correlate) with the method variance from the other latent factors. This is a justification for interpreting the meaning of the model's latent structure based on the operationalization of the manifest indicators that are utilized at the measurement level. The manifest measures used to operationalize the measurement model may necessitate changes in the structural model. When applying CFA and SEM techniques to training or any other type of performance, researchers should be aware that the operationalization of the measurement model in combination with the latent configuration could have a great impact on the construct-level interpretation.

Job Performance Research Summary

Research over the past decade has begun to map the job performance criterion space. After being considered a unitary construct for years, job performance has emerged as a multidimensional phenomenon. Although there are a number of job performance theories, Campbell's eight-factor model and Borman and Motowidlo's model of task and contextual performance are considered the predominate models in the field. A comparison of job performance models can be found in Table 1.

Campbell (1999) believes that many of the other models fit hierarchically within or above his framework. Borman and Motowidlo's two-factor model could be conceived as a higher-order model to Campbell's eight-factor model. However, Grant (1996) presents empirical evidence to the contrary. Both models specify performance content factors and direct and indirect determinants of performance. In both models, direct determinants are thought to mediate the relationship between indirect determinants (e.g., human attributes) and job performance content factors fully. The models disagree as to the number and nature of the direct determinants with Campbell's model specifying three—DK, PKS, & M—and the task-contextual model specifying six—task knowledge, task skill, task habits, contextual knowledge, contextual skill, and contextual habits. There has only been one published confirmation of the Campbell model's direct determinants. Neither model has been completely operationalized in one study. Other models—derived from empirical research—offer alternative views of the criterion space. The Wilson and Grant (1997) model—knowing the job, doing the job, and citizenship—may provide a strong challenge. Comparison of various studies demonstrates that differences in the measurement model can impact the interpretation and relationship of the latent factors.

Specifically, the following is known (or assumed) about the nature of job performance: (a) job performance is multidimensional; (b) job performance consists of measurable behavioral episodes under the control of an individual in the work environment that are meaningful to the employing organization; (c) there are several competing models of job performance that are thought to be hierarchically related (Campbell, 1999); (d) these job performance models specify performance content factors, and some specify performance determinants and their relationship to human attributes; (e) performance determinants are thought to mediate fully the relationship between human attributes and job performance; (f) the Campbell eight-factor model and the model of task and contextual performance are the two predominate models of performance; (g) the predominate job performance models have never been completed operationalized and empirically tested in the same CFA model; (h) job performance at one point in time is thought to predict more distal performance factors; (i) SEM techniques have been frequently applied to research the latent structure and nature of job performance; (j) the measures and model structure used in the SEM technique can impact the interpretation of the latent performance constructs and their relationships; and (k) the nature of job performance will continue to be an important research topic.

Although the state of job performance research has improved greatly in the past 12 years, many questions still remain. More research is needed to determinant the nature and structure of job performance. Many of the questions about job performance are still unanswered and may never be answered: (a) How many factors adequately describe the job performance criterion space? (b) Are the various job performance models hierarchically compatible? (c) If they are, are certain models or levels more appropriate in certain situations? (d) What is the structure and nature of the direct determinants of performance?

(e) What is the third performance determinant? Is it motivation or habits or citizenship? (f) Does the latent structure of job performance (or determinants) change over time? (g) Is the latent structure of job performance independent of job content and context? (h) Are determinants performance content specific or general? (i) Do direct determinants fully or partially mediate the relationship between human attributes and job performance? (j) What are the most appropriate measures to operationalize job performance constructs? (k) What is the relationship between job performance models and training performance? Can training performance and job performance have the same latent structure? and (l) How will the changing nature of work impact job performance models? The list of what is not known could continue for pages. These are some of the questions that should be addressed by future research—most of which are beyond the scope of the current research.

Foreshadowing what is to come, there are some issues from the job performance literature that are of particular interest: (a) the structure of performance; (b) the relationship between job performance determinants and factors and training performance factors; (c) the stability of the latent structure of performance over time; (d) the content specificity of performance factors and determinants; (e) the mediation of the relationship between human attributes and job performance; and (f) the nature of the third performance determinant. A review of training evaluation research is next.

Training Evaluation Research

This section reviews the training evaluation research literature. This review focuses on defining the basic terms, presenting the two basic categories of evaluation questions, and reviewing models of training criteria and related empirical research. The section concludes with a summary of what is known and not known in this area.

Training Evaluation Overview

As Salas and Cannon-Bowers (2001) report, training research has been very robust over the past decade, and several pivotal articles have expanded the conceptualization of training evaluation (e.g., Kraiger, Ford & Salas, 1993; Alliger, Tannenbaum, Bennett, Traver & Shotland, 1997). Since our interest is training criteria (i.e., training performance), the training evaluation and training effectiveness literatures are the focus of review in this section. Before reviewing the relevant literature, terms must be defined to create a shared understanding, starting with the definition of training.

According to Goldstein (1993), training can be defined "as the systematic acquisition of skills, rules, concepts or attitudes that result in improved performance in another environment" (p.3). This means training must be linked to performance in future training or on the job. For Broad and Newstrom (1992), "training consists of instructional experiences provided primarily by employers for employees, designed to develop new skills and knowledge that are expected to be applied immediately upon (or within a short time after) arrival on or return to the job" (p. 5). The primary difference between training and development is the idea that training is designed to be immediately useful on the job and development has no immediate application to the job. However, the definition of immediate varies greatly.

A job analysis, training needs assessment, or performance needs analysis should be used to develop training objectives. The training objectives specify the outcomes of training and suggest criteria for measuring successful training performance (Goldstein, 1993). Training objectives are typically stated in terms of knowledge and skill development, or on-the-job behavior (Taylor & O'Driscoll, 1998). Instructional objectives specify the behavioral

objectives of the training and describe what the trainees should be able to achieve or perform at the end of training (Goldstein, 1993). Training criterion measures should be related directly to the training objectives and to future performance.

When discussing training criteria, many different terms seem to be used interchangeably (Kraiger et al., 1993; Ramirez, 2000)—training evaluation, training effectiveness, validation, or assessment. Often, these terms have very different meanings. According to Kraiger and Colleagues (1993), evaluation is conducted to determine "whether training objectives were achieved and whether accomplishment of those objectives results in enhanced performance on the job" (p. 311), and training effectiveness seeks to discover "why training did or did not achieve its intended outcomes" (p. 311). Training effectiveness is a broader concept and encompasses training evaluation and its criteria. For Talbot (1992), evaluation refers to "the assessment of the total value of a training system, training course, or training programme in social as well as financial terms" (p. 26). Whereas, validation refers to a series of measurements to determine whether or not the training achieved the behavioral objectives. Assessment is used by human resources development (HRD) practitioners as a surrogate for "evaluation" to soften the perceptions of stakeholders (e.g., Swanson & Holton, 1999). Goldstein (1993) suggests evaluation involves two processes—establishing criterion measures and using research designs to determine the changes that occurred in the criteria during training and transfer. These criteria should pertain to immediate training success and to success in more distal performance environments like the work environment. Additionally, the characteristics of "good" job performance criteria presented earlier should be applied to developing training performance criteria. Since this research focuses on

training criteria (i.e., performance), the training evaluation literature will be the focus of our review and that term will be used instead of training effectiveness, validation, or assessment.

Sackett and Mullen (1993) suggest that evaluation is about answering two different categories of questions—one refers to how much change has occurred as a result of training, and the other relates to the attainment of a specified level of achievement or performance by the end of training. Training evaluation can focus on one or both of these goals. There are several situations where the measurement of change might be of interest: (a) evaluating the utility of the training in terms of organizational goals; (b) comparing two different training programs; and (c) researching the effectiveness of different training methods.

Situations, where measuring the level of performance is needed, have "two critical features" (Sackett & Mullen, 1993): (a) there is a clearly defined level of desired performance; and (b) there is an interest in documenting the performance of individual trainees. In other words, once trainees successfully complete the course, they are certified to have the desired level of proficiency in the training content. Trainees who fail the course are often given the opportunity to repeat the course (or the part of the course failed) or may be terminated from the course and possibly the job. The major difference between change-oriented and performance-level evaluation is that change-oriented is concerned with the functioning of the program and level-of-performance is concerned with trainee proficiency and program functioning (Sackett & Mullen, 1993). Another difference is that performance-level evaluation requires ongoing measurement of each trainee—constant assessment and grading of individual trainees—whereas change-oriented evaluation can happen as a discrete activity and stop once the research question has been addressed. A well-known change-oriented evaluation method is the pre-test post-test design (Kirkpatrick, 1996).

Overall, Sackett and Mullen (1993) suggest that there are several reasons why organizations undertake evaluation regardless of the type: (a) making decisions about the future use of training; (b) making personnel decisions about trainees; (c) contributing to the scientific understanding of training; and (d) providing data for political or public relations reasons. The authors note that the last two reasons for evaluations are often add-ons. In other words, organizations undertake evaluation to make decisions about training or the individuals in training, and research and public relations are just bonus items.

Training programs for some jobs, like police officers and soldiers, provide examples of situations where the level of trainee performance is of paramount importance. In these situations, training is designed not only to prepare trainees to perform the missions, duties, and tasks of the job but also to function as an additional screening mechanism to ensure a high level of proficiency in work situations where mistakes can have life-and-death consequences. Since police officers and soldiers may face situations where their lives are on the line, their proficiency in terms of critical skills becomes very important to staying alive and completing the mission successfully. Whether it is the police academy or U.S. Army Special Forces training, this type of training provides a minimum acceptable proficiency in the areas identified as critical to job performance, and the training continues on-the-job through mentoring and training exercises. Since candidates must graduate these programs with an acceptable level of proficiency, measurement of relevant training outcomes in terms of level of performance evaluation receives the focus. This research focuses on training criteria used for level-of-performance evaluation—in building a model of training performance, the level of proficiency is important, not necessarily the amount of change that occurs during training.

The Kirkpatrick Model

According to many authors (e.g., Kraiger et al., 1993; Salas & Cannon-Bowers, 2001), the most popular model used for training evaluation has been Kirkpatrick's (1979, 1996) four levels of training criteria: reactions, learning, behavior, and results. Many authors have criticized the Kirkpatrick framework (e.g., Swanson & Holton, 1999). However, many of the criticisms are the result of evaluators misunderstanding his model or lacking the skills to effectively evaluate training criteria at the various levels (e.g., Surface, 2000; Salas & Cannon-Bowers, 2001). Alliger and Janak (1989) suggest many of the misunderstandings and misuses of Kirkpatrick's model result directly from problems with the model. If taken as four categories of training criteria to evaluate—not a comprehensive model specifying relationships between outcomes or an exhaustive list of methods for evaluating outcomes—Kirkpatrick retains its usefulness. As Surface (2000) points out, "true evaluation involves posing mission-related questions and using practical research methodologies to answer them. Unfortunately, most HRD practitioners are not well versed in assessment and research methods" (p. 236).

Many of the same authors who criticize Kirkpatrick's model (e.g., Swanson & Holton, 1999) have modified his model, making the situation more confusing. Swanson and Holton (1999) collapsed the model into three categories—performance, learning, and perceptions—with performance focused on measures of outcomes, not behavior as Campbell (1999) and most other psychologists would insist. This may lead to many inappropriate and ineffective decisions about training programs and trainees—for example, the use of outcome measures makes giving developmental feedback difficult (Surface, 2000). One legitimate criticism of Kirkpatrick's model refers to his limited definition of learning (Kraiger et al.,

1993), which will be discussed momentarily. The Kirkpatrick model has also been criticized for its theorized—or implied—relationships between the levels of criteria that have only been partially supported by research (Alliger & Janak, 1989; Alliger et al., 1997).

Reactions refer to the perceptions of trainees about the training. These perceptions are usually measured at the end of training by administering a series of Likert-type items designed to assess satisfaction with various aspects of the training from the content to the instructor. Recent research has classified reactions into two categories—affective and utility (Alliger et al., 1997). Basically, affective reactions are how much the participant liked the training, and utility reactions refer to how useful the participant found the training in terms of the practical benefits to the trainee on the job. Alliger and colleagues (1997) conducted a meta-analytic study that demonstrated utility-type reactions where more strongly related to transfer of training and on-the-job performance than were affective-type reaction measures.

Kirkpatrick's (1996) second level—learning—is conceptualized as measures of acquisition of declarative knowledge or skill. Declarative knowledge can be measured using content-relevant tests. Work samples (i.e., highly controlled work-related simulations) can be used to measure skill. Kirkpatrick (1996) focuses on the change in knowledge and skill as a result of the training—he even suggests using a pre-test-post-test measurement methodology. For Kirkpatrick, the reason for evaluating training is to determine the effectiveness of a training program. He does not focus on level-of-performance evaluation.

Kirkpatrick's third level—behavior—refers to measuring the impact of training on job behavior (i.e., job performance). In the evaluation literature, behavior has traditionally been measured as transfer of training or actual job performance. Transfer of training is

defined as the effective and continuing use of knowledge and skills learned in training on the job (Broad & Newstrom, 1992).

Kirkpatrick's fourth level—results—refers to more distal outcomes of training that are related to the goals of the organization. Productivity, sales volume, or stock price would be examples of organizational results.

Beyond Kirkpatrick

Many researchers and practitioners have expanded or adapted the Kirkpatrick framework. Tannenbaum, Cannon-Bowers, Salas, and Mathieu (1993) expanded Kirkpatrick's model by adding attitude change and training performance to the list of criteria containing six "levels". Changes in attitudes as a result of training may increase or decrease motivation, self-efficacy, and commitment. Training performance refers to trainees demonstrating the desired behavior during training via role-plays, simulations, and work samples. In Europe, Talbot (1992) reports the main adaptation to Kirkpatrick's model is to add a level assessing departmental change between the behavioral level and the organizational results level. This might be thought of as adding a team performance level to Kirkpatrick, creating a fifth level in between behavior and results. In the HRD community, many practitioners have modified the model. One example, Swanson and Holton (1999), collapses the four Kirkpatrick levels into three—performance, learning, and perceptions. Their performance level includes Kirkpatrick's behavior and results by their own statement, but the authors focus more on the outcomes of behavioral performance, which goes against current thinking in psychology (Surface, 2000). Many of the practitioner adaptations are attempts to repackage Kirkpatrick, and these changes add little to the understanding of training criteria or evaluation. As Salas and Cannon-Bowers (2001) point out, several pivotal

articles have started to move training evaluation and criteria beyond Kirkpatrick. Two of these studies are discussed next—one used a meta-analysis to further our understanding of Kirkpatrick, especially reaction measures, and the other reconceptualizes learning criteria.

Alliger, Tannenbaum, Bennett, Traver and Shotland (1997) conducted a meta-analysis of the relationships among the levels of training criteria based on an augmented model of Kirkpatrick's four levels. This model expanded levels one and two and defined level three in terms of transfer. Reactions were split into affective-type, utility-type, and combined-type reaction measures. The model categorized learning into immediate knowledge, knowledge retention, and behavior/skill demonstration. Alliger and colleagues found modest correlations between the different types of training criteria, with stronger relationships between measures at the same level. Utility and combined reactions correlated more highly with job performance than did affective reactions. Surprisingly, utility reactions were found to be more related to job performance than learning measure were found to be. In general, the authors conclude, "reaction measures cannot be used as surrogates of other indicators" (p. 353). Another interesting finding is that learning measures tend to have lower reliability coefficients than reaction and performance measures. A moderator analysis demonstrated "that if training criteria do not overlap in content, convergence between or among them should not be expected" (p. 353). Convergence would be expected only when you had multiple measures of the same or highly related content. The authors suggest using multiple criteria with little content overlap would give a more complete view and coverage of the training content domain.

Kraiger, Ford, and Salas (1993) provide a multidimensional conceptualization of learning criteria—measures of Kirkpatrick's level two. The authors set out to develop a

"conceptually based scheme for evaluating learning outcomes" (p. 311). Their two assumptions are that learning criteria are multidimensional and that a construct-oriented approach to learning criteria is needed. Cognitive, skill-based, and affective capacities were posited as categories of criteria. Therefore, learning may be inferred from changes in these capacities, or meeting proficiency standards can be inferred from level-of-performance evaluation (Sackett & Mullen, 1993). According to Kraiger and colleagues (1993), learning has been "measured by the extent to which trainees have acquired relevant principles, facts, or skills and could be assessed using traditional multiple choice tests" (p. 311). Learning criteria have often been categorized as verbal knowledge and behavioral skills only. Kraiger and associates (1993)—"drawing from Bloom's (1956) and Gagne's (1984) taxonomies" (p. 312)—proposed these three categories of learning outcomes (see Figure 1) to move beyond the limited definition of declarative knowledge and skill.

"Cognition refers to a class of variables related to the quantity and type of knowledge and the relationships among knowledge elements" (p. 313, Kraiger et al., 1993). Cognitive capacities include verbal knowledge, knowledge organization, and cognitive strategies. Cognitive learning criteria encompass the measurement of declarative knowledge assessed by tests—one of the traditional criteria of training evaluation. However, cognitive capacity goes beyond declarative knowledge and includes dynamic processes for the acquisition and application of knowledge. When measuring verbal knowledge (i.e., declarative knowledge), the nature of the test can impact the construct measured—for example, a speed test measures the rate of knowledge access, and a power test measures the accuracy of knowledge access. Knowledge organization refers to mental models, knowledge structures, cognitive maps, and task schemata—all of which can represent the structure of a person's storage of facts,

principles, and other verbal information. Research has shown the type and complexity of the stored knowledge elements and the use of hierarchical structures to organize the elements differentiates experts from novices (Kraiger et al., 1993). Cognitive strategies refer to mental processes that "facilitate knowledge acquisition and application" (p. 315). Mnemonics—such as learning and recalling a grocery list by walking through the rooms of a house in your mind, visualizing objects to be purchased in each room—can be simple or complex mental tools to aid in the acquisition and recall of knowledge.

Skill-based criteria involve learning technical or motor skills (Kraiger et al., 1993). The definition can be expanded to cover contextual skills as well (Motowidlo et al., 1997). Traditionally, simulations or work samples have been used to measure skill acquisition. Kraiger and associates (1993) have categorized skill development into three phases: initial acquisition, compilation, and automaticity. Initial acquisition refers to the process by which declarative knowledge is transformed to procedural knowledge—in other words, knowing what to do becomes knowing how to do it. Acquisition of procedural knowledge must occur for the "reproduction of trained behavior" (p. 316)—in other words, procedural knowledge is a precursor for procedural skill.

Compilation—as the result of proceduralization and composition—occurs with continued practice of the trained skill. Proceduralization refers to creating a behavioral routine to accomplish a specific objective from smaller, discrete behaviors. Composition is the mental linking of successive steps in a procedure into a more complex process—this happens simultaneously with proceduralization and can continue beyond its end. Compilation results in faster, more fluid proficiency without as many errors. Measuring compilation requires the use of highly specific criteria that reflect maximum performance—

methods like target behavioral analysis, hands-on measures, and structured situational interviews are recommended (Kraiger et al., 1993).

Automaticity refers to an operational shift in behavior from controlled to automatic processing. Performance becomes more fluid, expert, and individualized to the person. As Kraiger and colleagues (1993) point out, "with automaticity, individuals have greater cognitive resources available to cope with extraneous demands" (p. 318). When a process or task becomes so well learned that it can be performed with little or no conscious control, automaticity for that task or process has taken place. One common example would be driving. Most people have been driving long enough that they can do other things—like a have conversation with a passenger—while driving because driving has become a well-patterned behavior requiring less cognitive resources under normal driving conditions than it initially did.

Affective learning criteria are the final category and relate to learning attitudes and motivational constructs. According to Kraiger and associates (1993), "Gagne (1984) included attitudes as a training outcome…[and]…defined an attitude as an internal state that influences the choice of personal action" (p.318). Affective outcomes are included because cognitive and behavioral measures provide only a partial picture of learning or training proficiency. Affective measures encompass attitudes, motivations and goals that are specified as objectives of training. This category tends to focus on two types of outcome measures—"those that target attitudes or preferences as the focus of change and those in which motivational tendencies are an indirect target of change" (p. 319, Kraiger et al., 1993). Attitudes can relate to measuring organizational commitment or generational differences.

Measurement of attitudes generally uses a Likert-type scale and assesses direction and strength of the attitude.

Motivational outcomes of interest include motivational dispositions, self-efficacy, and goal setting. The authors discussed trainees adopting a mastery orientation versus a performance orientation as the major disposition impacting training. Self-efficacy refers to a person's perceptions of their ability to perform a specific activity—this relates to performance by influencing activity choices, effort expenditures, and persistence. The final motivational process proposed by the authors is goal setting. According Kraiger and colleagues (1993), "the mechanisms presumed to operate through goal setting are also those that characterize motivated behavior: direction, arousal, and persistence of effort" (p. 321). Individual differences are posited to work through self-management processes (i.e., working towards goals), type and structure of goals, and quality of goals. The affective capacities factor sounds similar to one of Campbell's (1999) direct determinants of performance, motivation.

Empirical Research with the Kraiger, Ford, and Salas (1993) Model

Kraiger and associates (1993) encouraged researchers to adopt and empirically study their model. Some research has used the multidimensional Kraiger et al. (1993) model to develop training criteria. For example, Simon and Werner (1996) used the Kraiger et al. (1993) model to develop their training outcomes in a study evaluating three different approaches to computer training. The cognitive learning measures consisted of a test to measure general and procedural comprehension using multiple choice, true/false, and open-ended questions. Skill-based learning outcomes consisted of hands-on measures to assess skill-based learning and transfer. The affective measure consisted of an end-user computer

satisfaction inventory. Simon and Werner (1996) found high correlations between the measures of knowledge and skill, leading them to question if they had thoroughly measured the training performance domain. Ford, Smith, Weissbein, Gully, and Salas (1998) demonstrated the three Kraiger et al. (1993) learning criteria measures—knowledge acquisition, skilled performance at the end of training, and self-efficacy—were related to a transfer task (i.e., some researchers define performance in terms of training transfer). Kozlowski and colleagues (2001) employed multiple Kraiger et al. (1993) training criteria—declarative knowledge, knowledge structure coherence, training performance (i.e., simulation scores at the end of training), and self-efficacy—to study the effect of mastery versus performance goals and orientation on training performance utilizing a path analytic approach.

Two recent research studies (Colquitt et al., 2000; Tracey et al., 2001) have shown support for the Kraiger et al. (1993) multidimensional view of training criteria and for the view that training criteria—such as knowledge acquisition—mediate the relationship between individual and situational differences variables and more distal outcomes like job performance. Primarily, these two articles illuminate the role of self-efficacy and motivation in the training context—measures of these constructs can be used to operationalize affective capacity (Kraiger et al., 1993).

Colquitt and his associates (2000) conducted a meta-analytic path analysis to explore training motivation, its antecedents, and its relationships with training outcomes, such as declarative knowledge, skill acquisition, and post-training self-efficacy. The authors did a narrative review of the literature, developed two models—fully and partially mediated—and conducted a meta-analytic path analysis. Colquitt and colleagues (2000) adopted the Kanfer (1991, as cited in Colquitt et al., 2000) definition of training motivation—"we define training

motivation here as the direction, intensity, and persistence of learning-directed behavior in training contexts" (p.678). Training motivation is influenced by individual and situational characteristics. According to the authors, "training motivation differs from general motivation in terms of its context and its correlates" (p. 685). Although many individual characteristics have been related to learning outcomes, the authors point out that many of these characteristics are more frequently linked to learning outcomes through the intervening variable of motivation to learn. Motivation to learn is the focal point of their models.

In reviewing training outcomes in the literature, Colquitt and associates (2000) acknowledge that many studies have operationalized learning outcomes in terms of Kirkpatrick. The most frequent outcomes in the training literature are declarative knowledge and skill acquisition—level two learning outcomes. In terms of the additional learning outcome in the (Kraiger et al., 1993) framework, the authors report that the only measure of affective capacity that has been studied with any frequency is post-training self-efficacy. Colquitt and his colleagues (2000) chose to use the Kraiger et al. (1993) conceptualization of learning, but they examined Kirkpatrick's behavior level by including transfer of training and job performance in their models. Additionally, they included reactions in the models.

Colquitt and associates (2000) proposed two integrative models of how motivation to learn intervenes between individual and situational differences and learning outcomes—specifically declarative knowledge, skill acquisition, post-training self-efficacy, and reactions—and future outcomes like transfer and job performance. One model was fully mediated—the individual and situational differences variables can only impact learning outcomes through motivation to learn. The other was partially mediated—the individual and situational differences variables also have direct influence on the learning outcomes.

Although both models demonstrated adequate fit, a chi-square difference test (nested models) and the comparison of average residual sizes indicated the partially mediated model was a better fit than the fully mediated model—supporting the notion of direct and indirect influence. See Figure 2 for a view of the partially mediated model.

Colquitt and colleagues (2000) found several interesting individual findings. Conscientiousness was not significantly related to declarative knowledge and skill acquisition, but it was related to post-training self-efficacy. This mirrors the finding in the performance literature—conscientiousness predicts contextual performance but not task performance (Motowidlo et al., 1997). The study found relationships between the learning outcomes similar to those found by Alliger and colleagues (1997) with the exception of the learning and transfer relationship—the corrected correlations were higher than the Alliger et al. (1997) correlations, but Alliger et al. (1997) did not correct for unreliability. Motivation to learn was found to be a significant predictor of all four learning outcomes. Additionally, the partially mediated model explained 87% of the variance in declarative knowledge, 29% in skill acquisition, 86% in post-training self-efficacy, and 47% in reactions. The four learning outcomes explained 53% of the variance in transfer, and 81% of the variance in transfer was explained when the direct relationships of the individual and situational differences variables were included. Transfer was found to explain 35% of the variance in job performance. The findings support the multidimensional nature of training performance and the relationship of learning outcomes as partial mediators of more distal outcomes.

Tracey and associates (2001) found support for the role of pre-training motivation as a mediating variable between individual and situational differences characteristics and training outcomes with a group of managers in training. The CFA model included measures

for affective reactions, utility reactions, declarative knowledge, and application-based knowledge (i.e., procedural knowledge). The results support the findings of Alliger and associates (1997) and Kirkpatrick's implied relationships between training outcomes—"results from the confirmatory factor analysis of the proposed model showed that affective reactions were significantly related to utility reactions, utility reactions were significantly related to declarative knowledge, and declarative knowledge was significant related to application-level knowledge" (p. 19-20, Tracey et al., 2001). The authors revised their model slightly—adding paths to pre-training motivation from job involvement and work environment—to achieve better fit. Tracey and associates (2001) support the idea that training performance is multidimensional.

Finally, Table 2 provides examples of research utilizing the two most prominent training evaluation models—the Kirkpatrick model and Kraiger et al. (1993) model. Obviously, all of the research cannot be listed here. The striking difference relates to the number of adaptations and expansions of the Kirkpatrick model versus none for the Kraiger et al. (1993) model. Because the Kirkpatrick model is over 40 years old and the Kraiger et al. (1993) model is only nine years old, the Kirkpatrick model is better understood and used more, leading to many people trying to adapt it. Many practitioners have attempted to make the Kirkpatrick model more useful (e.g., Swanson & Holton, 1999). In recent years, many studies have started to use the Kraiger et al. (1993) operationalization of learning. Research over time will determine the effectiveness of the Kraiger et al. (1993) model in describing training performance.

Training Evaluation Research Summary

In summary, when conceptualizing and measuring training performance, models of training evaluation provide the best underpinning because of their focus on training criteria. Training criteria can be used to address two categories of evaluation questions (Sackett & Mullen, 1993): How much learning occurred as a result of training? And, does the trainee meet a specified level of proficiency or performance? For the past 40 years, the predominate evaluation model has been Kirkpatrick's four levels. In the past decade, the Kirkpatrick framework has been expanded or adapted (e.g., Alliger et al., 1997). Although all levels of training evaluation were reviewed, this research focuses on Kirkpatrick's level two criteria—the learning that occurs during training. When concerned with individual proficiency in training (i.e., training performance), the focus shifts to measures of learning criteria. Traditionally, learning has been conceptualized and measured as a change in declarative knowledge and/or skill—usually just declarative knowledge—during training. Kraiger et al. (1993) have expanded level-two evaluation to three categories of learning criteria—cognitive, skill-based, and affective. The Kraiger et al. (1993) model clearly suggests that training performance—as defined in terms of learning proficiency—is multidimensional. Various measures—including tests, work samples, and peer ratings—can be used to assess individual proficiency in these categories. Kraiger and associates (1993) suggest that certain measurement methods are more appropriate for certain categories of training performance (e.g., simulations or work samples for skill-based criteria).

The models of training evaluation (e.g., Kirkpatrick, 1979) and the empirical research (e.g., Colquitt et al., 2000) reviewed in this section suggest the following is known: (a) unlike job performance research that has its own literature, training performance research is

conducted in the context of training process and outcome research; (b) effective training evaluation requires multidimensional criteria; (c) the Kirkpatrick model has been the predominate training evaluation model for 40 years; (d) recently, the Kraiger et al. (1993) model has started to be more frequently used; (e) the goal of evaluation determines the criteria to be measured and the measurement method; (f) training performance is the learning demonstrated during training; (g) a multidimensional model of learning criteria, like the Kraiger et al. (1993) model, is required to measure training performance; (h) training performance, when defined by the Kraiger et al. (1993) learning criteria, partially mediates the relationship between human attributes and job performance (Colquitt et al., 2000); (i) training performance can be used to predict more distal performance (Goldstein, 1993); and (j) no training performance models have been fully operationalized and tested using CFA or SEM techniques.

The following are not known about training performance and are of particular interest: (a) the exact nature and structure of training performance; (b) more specifically, the exact number of training performance factors; (c) the importance of the level of content specificity (or independence) to measuring training performance and testing training performance models; (d) the change in the latent structure of training performance over time; (e) the relationship between training performance and job performance determinants and factors; (f) the most appropriate measures and operationalizations of training performance factors; (g) the impact of the application of CFA or SEM techniques to training performance modeling; and (h) integration of training performance with job performance models.

Integrating Training Evaluation and Job Performance Research

The goal of this section is to provide a rationale for integrating the training evaluation and the job performance literatures. The Kirkpatrick framework is used as an integration tool for training evaluation and job performance research and models. Then, the question "Can Kraiger et al. (1993) learning criteria be used to operationalize direct determinants of performance?" will be addressed. The issue of when training criteria can be used to model the direct determinants of job performance in the Campbell model is discussed as well. This section concludes with a summary.

The training evaluation and job performance literatures contain similar constructs that refer to knowledge, skill, and a third factor—motivation (Campbell, 1999), habits (Motowidlo et al., 1997), citizenship (Wilson & Grant, 1997), and affective capacities (Kraiger et al., 1993). In the training evaluation literature, these constructs are learning criteria (Kraiger et al., 1993)—they are evaluated to assess change due to training or against a desired level of proficiency (Sackett & Mullen, 1993). In the performance literature, these constructs are the direct determinants of job performance components (McCloy et al., 1994) or performance itself (e.g., Wilson & Grant, 1997).

The goal of training is for participants to successfully complete relevant knowledge-based, skill-based, and affective learning objectives. These learning objectives—that should have been identified from a training needs assessment or job analysis—define performance in training. Training does not occur in a vacuum. These outcomes are impacted by individual and situational characteristics (Kraiger et al., 1993; Colquitt et al., 2000; Ramirez, 2000; Salas & Cannon-Bowers, 2001) and impact more distal performance on the job (Goldstein, 1993). In essence, individual differences and situational characteristics—including the

training content and context—impact the participants' existing levels of knowledge, skill, and affective capacities. Assuming the training is appropriately designed to interact with individual and situational characteristics, pre-training levels of the relevant learning capacities (i.e., defined by objectives) should be increased at the end of training. This increased capacity should impact relevant job performance if the environmental conditions are correct. Training criteria have been found to partially mediate the relationship between human attributes and job performance (Colquitt et al., 2000).

Performance determinants can be direct or indirect. Performance factors are a function of direct determinants—declarative knowledge (DK), procedural knowledge and skill (PKS), and motivation (M) in Campbell's model—and direct determinants are a function of indirect determinants (i.e., individual and situational differences variables). In other words, Campbell's model stipulates that the relationship between indirect determinants (e.g., cognitive ability) and job performance components (e.g., "non-job-specific task proficiency") is fully mediated by the direct determinants. Others researchers agree with the fully mediated model (e.g., Motowidlo et al., 1997). Therefore, levels of the direct determinants of performance—DK, PKS, and M for Campbell—determine performance at any given instant.

To summarize, these models suggest that training performance—specifically learning criteria—mediate the relationship between human attributes and job performance and that direct determinants of job performance mediate the relationship between indirect determinants (i.e., human attributes) and job performance. These learning criteria—knowledge, skill, and affective capacities—and direct determinants of performance—DK, PKS, and M—are similar in definition and function. However, an interesting issue here is

that the training literature supports partial mediation and the job performance literature supports full mediation.

Using Kirkpatrick as an Integration Framework

Using the Kirkpatrick (1996) framework to integrate the different models should make the relationship between training performance and job performance determinants more apparent (see Table 3). Kirkpatrick specified four levels of training criteria: reactions, learning, behavior, and results. Although his framework is often criticized and misunderstood, it is a useful categorization and organizing framework nonetheless (Surface, 2000; Salas & Cannon-Bowers, 2001). When discussing the overlap between training and job performance, Kirkpatrick's learning and behavior criteria are the most relevant levels. Performance in training is achieving the expected level of proficiency on the relevant learning criteria, as specified by the training objectives. Job performance is behavior on the job that is meaningful to the organization and that is under a person's control.

Level One

Reactions—Kirkpatrick's first level—are attitudes toward training, not measures of attitudinal, affective, or motivational learning objectives. Although reactions have been shown to be predictive of more distal outcomes (Alliger et al., 1997; Colquitt et al., 2000), this level is not important to our discussion. There is no directly corresponding construct in the job performance literature, except maybe job satisfaction.

Level Two

Kirkpatrick's second level, learning, is typically thought of as declarative knowledge and skill acquisition. Learning outcomes have been expanded to include three categories of criteria—cognitive (e.g., declarative knowledge), skill-based, and affective criteria (Kraiger

et al., 1993). These learning criteria have been empirically used to measure training performance (e.g., Simon & Werner, 1996). Colquitt and associates (2000) demonstrated with a meta-analytic path analysis that learning criteria (e.g., skill acquisition) have direct impact on more distal evaluation criteria like job performance and partially mediate the relationship between individual and situational antecedents and job performance.

From job performance research, the direct determinants from Campbell's model—DK, PKS, and M—would fit at this level. These are very similar to the Kraiger et al. (1993) learning outcomes above and mediate the relationship between individual differences and job performance as well. In addition, Motowidlo and colleagues (1997) proposed performance determinants based on their contextual and task performance framework. They suggest six performance determinants: contextual knowledge, contextual skill, contextual habits, tasks knowledge, task skill, and task habits. Their model—influenced by McCrae and Costa (1996)—suggests that motivation is embedded in the entire system and, therefore, does not need to be modeled as a separate construct—however, habits are response patterns to situations that include motivational and non-motivational response patterns. Previous job performance research, such as Hunter (1983), has demonstrated the mediating relationship of knowledge and skill between individual differences and job performance.

Kirkpatrick (1996) indicates that four conditions must be in place for a person to change their behavior: (a) must have a desire to change; (b) must know what to do and how to do it; (c) must work in an culture that supports the change; and (d) the organization's reward systems must provide incentive for change. These conditions suggest that motivation, knowledge and skill, and fit with the organizational culture and policy are all necessary for a change in performance. Interestingly, this is very similar to how the performance

determinants influence job performance in the Campbell model. It also suggests that learning at level two is necessary but not sufficient for behavioral change. "Learning may not be manifest in subsequence job behaviors" because "more distal criteria, such as behaviors, are susceptible to environmental variables that can influence the use of trained skills and capabilities" (p. 26, Ramirez, 2000). Other research (e.g., Tracey, Tannenbaum & Kavanagh, 1995; Tracey, Hinkin, Tannenbaum & Mathieu, 2001) supports the importance of environmental or situational factors in determining the success of training in the work environment.

Level Three

Kirkpatrick's third level is behavior on the job—did the training impact how you perform your job (i.e., job performance)? In training criteria research, the behavior level is often operationalized as transfer of training or job performance (Colquitt et al., 2000). Training is designed to train participants to do their job. In the job performance literature, behavior is operationalized as behavioral performance factors, not the outcomes of the behaviors. Campbell's eight performance components or the task and contextual performance factors (Borman & Motowidlo, 1993) fit in this level. The performance content factors of any job performance model would fit at this level.

Level Four

Kirkpatrick's fourth level, results, refers to organizationally desirable outcomes of learning or behavior, like effectiveness, productivity, utility or profits. Things that Campbell and his colleagues (1993) are quick to distinguish from performance. However, Campbell (1999) suggests that job performance may be considered a determinant of organizational performance factors—although he does not specify how to define these organizational

factors. For Kirkpatrick and others, learning and job performance are distally related to organizational results. This level is beyond the scope of the current discussion.

Summarizing the Kirkpatrick Integration

After the integration analysis using the Kirkpatrick framework, a strong argument can be made that training criteria from levels two and three and components of the job performance models (i.e., direct determinants and content factors) overlap significantly at the conceptual level. Because most organizations use the same job performance measures to operationalize level three criteria for training evaluation as they do to operationalize job performance for performance management, the overlap between the behavior level of training criteria and the job performance content factors can easily be understood, justified, and supported—the only problem arises if you operationalize level three as transfer of training, instead of level of job performance. As for the integration of level two learning criteria and the direct determinants of performance, the conceptual argument supports the idea that they are similar constructs—especially when we are specifically comparing the Campbell determinants with the Kraiger et al. (1993) learning criteria. However, the justification for using learning criteria to operationalize the direct determinants of performance in the Campbell model needs more development.

The central issue is whether level-of-proficiency learning criteria can be viewed as determinants of job performance. As Sackett and Mullen (1993) suggest, when the goal is ensuring a standard level of proficiency, training measures can be viewed as measures of performance, and change in these learning capacities due to training is less important than the absolute level of the capacities. Goldstein (1993) points out that training performance can be used to predict (i.e., is a determinant of) future job performance just as trainability scores can

be used to predict training success. Campbell (1999) agrees that proximal dependent variables can be used to predict more distal performance measures. This argues level-of-proficiency training criteria are appropriate to operationalize the Campbell direct determinants of job performance as long as the measures meet the requirements of the performance model and the standards of "good" criteria. Two other issues are that the training be of sufficient duration to allow for typical and maximal performance to be measured and that the training criteria and performance criteria overlap in content enough to minimize content contamination and deficiency between the criteria.

Are Training Criteria and Direct Determinants of Performance Equivalent?

Reading the collective works on Campbell's model (Campbell et al., 1990; Campbell et al., 1993; McCloy et al., 1994; Campbell et al., 1996; Campbell, 1999) and the individual differences model of task and contextual performance (Borman & Motowidlo, 1993; Motowidlo et al., 1997) offers some support for viewing learning criteria as direct determinants of performance, but does not provide a definitive, straightforward answer to the question above. First, Campbell (1999) states that proximal dependent variables (i.e., criteria or performance) can be viewed as determinants of more distal performance variables. Although he was referring to job performance being a determinant (i.e., predictor) of organizational performance, the principle is the same for training criteria being a determinant of job performance.

As Campbell and his associates (1993) state, "training programs…can only affect performance by changing an individual's declarative knowledge, procedural knowledge and skill, and/or motivation" (p. 61). This statement supports this proposal's view that training criteria—when defined in terms of Kraiger and associate's (1993) learning criteria—can be

used to operationalize Campbell's direct determinants. Training objectives specify the changes, or levels of proficiency, in cognitive, skill-based and affective capacities. The training program is designed to influence these capacities. In turn, these levels of cognitive, skill-based, and affective capacities affect job performance, if the work situation allows them to have an impact (i.e., situational factors influence the on-the-job use of capacities learned in training). When we say training influences performance, we actually mean that the levels of proficiency in the cognitive, skill-based and affective capacities that were acquired through or certified during training have an influence on performance.

These two specific statements seem to support the notion that training criteria can be viewed as the direct determinants of job performance. However, several assertions in the writings of Campbell and his associates are less clear—they can be viewed as not supporting, conditionally supporting, or supporting the idea depending on the interpretation. The following Campbell model assumptions are discussed to demonstrate how they can be interpreted to support the learning criteria as direct determinants view: (a) determinants are not performance; (b) performance is defined as goal-relevant action; and (c) training criteria are determinant deficient.

Determinants Are Not Performance

McCloy and colleagues (1994) state very clearly that "determinants are not performance" and that training influences performance only through the direct determinants of performance—DK, PKS, and M. It is unclear if the authors mean that "determinants are not [job] performance", "determinants are not [any type of] performance" in the absolute sense, or "determinants are not performance [in the same model at the same time or using the

same measures in the same model]". Additionally, the following statement from Campbell (1999) elaborates on this idea:

> "Knowledge, skill and choice behavior are not themselves components of performance. Individual differences on any component of performance are a joint function of individual differences in knowledge, skill, and choice behavior" (p. 408).

In other words, direct determinants cannot be performance because performance is a function of direct determinants.

First, as Campbell (1999) and Goldstein (1993) suggest, proximal dependent variables can be viewed as determinants of more distal performance. Supporting this view, Kanfer (1990, as cited in Colquitt et al., 2000) suggests a distal-proximal framework for motivation—more distal individual differences variables influence performance through more proximal ones. *Determinants* are not *job* performance. *Training* performance is not *job* performance—unless your job is to be trained. Similarly, it would be inappropriate to view *job* performance as *team-level* performance in a model positing that individual job performance is a determinant of team performance—in this situation, job performance measures are determinants and not the performance of interest. The basic idea is that performance measures in one model can be measures of direct determinants of performance in another model. When determinants are operationalized using performance factors, those performance factors cease to be viewed as performance and must be viewed as a determinant factors with the properties for the determinants specified by the model in question.

Second, as for the specific statement that DK, PKS, and M are not performance because performance is a function of these determinants, this statement reflects a limited view of the definition of performance. Training objectives specify changes to or proficiency

levels of cognitive, skill-based and affective capacities that should result from training, and the training design creates a process to accomplish these objectives. The training design designates instructional methodologies (e.g., lecture) and learning strategies (e.g., spaced practice) to affect change in existing levels of relevant cognitive, skill-based and affective capacities if they are below the desired level of proficiency. Training criteria—when defined in terms of level-of-performance goals and the Kraiger et al. (1993) model—measure the trainees' level of proficiency in these capacities. From this standpoint, the level of proficiency on any training criterion measure is a joint function of previous levels of cognitive, skill-based, and affective capacities related to the training objective. This is consistent with Campbell's idea that performance components are a joint function of DK, PKS, and M—therefore, training performance as defined as learning criteria measures is a function of previous levels of learning on those constructs. If you expand Campbell's limited view of training criteria to recognize it as a form of performance, his statement supports the view that training criteria can be viewed as direct determinants and training performance.

Is It Goal Relevant?

A second unclear assertion revolves around a component of the definition of performance related to why the determinants are not performance. McCloy and associates (1994) state:

> "For the purposes of this model, performance was defined as goal-relevant action. Organizations, for the most part, do not reward individuals simply for possessing DK and PKS. The knowledge and skill must be used to accomplish the goal-relevant tasks" (p. 495).

At first glance, since training performance is defined in terms of learning criteria (i.e., basically DK, PKS and M), this statement could be problematic. In most training situations, this assumption would probably be correct because training is often only loosely aligned with relevant job-related goals. Training is often given as reward and has little to do with actual job tasks (Goldstein, 1993). There are very few situations were individuals are paid for their DK, PKS, and M. However, pay-for-skill is an example where pay is tied to DK and PKS. Situations where proficiency in training is necessary to be hired into a position or where skill proficiency is necessary for promotion to a new level or job would be examples where performance in training is definitely goal relevant. This is often seen in the manufacturing industry, police organizations, and special military units. Finally, if the training is based on a job analysis and designed to develop relevant DK, PKS, and M to improve or ensure job performance, then training performance, as defined as learning criteria, becomes very goal-relevant and rewarded.

Training Criteria Are Determinant Deficient

The Campbell framework views training as being "determinant deficient"—because all three of the direct determinants are not allowed to impact performance measures. The rationale is that motivation is held constant because of the standardized training situation and the measurement characteristics of knowledge tests and simulations (i.e., work samples). Measures of training performance are viewed as measures of maximum performance—allowing for variation in knowledge and skill to impact performance but not motivation. According to Campbell and associates (1993), "the standard job sample tries to hold the motivation determinants construct for all individuals (at a high level), whereas archived records do not" (p. 53). For measures—specifically work samples—to be considered

measures of maximum performance, they must meet the following criteria (Dubois, Sackett, Zedeck & Fogli, 1993): (a) constrain performance to the highest level; (b) participants must be given and accept standardized instructions; (c) the individuals must be monitored; and (d) the duration of measurement must be brief. The idea is that participants expend effort by accepting, perform at a high level because of being monitored, and persist because the course constrains them—therefore, motivation, or choice behavior, is not really operative. Additionally, Campbell and his colleagues (1993) suggest that training evokes "uniformly high motivation generated by the standardized situation versus the differential motivation across individuals in the actual work setting" (p. 51).

Does the standardization of training and the measurement characteristics of knowledge tests and work samples hold motivational levels constant? The easy answer is that it depends on the contextual factors associated with the training and their interaction with any given individual. Anyone who has conducted training programs—or taught classes—can relay stories of trainees or students who were not enthusiastic about the course or program or who did not perform at the level of their knowledge and skill on a test or activity. This is especially true for academic courses and long training programs with multiple opportunities to demonstrate performance.

In a short training situation with high demand characteristics for achievement, motivation would most likely be constrained to the highest level for each individual. However, if training is the actual work situation (i.e., job is to be trained) or takes place in the work environment, then motivation is more likely to differentiate across individuals. As the duration of the training and the number of measurements increases—remember for maximum performance the duration of measurement must be brief—motivation has more opportunity to

vary from day-to-day and measurement-to-measurement. This situation becomes more like an academic setting, such as college, were motivation does vary, especially in terms of preparation for announced assignments and day-to-day patterns of behavior, such as studying. Campbell (1999) even acknowledges that an argument can be made that academic grades are determinant sufficient.

In some training programs, like the Special Forces Qualifications Course (SFQC), meeting a standard of performance across one to two years of instruction is the paramount goal. Multiple measures of performance from various instructional modules represent a mixture of a person's maximal and typical performance levels. Over the duration of instruction, motivation is allowed to vary more for some activities and situations than for others. Like college, motivation comes into play in the amount of preparation for pre-announced tests and simulations of maximal performance and in typical day-to-day activities and performance. For these reasons, level-of-performance training criteria are viewed as more appropriate to operationalize Campbell's direct determinants than change-in-proficiency training criteria because job performance criteria typically measure level of performance. Additionally, long-term simulations that incorporate peer and cadre ratings ensure the assessment of individual level typical performance.

In conclusion, it can be argued that training criteria are not determinant deficient if the training and measurement context allow for motivation to vary across individuals. Beyond that argument, it is this researcher's opinion that if motivation is an individual difference, then regardless of the standardized situation there will always be some difference in motivational levels. The choice behavior of an individual should always be a function of individual and contextual factors. In most training instances, the situational demand

characteristics would not be salient and "powerful" enough to evoke the same motivational responses in all participants—one might argue it evokes the same behavior across participants, but the level of effort and persistence would still be under the individual's control.

When can training criteria be viewed as direct determinants?

The integration of the training evaluation and job performance literatures suggests a set of criteria for when training criteria measures can be used as direct determinants of performance. The following criteria are suggested for when using training criteria to operationalize training performance and the direct determinants of job performance is appropriate.

First, training performance is multidimensional and defined as three-factor model that approximates the framework of Kraiger and colleagues (1993)—cognitive, skill-based, and affective capacities—and the direct determinants of Campbell—DK, PKS, and M. Second, the measurement philosophy is that of the level-of-performance evaluation suggested by Sackett & Mullen (1993)—in other words, individual proficiency in the cognitive, skill-based and affective capacities is important to the organization. Third, training criteria must meet the assumptions for performance determinants in the relevant model of job performance in addition to the assumptions for training performance. In general, the criteria should meet the standards of good criterion measures; specifically, they should not be contaminated, deficient, or unreliable. Fourth, the training must be related to the activities on the job, and the training criteria content must overlap significantly with the content of the performance measures. The level of generality or specificity of the criteria should match as well. Situations where the training is a pre-requisite for doing the job are definite opportunities.

Fifth, training should be the only focus of the trainee, regardless of duration. The trainee should be paid for developing or demonstrating proficiency in relevant learning capacities. Sixth, training should be of sufficient duration and include enough measurement opportunities to allow for the observation and evaluation of maximum and typical performance episodes. In other words, the training should allow individual motivation to vary. The trainee should be free enough to choose to continue the training and the level of effort to exert to complete the training. The training environment must allow for the operation of previous levels of cognitive, skill-based and affective capacities. Seventh, multiple indicators from across the entire training process should be used to operationalize the three latent factors. The measurement model must be representative of the training and the performance content domain. Finally, the training environment should be reasonably standardized and should approximate the performance environment when appropriate for the development or demonstration of a context-specific skill.

These are a suggested set of standards for when it is appropriate to use training performance as direct determinants of job performance. The relevant literature has been reviewed, and an argument for integration has been presented. Research is needed to determine if all of the criteria are appropriate and the extent to which they should be applied.

Integration Summary

The goal of this section was to argue that training criteria—when defined and measured in terms of level of proficiency on Kraiger et al. (1993) learning criteria—could be both training performance and the direct determinants of job performance, especially as conceptualized in the Campbell model. As presented earlier, a strong rationale exists for this integration. This author's conclusion is that training criteria can serve as both. Training

context, the measurement model for the learning criteria, and its relation to job performance content determine the appropriateness of the dual conceptualization. A list of criteria for when training performance can function as performance determinants was suggested. Regardless of the relationship between training performance and job performance determinants, a conceptual question of interest is "Do training performance and job performance share the same latent structure, making them special cases of a general performance model?" Is there a general performance model that is content, process, and context independent? This section proposing the integration of training and job performance research ends by suggesting a potential answer for the last few questions. Since job performance often focuses on the knowledge, skill, and motivation to do the job (Kraiger, 1999), the similarities between various models—the learning criteria of the Kraiger et al. (1993) model, the performance determinants of the Campbell model, the determinants in the Motowidlo et al. (1997) model, and the performance factors in Wilson & Grant (1997)—suggest that performance might be best conceptualized as a general three-factor model. Can all performance be conceptualized by knowledge, skill, and affective constructs? This researcher suggests this is a strong possibility. Much more research is needed.

Although the integration highlights commonalities and provides a useful framework for discovery, it illustrates how little is known about the structure and nature of performance. Many questions abound and/or have yet to be addressed by empirical research. This study proposes four integration-inspired research questions in the next section related to the structure, stability, content specificity, and modeling of performance.

Integration Inspired Research Questions

Although the integration of training evaluation research and job performance research suggests possible answers to some questions, there are many questions that are unanswered and many assumptions that are untested. Although it is clear that training performance and job performance are multidimensional, what is the nature and structure of those dimensions? How many dimensions are there? How are they related? How is performance in each dimension best operationalized? If learning criteria that are measured under the level of proficiency philosophy can be considered training performance, is the Kraiger et al. (1993) model—cognitive, skill-based, and affective—the best conceptualization? Is the Campbell model the best conceptualization of performance? Does the structure of the performance model change over time? Is the structure of performance and the relationships between the performance dimensions independent of context and content? Do the performance and determinant factors in the model need to have an equivalent level of content specificity? Or, can general determinant factors be used to predict specific content factors? Can the Kraiger et al. (1993) criteria be used to operationalize Campbell's (1999) direct determinants of performance? Are training performance and job performance separate construct models, are they different cases of a general performance model, or is training performance a component of the job performance model? What methodologies should be utilized to model performance? How are these techniques best utilized? Does the conceptualization of performance influence the choice of technique? These questions have yet to be answered definitively.

This study investigates four issues related to modeling training performance by building on the training evaluation and job performance research reviewed. The four research

questions are introduced next.[1] First, the two research questions related to the latent structure of training performance are presented. Next, the two research questions related to performance modeling issues are discussed. Although important, these issues pertain to modeling performance in more general terms and do not relate directly to determining the overall latent structure of training performance. However, question three does address the content specificity issue with a single training performance factor, BE. For the fourth question, a brief review of the literature related to overfitting CFA and SEM models is presented to adequately frame the question.

What Is the Dimensionality of Training Performance?

A three-factor model is proposed to describe the latent structure of training performance. The model resembles the three learning criteria categories—cognitive (e.g., declarative knowledge), skill-based, and affective capacities—from the training literature (Kraiger et al., 1993), the direct determinants of performance—DK, PKS, & M—specified in the Campbell model (McCloy et al., 1994; Campbell, 1999), and the determinants of task and contextual performance (Motowidlo et al., 1997)—task knowledge, task skill, task habits, contextual knowledge, contextual skill, and contextual habits. This model is similar to the three-factor models confirmed by Grant (1996) and Wilson and Grant (1997) for job performance. At the most basic level, this question provides a complete operationalization and confirmation of the Kraiger et al. (1993) framework applied to level of proficiency criteria (Sackett & Mullen, 1993). The latent factors are labeled using the BE KNOW DO terminology from the U.S. Army's BE KNOW DO model of leadership (U.S. Army, 1999)—this terminology is discussed in the methods section.

Question One. What is the dimensionality of training performance? Specifically, does the BE KNOW DO model represent the latent structure of training performance in the SFQC for general SF soldiering content (i.e., measures from across training related to non-job-specific content)? How does the fit of the basic three-factor model compare with the fit of several alternative models?

Does the Latent Structure of Training Performance Change Over Time?

Schmitt and Chan (1998) and Campbell (1999) suggest that changes in the nature of performance over time need to be studied. Schmitt and Chan (1998) report that change in performance over time has been studied in three ways: (a) changes in mean performance, (b) changes in rank order of individual's performance, and (c) changes in the dimensionality of performance (i.e., latent structure). This study is concerned with changes in the latent structure of training performance over time and training context. The predictive relationships between the latent constructs measured during the initial phase and their counterparts measured during a later phase are of primary interest. MacCallum and Austin (2000) suggest that there are two types of longitudinal designs for studying multiple measurements across time. One involves modeling the relationships between different variables measured at different times, and the other involves repeated measurements of the same variables at different times. They point out that many model incorporate both philosophies.

Question Two. Does the latent structure of training performance change over time? Does the BE KNOW DO model of the latent structure of SFQC training performance change over time between Phase One and Phase Three? In other words, does the BE KNOW DO model provide adequate fit for performance in both phases? What is the

relationship between the latent structure of Phase One and that of Phase Three? Do Phase One constructs predict their counterparts in Phase Three?

Does one general factor or several specific factors describe BE?

Another important issue is whether performance factors and determinants are best operationalized as being performance content dependent (more specific) or performance content independent (more general). This question seeks to suggest whether the latent structure of performance factors is more appropriately modeled as several content-specific factors or as a general factor. The focus is really on which modeling strategy is more appropriate. In this specific case, the question is whether the affective (Kraiger et al., 1993), motivational (Campbell, 1999), or citizenship (Wilson & Grant, 1997) factor, BE, is best conceptualized as one general factor or as three content-specific factors from the Campbell model—demonstrating effort, maintaining personal discipline, and facilitation of peer and team performance. This question addresses the specificity of the latent factors from both the training criteria and direct determinants viewpoints for the BE factor. Both models may fit the data equally well, which would support the notion that performance can be conceptualized as both content specific factors and as a general performance factor. Such a finding would have important implications for modeling any type of performance in any context.

Question Three. Does one general factor or several specific factors describe BE? Does one general factor or several specific factors describe the latent structure of affective or citizenship training performance? Does one general factor or three of Campbell's more specific performance factors—demonstrating effort, maintaining personal discipline, and facilitation of peer and team performance—best represent the

latent structure of BE in SFQC? In general, are constructs best modeled as one general factor or several content specific factors?

What is the impact of overfitting on cross-validation?

Although the primary objective of the current research is to understand the structure and nature of training performance, the use of confirmatory factor analysis (CFA) and structural equation modeling (SEM) techniques to study the structure of performance provides an opportunity to address an important methodological issue that has profound implications for developing and validating any model of performance. The specific issue relates to using modification indices to change the original model to better fit the data of the current sample and the impact of this practice on model fit for the cross-validation. As Black (2001) points out, many researchers over fit their models to their sample data without confirming the adapted model on another sample. Millsap (2002) refers to the process of modifying the model based on a sample as respecification and criticizes it as misleading because it is no longer a confirmatory analysis and would most likely "not recover the true model" (i.e., would not cross-validate). These modifications are most problematic when alternative models are not tested and/or cross-validation samples are not utilized (MacCallum & Austin, 2001; Millsap, 2002). This section discusses the issue of overfitting the model to the data and proposes a methodological research question.

Black (2001) recommends three elements for modeling data successfully using SEM: (a) having a model with a strong theoretical basis—if not, use other techniques more suitable for exploratory analysis; (b) having a well-specified measurement model—"although SEM will accommodate measures with reliability of less than 100%, the researcher must always strive for the best possible measures for both theoretical and empirical concerns" (p. 27); and

(c) having a sound modeling strategy, including using nested models and other techniques to avoid the confirmation bias of SEM. As Black (2001) suggests, SEM is a "powerful tool for confirmation of a proposed model, but it is not nearly as well suited to developing a model" (p. 26). However, most SEM programs offer modification indices that suggest the addition or deletion of paths to model. This can be used for theory building. However, Black (2001) provides strong caution about making modifications—"the inclusion of a link between constructs in SEM has a much greater impact conceptually and empirically than the addition of a variable to a regression analysis" (p. 29). Other authors (e.g., Hatcher, 1994; MacCallum & Austin, 2000; Millsap, 2002) would agree and recommend that all modifications be justified theoretically.

Grant (1996) developed a three-factor model of job performance after her a priori models failed to confirm. She utilized the modification indices provided by the software package she utilized. Then, she cross-validated the model on several samples to ensure that the new model was not based on the patterns of variation unique to the original sample. Many researchers who modify models do not cross-validate (e.g., Tracey et al., 2001). Unfortunately, many researchers make modifications without theoretical or conceptual support based solely on empirical evidence that the change improves the model fit—without cross-validation of the modified model no one knows if the relationship will hold and change the theory, or if the modification is simply noise in the sample. As Black (2001) points about the Tracey et al. (2001) article, "again, theoretical support seems lacking…it seems that empirical results, not theory, are guiding analysis" (p. 29). At least, if a cross-validation sample is utilized, there is further support for the modification or there is not very much support for the modification.

In her discussion, Grant (1996) calls for a systemic exploration of the impact of model modification on the fit of the modified model to a cross-validation sample. To what extent does following the modification indices lead to poor fit on the cross-validation sample? According to Loehlin (1998), the technique of splitting the original sample, modifying the model on one subsample, and cross-validating on another subsample is useful. However, upon cross-validation, "the fit to the correlations in the opposite subsample typically improves and then deteriorates, suggesting that after awhile the factoring in the first subsample is merely fitting the idiosyncrasies of sampling error, making the fit in the second sample get worse instead of better" (p. 164). MacCallum, Roznowski, Mar, and Reith (1994) discuss several methods of cross-validation with covariance structural models, including partial validation strategies—an example might be cross-validating factor loadings but not factor covariances.

A study of a simple automatic model fit improvement strategy by MacCallum (1986; as cited in Loehlin, 1998) found that the strategy could not be recommended. The strategy, that consisted of a set of pre-established decision rules, resulted in eight of 20 samples having one legitimate change to the model—unfortunately, seven of those changes were incorrect. None of the samples achieved the "true" model—known a priori—in 20 tries. The strategy employed was the following: if the model fit is poor, make the single change that most improves the fit, repeat this until a non-significant chi square coefficient results, and delete any unnecessary path after testing each path.

As Grant (1996) suggests, more research should address the impact of overfitting the model to the initial sample—capitalizing on the unique aspects of the sample to increase fit—on the cross-validation of the modified model on other samples. A research question is

proposed to explore the impact of this issue. What is the impact of over modifying a structural equations model to fit the data of one sample on the cross-validation of the modified model on another sample?

Software programs typically provide tests—specifically the Wald test and the Lagrange Multipliers—that may be used to help develop a better fitting model (Hatcher, 1994). Often, researchers modify the model and report the modified model in journal articles without cross-validating the modifications on another sample (Black, 2001). Grant (1996) used the modification indices to create a post hoc three-factor model after her initial a priori models failed to confirm. She successfully cross-validated her model on another sample. However, her insights gained from model modification lead to this research question.

To what extent are researchers who overuse the modification indices capitalizing on characteristics unique to the sample and what impact does that have on cross-validation? Black (2001), in an invited critique of the only training research article to utilize CFA to partially operationalize and test the Kraiger et al. (1993) model (i.e., Tracey et al., 2001), indicated that the authors made modifications that were not theoretically justifiable because the modifications improved the model fit, and there was no cross-validation. Would the modifications in question cross-validate? This study will attempt to address the issue by over modifying and fitting the basic three-factor training performance model in research question one across several cross-validation samples.

Question Four. What is the impact of overfitting on cross-validation? What is the impact of modifying a CFA model to the point of capitalizing on the idiosyncratic characteristics of the sample on the cross-validation of the modified model on another sample? Specifically, what is the impact of following the fit indices to modify the

three-factor BE KNOW DO model on cross-validation of the modified model on another sample?

Chapter Two Summary

Chapter 2 reviews the relevant literature and introduces four research questions related to the latent structure of performance, the stability of the latent structure of performance over time, the content specificity of a performance factor, and overfitting a CFA model. In addition to reviewing all the relevant training evaluation and job performance literature and introducing the four research questions, chapter two suggests an integration of the training evaluation and job performance research literatures. This chapter argues that training criteria—defined in terms of Kraiger et al. (1993) learning criteria and measured for level of proficiency evaluation—can be considered training performance and the direct determinants of future job performance. This dual conceptualization provides the underpinning for future research modeling the entire employment process from recruiting to selection to training to job performance. Additionally, a set of criteria for when training criteria can be considered the determinants of performance is provided. The integration section concludes by suggesting that all performance might be best described and studied utilizing a general three-factor model—possibly knowledge, skill, and affective performance. However, much research is needed.

This research addresses a need for more research related to training evaluation and training criteria (Kraiger et al., 1993; Salas & Cannon-Bowers, 2001) and answers Schmitt and Chan's (1998) call for research into the Campbell model and its elements, especially the direct determinants of performance that are thought to mediate the relationships between individual differences and situation variables and job performance variables. This research

can be considered one of the first steps in addressing several macro-level questions posed in chapter 1 that are beyond the scope of the present study—for example, does training performance fully or partially mediate the relationship between human attributes and job performance? To address this question, a three-factor model of training performance that approximates the direct determinants must be confirmed. Additionally, this research answers Schmitt and Chan's (1998) call for research studying performance over time and performance modeling methodology issues and Grant's (1996) call for research into overfitting CFA models to the idiosyncratic characteristics of one sample.

CHAPTER THREE: METHODS

This section details the methodology utilized to research the nature of training performance in the Special Forces Qualifications Course (SFQC). Basically, this study opted for a confirmatory approach (Hatcher, 1994; McArdle, 1996) utilizing confirmatory factor analysis (CFA) and structural equation modeling (SEM) techniques to test a set of hypothesized models that describe the nature and structure of SF training performance. Archival data was collected from training files at the JFK Special Warfare Center and School (SWCS) at Fort Bragg, NC, and used to operationalize and test the research questions introduced in chapter 2. This section provides the following information: (a) an overview of the SF research context, including a discussion of the BE KNOW DO terminology and a description of SFQC; (b) a description of the participants; (c) the data collection procedure utilized; (d) the variables selected to operationalize the constructs; (e) the research models to be tested by question; and (f) the analytic strategy and procedure employed.

An Overview of the U.S. Army Special Forces Research Context

This section introduces the context of U.S. Army Special Forces (SF) research. First, the Army's BE KNOW DO terminology used to label the training performance constructs in this study are formally introduced and discussed. A rationale for why the *third factor* is best conceptualized as BE and not as motivation or citizenship is provided as well. The specific context for this research study is briefly presented. The basics of the *SF pipeline* and SF operations are described. This section concludes with a description of SFQC.

Introducing the BE KNOW DO Terminology

When referring to this study's proposed three-factor model of training criteria, the BE KNOW DO terminology from the U.S. Army's leadership field manual will be used (U.S.

Army, 1999). Although our model does not deal specifically with leadership, the terms have been adopted because they are similar from a philosophical standpoint to what our model attempts to embody in training performance criteria. Additionally, since the study uses an Army sample, these terms are familiar within the Army culture and conjure a sense of pride for soldiers of any rank. But, most importantly, the terms embody the meaning of the three latent factors in the training criteria and job performance determinant research.

According to Army Field Manual 22-100, the BE KNOW DO mantra is a clear and concise statement of what it means to be an Army leader—a model of leader performance. The manual goes on to suggest that the leadership framework should be considered to be greater than the sum of its three parts. According to the Campbell model, performance factors are functions—greater than the sum of their parts—of the three determinants, and latent constructs should be considered to be more than just the sum of manifest measures, although that is how they are defined.

BE

BE embodies a leader of character who embraces Army values and demonstrates those values through leadership actions. In the learning criteria model, the third factor is affective capacity—defined as motivational and attitudinal learning criteria. In the job performance literature, the third factor has been termed motivation (Campbell, 1999), habits (Motowidlo et al., 1997), and citizenship (Wilson & Grant, 1997).

In this study, BE represents the fit between the characteristics and patterns of behavioral responses of an individual—including motivationally related patterns of responses—and those characteristics and patterns of responses that are appropriate in the specific situation for the specific culture. This factor could be operationalized using peer

ratings, cadre ratings, critical incident measures (i.e., positive and negative spot reports), peer evaluations (i.e., rankings), and peer nominations (i.e., This is the soldier that I definitely would or would not want to be deployed with on a team.). These all represent the perception of the individual's fit with the performance situation and culture as judged by observers in the situation and culture. These are fairly straightforward and commonly used measures with the exception of nominations. Goldstein and Barlett (1977, as cited in Goldstein, 1993) used peer nominations during police training and found that the nominations were not correlated with training grades or other measures of training performance. However, the nominations were correlated with the amount of on-the-job training needed to perform at an acceptable proficiency (r = -.43). This suggests peer nominations tap into something not measured by tests and work samples—measures of knowledge and skill during training. Schwarzwald, Koslowski, and Mager-Bibi (1999) concur and provide evidence that peer nominations during training explain unique variation in future performance. Why conceptualize BE as "fit" and not affective capacity, motivation, habits, or citizenship? Clearly, BE as "fit" is related to all these other constructs.

What is the third factor? Is it motivation, affective capacity, work habits, characteristic adaptations, citizenship, or BE (i.e., fit)? This researcher would like to suggest that the third factor is BE, as described above, from the U.S. Army's leadership doctrine (U.S. Army, 1999). When the third factor is operationalized as ratings, awards, or disciplinary actions (e.g., McCloy et al., 1994) in a well-defined culture—such as the U.S. Army—the third factor can be interpreted as the "fit" between the individual's characteristics (as inferred from behavior) and the characteristics, culture, and norms of the organization or situation. This is very similar to the person-organization fit construct, described as one of the

top-three research areas in personnel selection (Borman et al., 1997). Kristoff (1996) provides a full description of the person-organization fit research.

In terms of direct determinants of job performance, declarative knowledge and skill components are universally accepted. However, the third performance determinant (or performance factor) is in dispute. Campbell (1999) calls the third factor motivation. Motowidlo et al. (1997) refer to the third determinant as task and contextual habits. Wilson and Grant (1997) believe it is citizenship. In the training evaluation literature, the third factor is considered to be affective learning criteria (Kraiger et al., 1993). Although a consistent label or definition has not yet emerged, most people agree that it is related to the contextual or citizenship aspects of performance and that there is some connection to motivation, or patterns of motivation, work styles, and preferences. This author puts forth the idea that the Army got it right in its BE KNOW DO leadership mantra. BE suggests that in the Army culture there is a certain way that leaders should behave to demonstrate that they have the character and attributes to lead. In other words, BE suggests a necessary fit between Army values and culture and individual character, values, and interpersonal attributes as demonstrated by salient behavior.

Think about an organization like the Army with a definite culture, value system, and a formal indoctrination of these values and norms. Most aspects of Army life reinforce these values and norms. There are certain human attributes—as demonstrated by patterns of behavior—that are perceived by organizational members to be more desirable. Awards, disciplinary actions, promotions and spot reports (i.e., reports of critical incidents for discipline or praise) are all designed to reward fit or to punish lack of fit with the organization's values, norms, objectives, and performance standards. Ratings, rankings, and

peer nominations—because the raters are indoctrinated in the culture—reflect the individual's fit with culture, norm, values, objectives, or standards of the organization as well as DK & PKS—they probably reflect fit—the third factor—more so than they do DK and PKS. It is overly simplistic to say the third factor is citizenship, motivation or habits—it is a combination of all these things but specifically it is the fit between the individual's level of these constructs and the requirements of the situation. The important thing is the perception of the raters of how all these factors fit with the values, norms, culture, objectives, and standards of the organization and the performance situation. For this study, the third factor—BE—is defined as the fit between the individual's characteristics and situational requirements in terms of character, values, work ethic/style, and motivational patterns—at least when you operationalize it using ratings, rankings, nominations, personnel actions (i.e., awards or disciplinary), or situation-specific critical incidents. The conceptualization of the third factor as "fit" should generalize across situations to various degrees depending on the characteristics of the situation in question (Beaty, Cleveland & Murphy, 2001; Hattrup & Jackson, 1996), with "fit" being better defined in strong situations, where the cues for successful behavior or performance are more salient.

KNOW

KNOW suggests that Army leaders should have a certain level of knowledge to do their jobs effectively and be competent. The leadership model suggests that knowledge is spread across and related to skill domains, suggesting the two are correlated, which is a fairly robust research finding (e.g., Hunter, 1983; Wilson & Grant, 1997). KNOW in the Army framework also encompasses procedural knowledge and skill as well. In our model, KNOW refers to the relevant declarative knowledge or cognitive criteria on which each trainee

should demonstrate a minimum level of proficiency by the end of training. KNOW can be operationalized using tests, the traditional measures of declarative knowledge, although Kraiger et al. (1993) suggest other measures of DK as well. The KNOW construct can only be operationalized using tests in this study.

DO

The DO concept suggests an Army leader does what is right to achieve success. The manual lists leader actions, including influencing, operating and improving. In this study, the DO factor will refer to the level of proficiency on skill-based criteria measured during training—in other words, do you have the skill and can you put it in action? Measures in this category include work samples, simulations, ratings of technical proficiency, and times to criterion measures among others. In this study, the choices for measures of DO are primarily times to criterion measures, skill proficiency ratings, and number of critical incident reports.

Special Forces Context

United States Army Special Forces (SF)—often referred to as the Green Berets—is an elite group of infantry soldiers that are selected and trained to execute a number of missions critical to the national security interests of the United States. These forces operate quietly, behind the scenes in countries around the world, conducting missions that range from training foreign troops to counterintelligence to providing humanitarian assistance. The precision with which SF soldiers perform their missions and their willingness to play the role of unsung silent partners have earned them a reputation as *quiet professionals*. SF soldiers, who volunteer for a rigorous selection and training program, are among the best the Army has to offer.

Entering SF begins with recruiting soldiers from regular Army units and using a prescreening process to qualify soldiers to attend Special Forces Assessment and Selection (SFAS) at Fort Bragg, NC. A more detailed description of SFAS can be found in Brooks (1997) or Marrs (2000), which is beyond the scope of this study. If a soldier meets the standards of SFAS, he proceeds on to the Special Forces Qualifications Course (SFQC).

SFQC is the training component of the SF pipeline and is divided into three distinct phases that can take up to two years to complete—depending on training needs, specialty area, course sequence, and need for recycling (i.e., repeating a portion of the course usually for poor performance or injury). Each enlisted soldier is assigned to one of four military occupational specialties (MOS), which are the weapons sergeant (18B), the communications sergeant (18E), the medical sergeant (18D), and the engineering sergeant (18C). The officers are assigned to the 18A MOS. The length of SFQC varies for the different MOS. Phases One and Three deal with general small unit tactics and SF mission training.[2] The second phase of SFQC instructs soldiers in their MOS. During SFQC, the soldier's only responsibility is the successful mastery training content and completion of the course, and the mission of SWCS is to ensure that SF soldiers are trained to standard to be successful in the field.

"Special Forces differ from conventional forces in that they are organized, trained, and equipped to achieve military, political, economic, or psychological objectives by unconventional means" (p. 5, Brooks, 1997). Another difference is that SF soldiers have a regional orientation and are assigned to one of five regionally oriented SF Groups. Because of their special operating circumstances (i.e., autonomous small groups operating independently in developing nations), SF soldiers are expected "to exercise more initiative,

self-reliance, and flexibility compared to conventional soldiers" (p. 5, Brooks, 1997). Because of their special skills and reputation for effectively and quietly getting the job done, SF soldiers receive complex and important missions, for which the penalties for failure can be relatively high—the costs can be high for the soldiers personally and for the national security interests of the United States (Brooks, 1997). This makes selecting the appropriate personnel and training them to the appropriate level of mastery paramount activities. Because this research focuses on SF training, a detailed description of SFQC follows.

Special Forces Qualifications Course (SFQC)

SFQC is the training component of the SF pipeline and is conducted by SWCS at Fort Bragg, NC. Soldiers who successfully complete the SFAS course can proceed to the SFQC. The length of time spent in SFQC varies by MOS and previous training and can range from one to two years. It should be noted that SFQC has recently been restructured slightly. The information provided here about the course structure describes the course during the period when the majority of the participants in the data set attended SFQC, a period prior to the restructuring.[2]

SFQC is conducted in three phases. Phase One of SFQC targets the instruction and development of general skills and knowledge related to land navigation and small unit tactics (SUT). A land navigation written exam covering concepts like map reading and several practice field-training exercises (FTX) are graded. The final "skill" examination for land navigation is administered at the individual level and involves navigating to four predetermined points within a set amount of time (i.e., the STAR FTX)—some participants take up to five tries over two trips through the course to pass (i.e., they were recycled during their first attempt). Small unit tactics are taught at the group level. Additionally, the small

unit tactics examination is administered both at the individual and group level. There is a written examination taken individually, and a practical examination done at the group level—for the practical exam, each person has a task as part of their team's mission and is evaluated against criteria for the mission, receiving a GO or NO GO.

During Phase Two of SFQC, the specific skills required by each MOS are trained. This study does not utilize Phase Two data, therefore, no description is provided. Phase Three involves further training and examination of specific Special Forces skills in the context of four of the five primary SF missions (Russell, Crafts, Tagliareni, McCloy & Barkley, 1994). Specifically, the missions of unconventional warfare (UW), foreign internal defense (FID), direct action (DA), and strategic reconnaissance (SR) are trained. Since UW missions involve the other three missions (Russell et al., 1994), UW is the primary focus in the Phase Three FTX—Robin Sage. Other aspects of the SF mission are taught and evaluated in different sections of Phase Three. The Isolation exercise prior to Robin Sage focuses on mission planning.

Robin Sage is a field simulation designed to present the same challenges found in operational settings. Before deploying on the FTX, the team spends time in "isolation" planning their mission—during this time they are observed and tested on their mission. During Robin Sage, the teams are sent out to conduct operations in an extremely realistic training exercise. Over the course of the Robin Sage exercise, the teams are constantly monitored and are rated on their performance at several points in time. Their primary mission is to infiltrate into the foreign territory, to link up with the guerilla forces, and to train them to fight an unconventional war against the occupying forces. The simulation is conducted in a rural area of the Piedmont region of North Carolina for over two weeks. The

entire area is designated as another country, Pineland, and both enemy and friendly forces and natives are role played by volunteers and soldiers from support units. During Robin Sage, each team is composed of twelve to fifteen men, each of whom fills a particular role within the team. These roles are associated with the five primary SF MOS and mission requirements.

Participants

Soldiers who graduated from the Special Forces Qualifications Course (SFQC) from late 1996 to early 2000 are the participants. These soldiers graduated from SFQC in classes 4-1996 to 1-2000—there are four classes a year, so 1-2000 is the first class of the year 2000, and 4-1999 is the last class of the year 1999. Data from 1441 SF soldiers were collected from their training folders. All five SF military occupational specialties (MOS) are represented in the sample: 294 participants from 18A (team leader); 359 from 18B (weapons sergeant); 321 from 18C (engineer); 202 from 18D (medic); and 265 from 18E (communications sergeant). Since the research focuses on training performance that is consistent for all SF soldiers, a detailed discussion of each SF MOS is beyond the scope of this study. The majority of soldiers in the study graduated SFQC in 1998 (n = 502) and 1999 (n = 520), with 323 graduating in 1997. Only 84 participants who graduated in 2000 and 12 who graduated in 1996 were included in the sample. The original intent was to include only graduates from 1997 through 1999 but some folders from 1996 and 2000 were available and therefore included.

Data Collection Procedure

This section covers the issues and procedures related to collecting the data used to operationalize the training performance models in this study.[3] The following topics are

presented and discussed: (a) the initial review of the available data; (b) the potential manifest indicators of the constructs; and (c) the collection and processing of the data.

Reviewing the Available Data

In determining the data review and collection strategy, a source for working with archival data (Elder, Pavalko, & Clipp, 1993) was consulted as a guide. The first step in evaluating the data collection options was to conduct a review of all existing data sources and to determine whether they were sufficient to address the research goals and questions. All sources of available SFQC data were identified and evaluated for compatibility with the project goals as part of the initial data review. Although some training data existed in computerized databases, the majority of the SFQC data of interest was archived in the training folders of individual soldiers who graduated from the course. The results of the data review indicated the data collection should focus on extracting data from the folders. Although limited, the available computerized data were requested as well.

Since the folders were the primary focus of the data collection effort, rosters of SFQC graduates from each class for each SF MOS from SFQC Class 2-1997 to 4-1999 were obtained. These classes were selected to maximize the chance of a match with the field performance data for modeling the entire SF pipeline. Plus, the initial review determined these years had the best chance of containing a consistent set of measures. The folders were pulled from multiple locations and organized for in-depth review. A comprehensive evaluation of the measures contained in the folders was conducted. A sample of three folders from each training class for each MOS was selected at random and evaluated. The measures from the sample of folders were assessed to determine consistency and standardization within and between the SF MOS. It was determined to collect every measure in the folders that met

the goals of the research and that were consistent enough to ensure minimal sample size for the analyses.

Identifying Potential Measures of the Constructs

After the initial review of the folders, a list of measures was created to guide the data collection. However, the specific measures that would be used to operationalize the research questions were unknown at that time. These measures would not be identified for certain until after the data collection was completed. The list of the potential measures identified from Phase One and Phase Three is presented in Table 4.[2] To ensure that there was enough data available in the folders to operationalize the research constructs, the potential indicators were grouped by construct. Table 5 provides a list of what was thought to be the most likely manifest measures for each training performance construct. After being satisfied that the data were available to support the research questions, the Herculean task of collecting the data started.

Collecting the Data

Several methods of data collection from the folders were explored. The optimal method would have been to enter the data directly into a database on a computer. However, resource limitations (i.e., no computers and limited labor) prohibited this approach. To allow for maximum efficiency under the circumstances, a computer-scanned form was developed and used to capture the data. This form contained blanks to capture the measures from Phases One and Three identified in the previous section. These measures were common across all five of the MOSs. Additionally, the MOS-specific measures from Phase Two were included for four of the MOSs—18B (weapons), 18A (officer), 18C (engineer), and 18E (communications). 18D (medic) Phase Two measures were not included due to the vast

number of high quality variables in the medical training database. The four-sided form allowed for multiple people to collect data independently. Appendix A contains a reproduction of the form.

Prior to beginning, all folders were moved to a centralized location for data collection. At first, a storage room at SWCS was used. However, due to access and security concerns, the folders were moved to the ARI offices at the United States Army Special Operations Command (USASOC) Psychological Applications Directorate (PAD). At PAD, a high level of control and security were maintainable. Additional labor was used to help with the data collection. These workers, who were trained and supervised by the researcher, collected the data using the form in Appendix A. The researcher collected data as well. The transient nature of the labor pool, the lack of organization/standardization in most folders, the large numbers of data points, and the requirement of making judgments about how to record certain pieces of information made data checking prior to scanning the forms a necessity. The time to complete relatively well-organized and straightforward folders ranged from 20 to 30 minutes. Due to the large number of errors, the researcher spent nearly four months reviewing each folder and correcting any mistakes on the corresponding form. After the data was checked for errors, all the forms were reviewed to facilitate the scanning process. Stray marks were erased, and bubbling mistakes were corrected. After the extensive error and "bubble" checking process, the forms were scanned into five data files by class and MOS, one for each MOS. After scanning, the data files were reformatted and cleaned. Scanning problems with several cases and duplicate cases were investigated and resolved. To check the integrity of the data, three forms were drawn at random from each class for each MOS (165 forms total) and checked against the corresponding entries in the databases. For three classes

in each MOS, all the forms were checked against the database. This demonstrated that the random checking was sufficient to catch problems. After the post-scanning data checking was complete, the five MOS datasets were integrated into one overall SFQC database containing data for 1441 soldiers.

Variables Selected to Operationalize Study Constructs

The exact manifest indicators to operationalize the constructs specified by the research questions (see chapter 2) were not identifiable until after the data collection was completed. After the data were extracted from individual training folders and aggregated in a database, several criteria guided variable selection. Construct definitions, psychometric properties, the characteristics of "good" criteria presented in chapter 2, and potential sample size were utilized to select variables for inclusion in each question. However, as Lance and Vandenberg (2002) suggest, the paramount selection criteria was the degree to which the variables represented the a priori definition of the construct. Early in the process, the data collection was guided by a list of potential variables that could be used to operationalize the constructs (see Table 5). Most of these were found to be viable measures. Table 6 presents the manifest indicators selected by research question and construct. The data collected provided little choice in terms of indicators for the DO and KNOW constructs. As for BE, numerous data points were available. However, peer ratings and peer rankings provided the best data for BE in terms of construct definition, psychometric properties, and sample size. Table 7 displays the descriptive statistics for all manifest indicators used in the research models for the entire SFQC project database prior to eliminating incomplete cases. It should be reiterated that this study utilizes archival data and the researcher was not present for instrument development or initial data collection. Each type of measure is described next.[4]

Peer Ratings

Peer ratings have been shown to account for unique and significant variance in objective performance measures (Conway, Lombardo, & Sanders, 2001). In this study, a standardize set of peer ratings were collected at three points during SFQC—the end of Phase One, after the Isolation exercise in Phase Three, and after the Robin Sage FTX in Phase Three. Each member rated his fellow team members in six different categories, and the aggregate ratings from the team were recorded. The teams consisted of the same membership for Isolation and Robin Sage, unless there was attrition. The membership of the Phase One team was different. The rating form provides behavioral definitions for each rating construct and utilized a five-point scale from *poor* to *outstanding*. Each individual rated all his peers on a single, double-sided form. The six rating categories (with behavioral examples) were physical performance (*strength; endurance; coordination; ability to function with little sleep; operation under stress*), effort and persistence (*keeps going when things get tough; works hard even when cadre are not around; is always determined to succeed*), social interaction (*able to resolve conflicts and defuse difficult situations; can read people and social situations; interacts well with people of very different backgrounds; knows when to back off or be aggressive, when to be serious or funny*), teamwork (*contribution to the team effort; loyalty to the team; trustworthy*), leadership performance (*planning patrols, issuing OPORDS, Isolation planning, directing, controlling, supervising team members; focusing unit on task at hand; coordinating unit actions*), and tactical performance (*battle drills; land navigation; reconnaissance; combat tracking patrols; use of M60, AT4, Claymores; infiltration; air ops; linkup ops; instructional techniques*). As mentioned in the discussion of fit, ratings reflect the judgments of individuals in the situation as to the extent the ratee's

level of performance fits with the requirements of the situation or expectations of the culture. Schmitt and Chan (1998) provide a brief review of using ratings as performance criteria that suggests ratings have some issues but are value sources of performance information.

Tests

Tests are one of the most traditional measures of declarative knowledge (Kraiger et al., 1993) and are often recommended for evaluating training (e.g., Kirkpatrick, 1996). Four tests were included in the SFQC database—the land navigation exam, small unit tactics (SUT) exam, isolation exam, and comprehensive exam. Descriptive statistics for the tests are presented in Table 7. All tests were constructed by subject matter experts with operational experience in SF and conform to the standards established by the Training and Doctrine Command (TRADOC). Item type varied according to the test and included multiple choice, short answer and calculation types. The land navigation and SUT exams are descriptive in terms of their content and occur in Phase One. The land navigation test consisted primarily of items related to the knowledge and use of maps. The SUT exam covered knowledge of procedures and doctrine related to SUT including ambushing and patrolling. The isolation items related to mission planning and mission planning for the Robin Sage exercise. The comprehensive exam is a composite of items/tests that cover various topics related to the primary SF missions, including knowledge of procedures and doctrine related to air operations and unconventional warfare. Both the isolation and comprehensive exams take place in Phase Three. The number of items varies across tests. For example, the land navigation test has 20 items and isolation test has 50 items. For all four tests, a minimum score is required to continue training. Soldiers who fail to meet standard can be recycled to repeat the training. For soldiers who recycled, the score from their first attempt of each test

was utilized. Again, the researcher was not present for the data collection to ensure standardize across administrations. Since item-level responses were not maintained in the archive, no item-level data or analysis was available. Therefore, no measures of internal consistency reliability (e.g., Alpha) were available for the tests. Fortunately, because of using CFA, this lack of item data was not a problem. Finally, sample items and detailed descriptions were not provided to maintain test security as requested by the organization.

Peer Rankings

After the Isolation and Robin Sage exercises in Phase Three, each member of the team was asked to rank order all their peers on the team from best to worst in terms of overall performance. There are two separate rankings in the dataset, one for Isolation and one for Robin Sage. The number of people providing the rankings varies from team to team, so the number of team members providing ratings was captured as well. Again, since people in the same situation who share a similar organizational culture and training experience are making the judgments, this measure reflects the degree of fit with the requirements of the situation.

Peer Nominations

Peer nominations are used frequently in military organizations (e.g., Schwarzwald et al., 1999). Typically, members of the team or group are asked to nominate a certain number of their peers who exhibit either the best or worst of some type of performance or characteristic. Peer nominations in training have been found to predict later job performance outcomes (Schwarzwald et al., 1999). In this study, each team member was asked to select the two peers with whom he *would want to deploy* and the two with whom he *would not want to deploy*. In other words, each member has to choose two people he would want to “go to

war with" and two he would want to leave behind. Two sets of *would want* (positive) and *would not want* (negative) nominations were captured, one for Isolation and Robin Sage.

Spot Reports

Spot reports reflect observations of critical incidents of "positive" and "negative" behavior. The incidents captured by the SFQC cadre are typically related either to failures in task (skill) performance or in personal discipline (e.g., feel asleep at post). Positive and negative spots are recorded during Phase One and Phase Three, yielding four separate counts.

Times to Achieve Criterion

A field training exercise (FTX) is a simulation that provides trainees with the opportunity to demonstrate their proficiency in a realistic situation. The FTX provides the cadre with opportunities to evaluation the performance of the trainees in a realistic situation. For each FTX, there is a behaviorally defined level of proficiency required to pass the exercise. Meeting this standard (or not) is often recorded as only a GO or NO GO. Generally, after receiving a NO GO, a candidate receives feedback and is given another opportunity to retest. Some candidates are even recycled into a later section of the course in order to receive additional instruction and attempt the FTX again. Since all study participants are graduates and received a GO, this criterion is of limited value for our research. However, the number of iterations of the FTX required by a participant to meet standard is a useful measure of training performance. Times to criterion measures represent skill acquisition and proficiency—the number iterations of the exercise required to demonstrate mastery. This is consistent with the level of proficiency philosophy of SFQC. Two times to criterion measures were included in the dataset. The STAR FTX measures land navigation skill. The trainee must find a certain number of geographic points over rural terrain in a specified amount of

time while observing certain rules (e.g., a candidate automatically fails the exercise if he is found walking on a road, “Road Kill”). The small unit tactics (SUT) FTX is a team event, but each individual receives an evaluation. The SUT FTX covers various team-level combat activities like patrolling or ambushing. Both times to criterion measures are counts of the number of iterations required by the trainee to meet the standard of skill proficiency.

Research Models with Manifest Indicators by Question

With data collection and variable selection complete, this section presents the conceptual models to be tested for each research question.[5] The models show the specific manifest indicators used to operationalize the constructs for the analytic procedures presented in the next section. Due to the nature of the data, the exact operationalizations of the models were not known until the data was collected and fully described. Therefore, the fully operationalized models could not be presented earlier in the dissertation. The research models are introduced and described in order by question with the questions related to the latent structure of training performance discussed first and the questions related to modeling issues presented next.[1] Table 6 lists the manifest indicators for each question by construct.

Question One: What Is the Dimensionality of Training Performance?

Figure 3 represents the general, three-factor BE KNOW DO model of SFQC training performance, operationalized with the initial set of manifest indicators. The model presented resembles the Kraiger et al. (1993) model or the Wilson & Grant (1997) model from the job performance literature. Figure 4 depicts a model in which multiple method factors have been added to the BE KNOW DO model. Although this model is not a complete multimethod-multitrait model (i.e., all the performance constructs are not measured by all methods), this question presents an opportunity to model the variation associated with the different

measurement methods as well as that associated with the training performance factors. A third model—based on the McCloy et al. (1994) conceptualization of Campbell's direct determinants—is pictured in Figure 5. How does a Campbell inspired BE KNOW DO model differ from the initial model? McCloy and his associates (1994) suggest manifest measures are functions of multiple determinants—DK, PKS, and M. The initial BE KNOW DO model suggests each manifest measure is a function of exactly one of the three latent constructs—BE KNOW DO. The initial BE KNOW DO model in Figure 3 is arguing that certain measures are better to operationalize different latent constructs. Whereas, the work of McCloy and his associates recognizes each indicator as a function of one to three of the determinants based on its measurement method. All measures are posited to be a function of knowledge. However, others are a function of knowledge and skill, and some are a function of knowledge, skill, and motivation. The independent clusters philosophy (McDonald & Ho, 2002) utilized in the initial BE KNOW DO model relegates any variation attributed to the other latent factors to the uniqueness term. A fourth model—based on the idea that isomorphic content may contaminant the determinants—is presented in Figure 6. The model adds a general SF soldiering proficiency factor to the initial BE KNOW DO model. This content factor is similar to Campbell's (1999) non-job-specific task proficiency performance factor. All models will use the same manifest indicators to allow for comparisons between the alternative versions.

Question Two: Does the Latent Structure of Training Performance Change Over Time?

Figure 7 illustrates the BE KNOW DO model of Phase One SFQC training performance with the initial manifest indicators. Figure 8 illustrates the BE KNOW DO model of Phase Three SFQC training performance with the initial set of manifest measures.

Figure 9 illustrates a model depicting the posited relationship between Phase One and Phase Three performance constructs. It should be noted that the Phase One and Phase Three operationalizations of the DO factor differ. For Phase One, the DO factor is operationalized with land navigation and small unit tactics FTX performance measures as well as negative spot reports. For Phase Three, because of a derth of skill measures recorded in the folders, the tactical skill and leadership skill peer ratings and negative spot reports are used to operationalize the DO factor. The difference in operationalization may impact the relationship between the Phase One and Phase Three DO constructs. This situation creates the possibility for the uniqueness terms to be correlated because ratings and spot reports are not pure measures of DO and might share variance with other constructs and measures in the model.

Question Three. Does One General Factor or Several Specific Factors Describe BE?

Figure 10 displays the one-factor conceptualization of BE with the initial set of manifest indicators. Figure 11 shows the three-factor conceptualization of BE with the initial manifest indicators. The three content-specific BE factors are based on three of the Campbell et al. (1993) performance content factors—demonstrating effort, maintaining personal discipline, and facilitation of peer and team performance. The operationalization of personal discipline with negative spot reports and peer nominations leaves much to be desired. Although spot reports do record critical incidents related to personal discipline, they primarily record critical incidents related to task (or skill) failure. Therefore, the negative spot reports are effective indicators of personal discipline to extent that the task/skill failure was related to lapses in personal discipline. The same applies for peer nominations. These are

effective indicators of personal discipline to the extent that judgments are influenced by observations of failures in personal discipline.

Question Four: What is the impact of overfitting on cross-validation?

No separate model diagram is pictured for question four because the primary BE KNOW DO model from question one is used to investigate the impact of overfitting on cross-validation. Figure 3 presents the initial BE KNOW DO model. The next section discusses the procedures utilized to test these research models.

Analytic Procedure

This section presents the analytic procedures used to address the research questions presented in chapter 2 and test the models introduced in the previous section.[6] The section begins with a discussion of the general modeling conventions and strategy adopted for this study. The additional data processing required prior to addressing the research questions is presented, including the recoding and transformation of the data to facilitate modeling. The analytic procedures utilized to address each research question follow. The section concludes with a discussion of the evaluation of model fit, which specifies the fit criteria used for judgments in this study.

Modeling Conventions and Strategy

To address the research questions of this study, a confirmatory strategy—where the researcher proposes (a priori) and tests structural equation models of the theorized structure and relationships between constructs (Loehlin, 1998)—was adopted. A confirmatory approach requires the use of software capable of performing analytical techniques like confirmatory factor analysis (CFA) and structural equation modeling (SEM). Many programs are available. This study utilized CALIS included in SAS, version 8.2, to test all CFA and

SEM research models. Hatcher (1994) provides a description of CALIS and its capabilities as well as many instructional examples. Before discussing the modeling strategy, some specific conventions were adopted for all of the models in this study.

First, all CFA analyses were performed on correlation matrices and used maximum likelihood (ML) parameter estimation. According to MacCallum & Austin (2000), fitting the model to the correlation matrix is acceptable as long as the statistical software does not treat it as a covariance matrix. Using the correlation matrix is the default for CALIS (Hatcher, 1994). Additionally, MacCallum & Austin (2000) note using the correlation matrix eases interpretation. However, Hatcher (1994) notes that fitting the model to the correlation matrix is more likely to provide invalid standard errors for parameter estimates affecting the accuracy of parameter significance tests. To identify any potential problems, several of the models were fit to the appropriate covariance matrices, yielding parameter estimates identical to those obtained from fitting the models to the correlation matrices. Additionally, for the final models in the study, most of the parameter estimates have significance estimates much higher than the minimum significance level ($t = 1.96$)—therefore, changes in significance would not make a difference in interpretation. ML estimation is frequently used because it provides accurate parameter estimates when normality assumptions are violated.

Second, with the exception of three models, all models tested were independent clusters models in which no indicators loaded on more than one latent variable (McDonald & Ho, 2002). These clusters of items are also referred to as congeneric item sets (Millsap, 2002). For question one, the alternative models (see Figures 4, 5, and 6) have items associated with more than one latent variable. Third, most models tested were CFA measurement models for which all the latent variables were correlated, the factor variances

were constrained to equal one, and no predictive relationships were specified (Hatcher, 1994). However, for question one, although no predictive relationships were included, the alternative models did not posit correlating all of the latent variables. For example, the Campbell version of BE KNOW DO (see Figure 5) hypothesized uncorrelated factors, and the BE KNOW DO model with method factors (see Figure 4) was proposed with performance factors and method factors not correlating across their respective categories (i.e., performance factors did not correlate with method factors). Question two, exploring the dimensionality of phase-specific training performance and its relationship across time, tested measurement models for each of the phases and predictive (structural) models when the models from Phase One and Phase Three were integrated. All other models met the definition of measurement models. Now that the general modeling conventions and specific instances of their deviation have been mentioned, the basic modeling strategy designed to eliminate confirmation bias (MacCallum & Austin, 2000; Black, 2001; Millsap, 2002) is presented.

One of the problems with confirmatory techniques, like CFA, is *confirmation bias*, which refers to viewing a preferred model as the "correct" one while failing to utilize methodologies that might disconfirm this view. For example, failing to test theoretically justifiable alternative models might lead to confirmation bias. Even if the model confirms and the fit is acceptable, the model must be viewed as one of many different possibilities, not the only one (MacCallum & Austin, 2000; Black, 2001; Millsap, 2002). The following three methodologies were used in combination to address the research questions as appropriate: (a) the comparison of theoretically justified alternative models (Loehlin, 1998; MacCallum & Austin, 2000; Black, 2001); (b) the cross-validation of models on additional samples (Loehlin, 1998); (c) the use of nested models and the chi square difference test (Hatcher,

1994; MacCallum & Austin, 2000).

According to MacCallum and Austin (2000), an appropriate strategy for model specification is the model comparison methodology. The researcher specifies a number of a priori models and fits each on the same data. Each model is evaluated and compared after it has been fit to the data. These models may be based on competing theories or conflicting research findings. As Black (2001) points out, testing multiple models offers much stronger evidence than just testing one model. As reiterated by other authors (e.g., MacCallum & Austin, 2000), the model comparison strategy makes a stronger case for the researcher's preferred model if it provides a better fit to the data than the alternative models tested. Otherwise, if only one model is tested, no conclusion can be drawn about the model other than how well it fits the data. There could be many other models that fit the data better. To provide for the most meaningful comparison, all the alternative models should utilize the same set of indicators.

Another strategy for model specification is to cross-validate a model on a number of samples (MacCallum, 1995; Loehlin, 1998). If confirmed for every sample, the confidence in the model to explain the phenomena is greatly increased. This is similar to the cross-validation of multiple regression equations across different samples. Some change in model fit (i.e., similar to shrinkage in selection validation studies with multiple regression) from sample to sample is expected. Therefore, if the model fits multiple samples well, a very stronger case is made for the model. If you do not have several samples for cross-validation, it is permissible to split the sample into two or more samples as long as you have enough participants and do it randomly. It is extremely important to cross-validate if the researcher engages in model modification (MacCallum & Austin, 2000; Black, 2001; Millsap, 2002). In

practice and publication, idiosyncratic modifications to improve fit are the most problematic when the researcher does not cross-validate the modified model because the idiosyncratic modifications cannot be validated (or disconfirmed) through comparison.

Nested models occur when you are testing two or more models and one of the models fits completely within the other (Hatcher, 1994). Since one model may be embedded in the other, you can perform the chi-square difference test to determine which model has the better fit. Colquitt and associates (2000) used the chi-square difference test because their fully mediated model was embedded within their partially mediated model. When the degrees of freedom difference between the two models is one or greater, the value of the chi-square difference can be evaluated for significance, determining if one model provides a significantly better fit than the other. When models are compared with equal degrees of freedom (e.g., the same model compared across different samples), there is no test of significance. In this study, these three methodologies were utilized in various combinations to increase the confidence in the findings.

As an a priori philosophical decision, modifications could not include correlating error terms or creating cross-loading indicators not already specified by the model (i.e., some alternative models specified them a priori). This maintains the independent cluster or congeneric item set philosophy specified earlier (McDonald & Ho, 2002; Millsap, 2002). Therefore, the only modifications available were to drop non-significant paths in the model or add new indicator variables, which is not a recommended option because it alters the definition of the construct. Although correlating error terms was not permitted, the correlation of disturbance terms in SEM was considered an acceptable modification (see Figure 9.2, Millsap, 2002, for an example). Millsap (2002) cautions that adding “nonzero

covariances" between disturbance terms should only be considered when theoretically justified and when the identification status remains unchanged. Disturbance terms account for causal inferences on the latent variables not accounted for by the structural model. An unspecified construct, like method or situational variance, would be an example.

Finally, this study adopted a practical post hoc philosophy. If the results of a proposed research question were found to be in question and a practical analysis to address this question was available, then the analysis was performed to the extent that time and resources allowed.

Additional Processing of the Data

Prior to testing the research models, the data needed some additional preparation. Some measures need to be recoded and/or transformed. Additionally, because each question was operationalized with different manifest indicators (see Table 6), a data subset had to be created for each research question.

Recoding and Transforming the Data

Based on the descriptive statistics (see Table 7), recoding and/or transforming some variables were deemed appropriate. Table 8 reports the variables recoded and/or transformed with the specific procedures utilized and post-procedure descriptive statistics. In general, the recoding took the form of reversing the data so increasing values indicate higher levels of training performance. When each variable describes performance in the same direction, negative parameter estimates due solely to the difference in the direction of measurement should be eliminated. Therefore, negative parameter estimates can be more accurately attributed to either a true inverse relationship between the variables or a problem with the indicator, the model, or parameter estimation. Several variables (e.g., spot reports) are counts

of critical incidents. Distributions of count data frequently have problems with skewness (Howell, 1992). Howell (1992) recommends several forms of the square root transformation for counts to decrease skewness and to normalize/stabilize variance. This study employed the square root of X + 1 for all counts. Additionally, the peer rankings had to be converted to proportions to be meaningful. The peer rankings are aggregates of the rankings provided by each team member during Robin Sage or Isolation. Since the size of teams is variable, a proportion of an individual's ranking to the number of team members being ranked must be used instead of the aggregate ranking number. Howell (1992) suggests using the arcsine transformation for proportional data. The rankings in this study were transformed accordingly.

Subsets for Each Question

Each research question required a different operationalization of the constructs to adequately address its objectives. Only questions one and four used the same subset of data. Creating one subset of the SFQC database that included all variables utilized in the four questions would have reduced the useable data for all questions to that of the question with the least number of participants with complete and usable data. Again, the archival nature of the data dictated what was available. It was decided to draw subsets of data by question to optimize the sample size for each question. The variables required for each question are presented in Table 6. The question one subset contained complete data on 822 participants. This subset was used for question four as well. Complete data for question two was available for 558 participants. The subset utilized for question three contained 685 complete cases. Tables B1 through B11 in Appendix B contain descriptive statistics and zero-order correlations for all subsets and samples utilized to address the four research questions.

Analytic Procedure by Research Question.[1, 7, 8]

The basic analytic procedure utilized for this study involved fitting an a priori research model to an initial sample and making justified modifications until a final model was accepted. Then, the final model was cross-validated on the other samples. Elaborations of and/or deviations from this basic procedure are described for each question.

Question One

For question one, four completing models were tested: (a) BE KNOW DO (Figure 3); (b) BE KNOW DO with method factors (Figure 4); (c) the Campbell version of BE KNOW DO (Figure 5); and (d) BE KNOW DO with a general SF soldiering factor (Figure 6). It should be noted that the models are referenced by their descriptions, by their figure designation (e.g., the basic BE KNOW DO model is Figure 3) or by their letter from the preceding sentence (e.g., the basic BE KNOW DO model is model A). Four samples (n = 205, 205, 205, and 207) were drawn from the question one data subset (N = 822) in order to cross-validate the models. Cross-validation on multiple samples is highly desirable to demonstrate that results are not idiosyncratic to any one sample (Black, 2001; Lance & Vandenberg, 2002; MacCallum, 1995). Table 9 displays the demographic composition of the question one samples in comparison to the question one subset. Table B1 presents the descriptive statistics and zero-order correlations for the question one subset. Tables B2 through B5 present the descriptive statistics and zero-order correlations for samples 1 through 4.

The initial BE KNOW DO model, as presented in Figure 3, was fit to sample 1. The initial model was evaluated, and modifications were made as appropriate. Then, the new model was tested. Once modifications were completed and accepted, the final model was

cross-validated on the other three samples. The modifications conformed to the a priori model modification philosophy. As suggested in the literature (e.g., MacCallum & Austin, 2000), alternative models were tested to limit the possibility of confirmation bias. For each alternative version of the BE KNOW DO model, the initial model as presented in the appropriate figure (Figures 4 through 6) was fit to sample 1. Each initial alternative model utilized the same indicators as the initial BE KNOW DO model (see Table 6 or Figure 3). Each final alternative model was constrained to use the same manifest indicators as the final BE KNOW DO model (Figure 12). For each alternative model, the final version was fit to all four samples. This provided the most standardized comparison between the multiple versions of the BE KNOW DO model. This arrangement allowed for comprehensive comparisons of each model across all four samples or of the four models as fit to each sample.

Question Two

For question two, two samples (n = 279) were drawn from the question two data subset (N = 558). Table 10 presents the demographics for the question two samples in comparison to the overall subset. Table B6 presents the descriptive statistics and zero-order correlations for the overall subset. Tables B7 and B8 present the descriptive statistics and zero-order correlations for samples 1 and 2. A technique similar to question one was employed. Basically, the BE KNOW DO model was applied to data from Phase One (Figure 7) and Phase Three (Figure 8). A model showing the relationship between Phase One BE KNOW DO constructs and Phase Three BE KNOW DO constructs (Figure 9) was tested as well. The initial model for Phase One (Figure 7) was fit to both samples. The model was modified based on the a priori model modification philosophy. The modified model was fit to both samples. The same procedure was followed for the Phase Three (Figure 8) model. SEM

was used to test the predictive model (Figure 9). The initial predictive model (Figure 9) was fit to both samples. Modifications based on the Wald test (Hatcher, 1994) were fit to both samples. The modification process continued for an additional iteration.

Question Three

For question three, two samples (n = 340 and 345) were drawn from the data subset. Table 11 displays the demographic composition of question three samples as compared to the overall subset. Tables B9 through B11 present the descriptive statistics and zero-order correlations for the question three dataset and samples. Two versions of the BE factor are fit to the data—a one-construct version (Figure 10) and a three-construct version (Figure 11). The initial one-factor BE and initial three-factor BE models were fit to both samples. Then, both models were modified and fit to both samples, maintaining the same set of indicators across both models. For comparison purposes, it was extremely important to maintain the same set of manifest indicators for both the one- and three-factor versions of BE. Additionally, the need to operationalize all three of the BE constructs for modeling limited the possible modifications.

Question Four

For question four, the original BE KNOW DO model (Figure 3) was fit to the samples with a different purpose. It should be noted that the procedure utilized for this question was changed due to logistical constraints.[8] The procedure as originally proposed was very elaborate and was found to be untenable. The ultimate intent of question four was to determine if idiosyncratic modifications made to improve model fit would cross-validate to other samples. The new procedure accomplished that intent.

Several assumptions guiding the procedure for question four need to be stated. First, these idiosyncratic modifications are characterized as overfitting. The researcher capitalizes on some idiosyncratic characteristic of the sample to improve the results (fit indices and parameter estimates) to make the research more profound or publication worthy. The researcher probably will not decrease model fit or undermine the basic model she has proposed. Therefore, modifications that decrease fit or undermine the model should not be accepted. Second, there are several different modification indices (e.g., LaGrange Multipliers PHI; Hatcher, 1994) that can be used. To make the process more manageable, rules for using the modification indices need to be developed to guide the model modification process. Third, criteria for when to stop modifying and to accept the model need to be specified. Finally, modifications considered problematic are the ones that have little or no theoretical or empirical justification (MacCallum & Austin, 2000; Black, 2001). Therefore, the modification recommendations that meet the criteria should be followed regardless of justification in order to best approximate overfitting.

The initial BE KNOW DO model presented in Figure 3 and the four samples from question one were utilized for this question as well. Table B2 through B5 in Appendix B present the descriptive statistics and zero-order correlations for the four samples. The following procedure was adopted post hoc to adjust the research question to reflect the practical nature of the problem and to allow the issue to be addressed in a parsimonious manner. First, the model in Figure 3 was fit to each sample starting with sample 1. Second, the model was modified and fit to the initial sample iteratively until the criteria for stopping the modification process was met. Third, the accepted modified model was fit to the other three samples. For example, when sample 2 was the initial sample, the accepted modified

model was fit to samples 1, 3, and 4. Fourth, the process was repeated until all four samples rotated through and had been the initial sample being modified. This allowed the fit indices for the four over fit models to be compared across samples. This procedure is somewhat similar to one used by MacCallum (1986; as cited in Loehlin, 1998) with simulated data.

The modifications did not violate the following principles and rules. First, the basic structure of the model was maintained. No factors were deleted or added. No indicators important to construct definition were deleted. Second, the model was maintained as an independent cluster model in which no indicators loaded on more than one latent variable (McDonald & Ho, 2002). Since items were not allowed to load on more than one factor, the LaGrange Multiplier GAMMA (Hatcher, 1994) was ignored as a modification index. Third, all model modifications accepted maintained valid parameters estimates (.00 to 1.00). Again, no researcher would over-modify a model to create parameter estimates that were not publishable. Fourth, the modifications yielded significant and non-trivial parameter estimates for all parameters that had significant and non-trivial parameter estimates. In other words, no changes were made that undermined the initial state of affairs. Hatcher (1994) indicates the importance of significant and non-trivial parameter estimates. Fifth, two indices were selected to guide the modifications for this question—the Wald (path deletions; Hatcher, 1994) and the LaGrange Multipliers PHI (correlating error terms; Hatcher, 1994). Sixth, only the top two LaGrange suggestions (that did not including latent variables) were made per iteration. All Wald suggestions that did not violate any of the aforementioned rules were made. The choice to limit modifications was made for parsimony. An initial test of the procedure discovered that making too many modifications was chaotic, resulting in many suggested additions being suggested for deletion in the next iteration. Finally, the criterion

for stopping the modification process was defined as two successive iterations of modifications that failed to produce increased fit or that violated the above rules. The last model that increased fit and did not violate the rules was selected for cross-validation.

Evaluating Model Fit

Because this research utilizes CFA and SEM techniques, includes some model modification, and compares alternative models, a discussion of the criteria selected to evaluate model fit is presented. Discussing criteria for model fit is important because the degree of fit impacted decisions about modifying and accepting models during the analyses. The discussion ends with a statement of evaluation criteria utilized in this study. In terms of evaluating and comparing the models, Hatcher (1994) suggests the following criteria: (a) the *p* value for the model chi-square test should be non-significant; (b) the chi-square/degrees of freedom ratio should be less than 2; (c) the *t* values for each of the factor loading and path coefficients should exceed 1.96 and standardize factor loadings should be nontrivial in size; (d) R-square values should be relatively large; (e) distributions of normalized residuals should be symmetrical and centered on zero; and (f) the model should demonstrate high levels of parsimony and fit as indicated by the fit indices. Of course, when appropriate, the chi-square difference test should be used to determine whether the fit of one model is significantly better than another.

The selection of fit indices is an important decision. There are numerous indices (e.g., CFI, comparative fit index) available to judge the fit of a CFA or SEM model (Vandenberg & Lance, 2000; Lance & Vandenberg, 2002). The relevant literature recommends reporting the model chi-square statistic as well as a variety of other fit indices (MacCallum & Austin, 2000; Millsap, 2002; Lance & Vandenberg, 2002; Vandenberg & Lance, 2000). The chi-

square is often used as the primary fit index although it is greatly influenced by sample size (Hu & Bentler, 1999; Jackson, 2001; Lance & Vanderberg, 2002; Millsap, 2002). Therefore, with larger sample sizes, a statistically significant chi-square value does not necessarily indicate the model is a poor fit for the data (Millsap, 2002). To help standardize the chi-square value when comparing across models, the model's chi-square can be divided by its degrees of freedom, with lower values indicating better fitting models. Hatcher (1994) suggests the chi-square and degrees of freedom proportion should be less than 2 for a good fitting model. A variety of authors (e.g., Lance & Vandenberg, 2002) suggest using the following fit indices in conjunction with the chi-square: (a) the comparative fit index (CFI; Bentler, 1990); (b) the Tucker-Lewis index (TLI) also known as the non-normed fit index (NNFI; Vandenberg & Lance, 2000); (c) the standardized root mean square residual (SRMSR; Hu & Bentler, 1998, 1999); and (d) the root mean square error of approximation estimate with the accompanying 90% confidence interval (RMSEA; Browne & Cudeck, 1993; Vandenberg & Lance, 2000). According to Vandenberg and Lance (2000), the advantage of the NNFI over other fit indices comes from its immunity to the impact of sample size and its sensitivity to parsimony (penalizes for complexity). Hu & Bentler (1998, 1999) recommend the TLI (NNFI) for the additional reasons of sensitivity to model misspecification and relative insensitivity to violations of normality and to differences in estimation methods. The SRMSR is sensitive to misspecifications among the factor covariance in the model (Vandenberg & Lance, 2000). The RMSEA is sensitive to misspecifications in factor loadings (Vandenberg & Lance, 2000). Additionally, a 90% confidence interval helps in the interpretation of the RMSEA.

Models in this study were evaluated using the chi-square, the chi-square divided by its degrees of freedom, the CFI, the NNFI, the SRMSR, and the RMSEA. Additionally, although not used to evaluate model fit, the goodness-of-fit (GFI) and the normed fit (NFI; Bentler & Bonnett, 1980) indices are reported in tables for comparison. Both are strongly influenced by sample size. A non-significant chi-square value is desired as well as a chi-square/degrees of freedom ratio of less than 2. For the CFI, NNFI, and NFI, values above .90 are considered reasonable fit (or the lower boundary of good fit), and values above .95 are considered good or excellent fit (Hu & Bentler, 1999; Lance & Vandenberg, 2002; Vandenberg & Lance, 2000). For the SRMSR, Hu and Bentler (1999) suggest values below .08 represent good fitting models. Vandenberg and Lance (2000) suggest .10 as the upper limit for good fit for the SRMSR. For the RMSEA, Hu and Bentler (1999) suggest values below .06 represent good fit. Vandenberg and Lance (2000) offer .08 as the upper limit of good fit for the RMSEA. Millsap (2002) suggests SRMSR and RMSEA values less than .05 indicate good fit and values less than .08 indicate reasonable fit. However, Millsap (2002) acknowledges that these cutoffs are only rules of thumb and must be interpreted in the context of the analyses being conducted. Basically, for the SRMSR and RMSEA, smaller values represent better fit. For the CFI, NNFI, NFI, and GFI, values closer to 1.00 represent better fit.

CHAPTER FOUR: RESULTS

This section presents the results of the study. The results are reported by research question. Within each question, the findings are organized by model and/or sample (as appropriate), and post hoc analyses for that question, if any, are presented in an appendix. The results section is organized as follows: (a) question one (What is the dimensionality of training performance?), (b) question two (Does the latent structure of training performance change over time?), (c) question three (Does one general factor or several specific factors describe BE?) and (d) question four (What is the impact of overfitting on cross-validation?).[1]

Question One: What Is the Dimensionality of Training Performance?

Question one addresses the dimensionality of training performance for the entire Special Forces Qualifications Course (SFQC). Confirmation of a three-factor training performance model similar to the Kraiger et al. (1993) training evaluation model and other models from the job performance modeling literature (e.g., Wilson & Grant, 1997) is the goal of question one. In addition to the primary model, several other versions of the three-factor model inspired in part by the work of Campbell were tested for comparison. The initial conceptual models with the manifest indicators are presented in Figures 3 through 6. Table B1 presents the descriptive statistics and zero-order correlations for the question one subset. Tables B2 through B5 present the descriptive statistics and zero-order correlations for samples 1 through 4.

BE KNOW DO Model

Initial Model

Figure 3 presents the baseline BE KNOW DO model. This model represents the primary model of this study. The model was fit to sample 1 resulting in a highly significant

chi-square statistic, 76.70 (41), $p = .0006$. However, the chi-square/degrees of freedom ratio was 1.87. The other fit indices indicated the model was on the verge of providing an adequate fit of the data (CFI = .90; NNFI = .86; SRMSR = .06; RMSEA = .07 [.04 - .09]). The NNFI below .90 was the most troubling fit statistic. Table 12, section 1 presents the fit indices for the initial model fit to sample 1. Table 13 presents the standardized parameter estimates for the model fit to sample 1. All parameter estimates were significant with the exception of the loading for Phase 3 negative spot reports on the DO factor (.07; $p > .05$).

Final Model

The Wald test (Hatcher, 1994) identified the Phase 3 negative spot reports variable as an indicator that should be removed from the model. No other modifications were justified. Figure 12 presents the final version of the BE KNOW DO model. The final version with the specified deletion was fit to sample 1 resulting in a slightly less significant chi-square, 60.16 (32), $p = .0019$, with a chi-square/degrees of freedom ratio of 1.88. The difference in the final model chi-square and the initial model chi-square was on the verge of being significant, 16.54 (9), $p > .05$ ($p = .05$ at 16.92). The other fit indices demonstrated a slight improvement over the initial model (CFI = .92; NNFI = .88; SRMSR = .05; RMSEA = .07 [.04 - .09]). Taken together, the fit indices with the exception of the significant chi-square and the NNFI suggest the model provides an adequate to good fit of the data. Table 12, section 1 presents the fit indices for the final model fit to sample 1. All parameter estimates in the final model were significant at $p < .05$. As can be seen in Table 13, the parameter estimates for the initial and final models for sample 1 differ very little. For both versions of the model, the KNOW and DO constructs had the largest correlation (.92, .91 respectively).

Cross-validation. The final model was fit to each of the other three samples. Table 12, section 1 presents the fit indices for all samples. Although diagrams with parameter estimates are not provided, Figure 12 can be used to visual the model with the standardized parameter estimates provided in Table 13. Cross-validation on sample 2 resulted in a chi-square statistic that approaches non-significance, 47.73 (32), p = .036, with a chi-square/degrees of freedom ratio of 1.49. The other fit indices demonstrated the final model provided a good fit to the sample 2 data (CFI = .96; NNFI = .94; SRMSR = .05; RMSEA = .05 [.01 - .08]). Fitting the final model on sample 3 resulted in a more significant chi-square statistic, 64.65 (32), p = .0006, with a chi-square/degrees of freedom ratio of 2.02. The other fit indices demonstrated the final model provided an adequate fit to the sample 3 data (CFI = .93; NNFI = .90; SRMSR = .06; RMSEA = .07 [.05 - .10]). Finally, the model was tested on sample 4 resulting in a less significant chi-square statistic (than for samples 1 and 3), 56.07 (32), p = .005. The chi-square/degrees of freedom ratio was 1.75. The other fit indices indicated the model provided an acceptable to good fit of the sample 4 data (CFI = .93; NNFI = .90; SRMSR = .06; RMSEA = .06 [.03 - .09]). Although the chi-square was significant for all samples, the fit indices across the four samples suggested that the final version of the BE KNOW DO model provided an adequate to good fit for the data. Importantly, only the NNFI for sample 1 was below the threshold of adequate fit (NNFI = .90). All other fit indices achieved the minimum levels of acceptable fit or better.

BE KNOW DO with Method Factors

Initial Model

Figure 4 presents the basic BE KNOW DO model with correlated method factors. In addition to the three training performance factors, this model adds a factor for each

measurement method (e.g., ratings) represented by the model's manifest indicators. The initial model was fit to sample 1, and the model would not converge under any circumstances (i.e., increasing iterations, increasing function calls, and changing the convergence technique). The model with correlated method factors did not converge for any of the other samples as well. Then, the method factors were modeled as being uncorrelated (see Figure 13). Initially, the uncorrelated version did not converge. The convergence properties were altered to facilitate convergence, that is the number of both function calls and iterations were increased to 1000 (50 is the default for both) and the convergence technique was changed as well. The initial model with uncorrelated method factors was fit to the sample 1 resulting in a highly significant chi-square, 61.79 (30), p = .0006, with a chi-square/degrees of freedom ratio of 2.06. The other fit indices suggested a mixed interpretation of whether the model provided an adequate fit of the sample 1 data (CFI = .91; NNFI = .83; SRMSR = .05; RMSEA = .07 [.05-.10]). Table 12, section 2 presents the fit indices for the initial model tested on sample 1. When fit to sample 1, this model generated a number of warning messages, including "the covariance matrix for the estimates is not full rank" and "some parameter estimates are linearly related to other parameter estimates", which most likely indicates a problem with the model. Table 14 presents the standardized parameter estimates for the initial model fit to sample 1. By using Table 14 and Figure 13, the model can be visualized with standardized parameter estimates. Many of the parameter estimates were found to be non-significant and trivial in size. Hatcher (1994) indicates that parameter estimates should be significant at $p < .05$ and should be non-trivial in size. Otherwise, the model is not a good fit for the data and/or may have problems. Additionally, several of the parameter estimates were inexplicably negative. Given the warning messages and the

parameter estimates, the model was probably not identified, and the results of the initial model should be viewed with suspicion.

Final Model

Figure 14 presents a conceptual diagram of the final model. The methods version of the final BE KNOW DO model was fit to sample 1 resulting in a less significant chi-square statistic, 45.42 (23), $p < .0035$, with a chi-square/degrees of freedom ratio of 1.98. The other fit indices suggested the model provided an adequate fit of the sample 1 data (CFI = .95; NNFI = .87; SRMSR = .05; RMSEA = .07 [.04-.10]), although the NNFI is slightly below the threshold for adequate fit. Table 12, section 2 presents the fit indices for the final model tested on sample 1. Table 14 presents the standardized parameter estimates for the final model fit to sample 1. Again, this model generated the same warning messages and suspicious parameter estimates.

Cross-validation. The methods version of the final BE KNOW DO model was fit to the additional three samples. For each sample, the convergence properties were altered in the same way as sample 1 to facilitate convergence. For all three samples, running the model generated the same warning messages and suspicious parameter estimates. Table 12, section 2 presents the fit indices for the final model from each sample. Table 14 displays the standardized parameter estimates for the cross-validation samples. Although the fit indices look good for all three cross-validation samples, it would be pointless to interpret the results of the model. Given the findings, the methods version of the model was most likely not identified. Although the model was overidentified according to the counting rule (Hatch, 1994; Kaplan, 2000), the condition of having more informants than parameters to be estimated is necessary but not sufficient for declaring the model identified. Hatch (1994)

suggests using different start values for parameter estimation to test for identification. If the parameter estimates vary for different start values, the model is not identified. However, yielding consistent parameter estimates with different start values still does not guarantee identification. The start values were varied, and the BE KNOW DO model with uncorrelated method factors was fit to the four samples. The parameter estimates for the indicator loadings on the method factors and the correlations between the latent variables were found to change with the start values, confirming the model was probably not identified.

Campbell's BE KNOW DO

From the job performance modeling literature, the Campbell (1999) model posits that all performance content factors are a combination of three direct determinants—DK, PKS, and M. This version of the BE KNOW DO model was designed to test Campbell's ideas about the direct determinants. Figure 5 shows the conceptual diagram of the Campbell version. For all samples, the convergence criteria had to be altered (iterations set to 1000, function calls set to 1000, and technique changed to DD) in order for the model to converge. Table 12, section 3 presents the fit indices related to the Campbell version, and Table 15 displays the standardized parameter estimates for all four samples.

Initial Model

Figure 5 displays the Campbell version with the initial set of indicators. This model was fit to sample 1 data and generated warning messages related to the covariance matrix for the estimates not being full rank. Varying the start values of the parameter estimates for the initial model fit to sample 1 resulted in different parameter estimates for the BE and DO factors. Several of the parameter estimates were out of bounds (above 1.00) as well.

Although the fit indices were very good, the evidence suggested the model was not identified. Therefore, the fit indices and parameter estimates should be disregarded.

Final Model

Figure 15 presents the Campbell version with the final set of manifest indicators. When the final version was fit to sample 1, the same warning messages suggesting the model was not identified were generated. The start values were varied and resulted in the parameter estimates for BE and DO varying with the start values. Several of the parameter estimates were out of bounds as well. Again, although the fit indices were very good, the evidence suggested the model was not identified and, therefore, the fit indices and parameter estimates should be disregarded.

Cross-validation. At this point, the Campbell version with the final set of indicators (Figure 15) was fit to the remaining three samples. Surprisingly, the final model failed to generate warning messages for samples 2 and 4. Sample 3 had the same problems as sample 1. The start values were altered for all three samples. The parameter estimates varied for sample 3 but not for samples 2 and 4. It was concluded that the final model was not identified for sample 3. The parameter estimates for samples 2 and 4 (see Table 15) were examined. For sample 2, the parameter estimates included many that were non-significant and trivial as well as one that was out-of-bounds. Since the counting rule and the rank condition are necessary but not sufficient for identification, the parameter estimates suggest that the final model was not identified for sample 2. No problems were found to suggest that the final model was not identified for sample 4. As can be seen in Table 12, section 3, fitting the final model to sample 4 resulted in a non-significant chi-square statistic, 31.64 (24), $p = .14$, with a chi-square/degrees of freedom ratio of 1.32. The other fit indices confirmed that the model

provided a good fit for the sample 4 data (CFI = .98; NNFI = .96; SRMSR = .05; RMSEA = .04 [.00-.07]). However, the Campbell version of the BE KNOW DO model was most likely not identified.

BE KNOW DO with a General Soldiering Factor

This version of the BE KNOW DO model was inspired by the non-job-specific task proficiency performance content factor in the Campbell model (Campbell, 1999; Campbell et al., 1993). Figure 6 displays the conceptual diagram with the initial manifest indicators for the model. In this case, the non-job-specific task proficiency is general soldiering task proficiency (e.g., land navigation). When the initial model was fit to sample 1 and the final model was fit to all the samples, none of them converged without altering the convergence properties to some degree (e.g., samples 3 and 4 only required increased iterations and function calls). Table 12, section 4 presents the fit indices for this version of the BE KNOW DO model for all four samples. Table 16 displays the standardized parameter estimates. Figure 6 displays the model with the initial set of manifest indicators, and Figure 16 displays the model with the final set of manifest indicators.

Initial Model

After adjusting the convergence properties, the initial model was fit to sample 1 resulting in a significant chi-square statistic, 46.93 (30), p = .03, with a chi-square/degrees of freedom ratio of 1.56. The other fit indices confirmed the model provided a good fit for the data (CFI = .95; NNFI = .91; SRMSR = .04; RMSEA = .05 [.00-.06]). However, many of the parameter estimates were non-significant and trivial in size (see Table 16). Additionally, many of the parameter estimates were inexplicably negative. Given these facts, the fit indices

and parameter estimates must be regarded with suspicion because the model was probably not identified.

Final Model

The final model was fit to sample 1 resulting in a significant chi-square statistic, 35.01 (22), $p = .04$, with a chi-square/degrees of freedom ratio of 1.59. The other fit indices suggested a good fit (CFI = .96; NNFI = .91; SRMSR = .04; RMSEA = .05 [.01-.09]). However, many of the standardized parameter estimates were non-significant and trivial in size, and the correlation between KNOW and DO (-1.17) was out of bounds (> 1.00). Although the model met the counting rule and start value test and did not generate warning messages, the model was probably not identified for sample 1. The fit indices and standardized parameter estimates should be viewed with suspicion.

Cross-validation. Fitting the final model to sample 2 resulted in a series of warning messages concerning violations of the rank condition. Several of the standardized parameter estimates were found to be out of bounds (> 1.00). For example, the correlation between BE and KNOW was -1265. The final model was not identified for sample 2, and the fit indices and parameter estimates should be disregarded. The final model was fit to sample 3 and yielded a significant chi-square statistic, 35.46 (22), $p = .04$, with a chi-square/degrees of freedom ratio of 1.61. The other fit indices suggested the model was a good fit (CFI = .97; NNFI = .94; SRMSR = .04; RMSEA = .06 [.02-.09]). However, the standardized parameter estimates suggested a problem. For example, the correlation between KNOW and DO (-1.12) was out of bounds (> 1.00). Although the model passed the counting rule and the start value test and did not generate any warning messages, the final model was probably not identified for sample 3, and the fit indices and parameter estimates should be viewed with suspicion.

The final model was fit to sample 4 and resulted in a non-significant chi-square, 27.66 (22), *p* = .19, with a chi-square/degrees of freedom ratio of 1.26. The other fit indices confirmed a good fitting model (CFI = .98; NNFI = .97; SRMSR = .04; RMSEA = .04 [.00-.07]). Table 12, section 4 presents the fit indices for sample 4, and Table 16 displays the standardized parameter estimates for sample 4. Several parameter estimates were non-significant and trivial in size. The fit indices and parameter estimates for sample 4 should be viewed with suspicion. This concludes the proposed analyses for question one. A number of post hoc analyses relevant to the first research question were conducted to explore the BE KNOW DO model further (see Appendix C).

Question Two: Does the Latent Structure of Training Performance Change Over Time?

Although the main thrust of question two is the relationship of training performance across phases of SFQC, the BE KNOW DO model had to be confirmed for each phase of training before the relationship between the phases could be explored. Tables B6 through B8 present the descriptive statistics and zero-order correlations for the question two subset and samples.

Phase One Performance

Initial Model

Figure 7 displays the initial conceptual model of performance in Phase One of SFQC. The model is similar to the question one BE KNOW DO model except the manifest indicators were measured only during Phase One training. This phase of training relates primarily to land navigation (LN) and small unit tactics (SUT). Due to the limited number of declarative knowledge tests available, only two indicators were included in the KNOW factor—one for each of the two major content areas. The initial model was fit to sample one

and yielded a very significant chi-square statistic, 52.79 (17), $p < .0001$, with a chi-square/degrees of freedom ratio of 3.11. The other fit indices for the initial model on sample 1 indicated good fit (CFI = .97; NNFI = .94; SRMSR = .05; RMSEA = .09 [.06-.11]), although the RMSEA was slightly above the .08 for adequate fit. The initial model was fit to sample two resulting in a less significant chi-square statistic, 31.26 (17), $p = .0186$, with a chi-square/degrees of freedom ratio of 1.84. The other fit indices for the initial model on sample 2 indicated good fit (CFI = .99; NNFI = .98; SRMSR = .05; RMSEA = .06 [.02-.08]). Table 17, section 1 reports the fit indices for the initial model from both samples. Table 18 presents the standardized parameter estimates for the initial model from both samples. Although the parameter estimates for sample one were found to be significant, the parameter estimates for sample 2 were slightly problematic. The negative spot reports variable produced a non-significant, negative loading. This negative loading influenced the sign of the parameter estimates in the sample 2 DO construct and the correlations between the latent variables. Since the negative spot reports variable was considered problematic a priori, it was decided to remove the variable. The Wald test (Hatcher, 1994) for sample 2 agreed with this assessment.

Final Model

The final Phase One model dropped the negative spot reports variable. This model was fit to sample 1 producing a less significant chi-square than the initial model did on sample 1, 33.07 (11), $p = .0005$, with a smaller chi-square/degrees of freedom ratio of 1.84. The other fit indices for sample 1 improved as well (CFI = .98; NNFI = .96; SRMSR = .05; RMSEA = .09 [.05-.12]), although the RMSEA remained the same at slightly above the threshold of adequate fit. Since the RMSEA is sensitive to the misspecification of factor

loadings (Vandenberg & Lance, 2002), this result may have been related to an indicator that could have loaded on two factors or to the fact that two latent variables have only two indicators each, which is typically not recommended (Hatcher, 1994). The final model was fit to sample 2 resulting in a statistically non-significant chi-square value, 9.37 (11), $p = .5877$, with a chi-square/degrees of freedom ratio of .85. Both the chi-square and the chi-square/degrees of freedom ratio indicated a model that is an excellent fit for the data in sample 2. The other fit indices for the final model on sample 2 indicated excellent fit as well (CFI = 1.00; NNFI = 1.00; SRMSR = .03; RMSEA = .00 [.00-.06]), with the CFI, NNFI, and RMSEA indicating the best possible fit. The change in the chi-square from the initial model to the final was significant for both samples, 19.72 (6), $p < .005$, for sample 1, and 21.89 (6), $p < .005$, for sample 2. This indicates that dropping the negative spot reports variable significantly increased the model fit. Table 17, section 1 reports the fit indices for the final model from both samples. Table 18 presents the standardized parameter estimates for the final model from both samples. Figure 17 displays the final model for Phase One. All parameter estimates for the final model were significant at $p < .05$ with most being significant at $p < .01$.

Phase Three Performance

Initial Model

Figure 8 displays the initial conceptual model of performance in Phase Three of SFQC. The model is similar to the question one BE KNOW DO model except the manifest indicators were measured only during Phase Three training. Due to the limited number of declarative knowledge tests available, only two indicators were included in the KNOW factor—the Isolation exam (mission planning) and the comprehensive exam covering several

topics including UW and air operations. Finding indicators of skill (DO) from Phase Three was problematic because there were limited measures captured in a quantitative metric. The measures selected consisted of negative spot reports and peer ratings of tactical and leadership skills. As a result, the Phase Three DO construct did not represent the exact same construct as Phase One DO. Otherwise, the Phase Three constructs tested were similar to those from Phase One.

The initial conceptual model was fit to sample 1 and resulted in a highly significant chi-square statistic, 46.98 (17), $p < .0001$, with a chi-square/degrees of freedom ratio of 2.76. Although the chi-square and chi-square/degrees of freedom ratio suggested the model did not provide an adequate fit, the other fit indices for sample 1 suggested otherwise (CFI = .98; NNFI = .97; SRMSR = .05; RMSEA = .08 [.05-.11]). The initial conceptual model was fit to sample 2 as well resulting in a highly significant chi-square value, 52.09 (17), $p < .0001$, with a chi-square/degrees of freedom ratio of 3.06. The other fit indices suggested the model provided a good fit (CFI = .98; NNFI = .97; SRMSR = .05; RMSEA = .09 [.06-.11]), with the exception of the RMSEA being slightly above the threshold for an adequate fit. Table 17, section 2 presents the fit indices for the initial Phase Three model for both samples. Table 19 displays the standardize parameter estimates for the initial model fit to both samples.. The parameter estimates were fairly consistent for the BE and DO factors, with the exception of the negative spot reports variable. The factor loading for negative spot reports (.10) was found to be non-significant ($p > .05$) for sample 1. The parameter estimate (.25) was significant at $p < .05$ for sample 2. The pattern of the KNOW indicator loadings reversed from sample 1 to sample 2. The latent relationship between BE and KNOW for both samples

1 and 2 (.23, .16 respectively) was found to be non-significant. The strongest latent relationship was found between BE and DO (.83, .87).

Final Model

The Wald test (Hatcher, 1994) for sample 1 recommended dropping both non-significant paths—negative spot reports to DO and BE to KNOW. The final model dropped the negative spot reports variable. This model was fit to sample 1 resulting in a less significant chi-square value for sample 1, 27.43 (11), $p = .004$, with a smaller chi-square/degrees of freedom ratio of 2.49. The other fit indices for sample 1 improved as well and suggested a good fitting model (CFI = .99; NNFI = .98; SRMSR = .04; RMSEA = .07 [.04-.11]), although the RMSEA was in the adequate fit range instead of the good fit range. Again, the RMSEA may be impacted by the two latent variables with only two indicators. The final model was fit to sample 2 yielding a less significant chi-square, 35.29 (11), $p = .0002$, with a larger chi-square/degrees of freedom ratio of 3.21. The other fit indices improved slightly (CFI = .99; NNFI = .98; SRMSR = .04; RMSEA = .09 [.06-.11]), with the exception of the RMSEA that remained the same at slightly above the threshold of adequate fit. The change in the chi-square from the initial model to the final was significant for both samples, 19.55 (6), $p < .005$, for sample 1, and 16.80 (6), $p = .011$, for sample 2. This indicates that dropping the negative spot reports variable significantly increased the model fit for both samples. Table 17, section 2 reports the fit indices for the final model as well as for the initial model from both samples. Table 19 presents the standardized parameter estimates for the final and initial models from both samples. Figure 18 displays the Phase Three model with the final set of manifest indicators. All parameter estimates greater than .23 for the final model were significant at $p < .05$ with most being significant at $p < .01$.

Relating Performance Across Phases

The results from the CFA analyses indicated that the three-factor BE KNOW DO model did an adequate job of modeling performance in both phases. Therefore, the SEM model showing the predictive relationship between Phase One and Phase Three constructs could be tested. The results for this section are presented by model (e.g., initial model). Due to the complex nature of the models, presenting standardized parameter estimates in tabular form was not attempted. Instead, a diagram displays the parameter estimates for each model fit to each sample (see Figures 19 through 24). Additionally, although these are structural models, the loadings for the manifest indicators are presented along with the latent paths (i.e., typically the item loadings are only presented in measurement models).

Initial Model

Figure 9 displays the conceptual diagram with manifest indicators for the initial model tested. This model was constructed by adding the initial Phase One and Phase Three models together. Although negative spot reports were dropped as an indicator from both the Phase One and Phase Three models, the initial structural model between the two phases includes them as indicators. The initial model posited several paths between the latent variables in Phase One and those in Phase Three. Namely, each construct in Phase One was thought to predict the similar construct in Phase Three (e.g., Phase One DO will predict Phase Three DO). Additionally, Phase One KNOW (declarative knowledge) was thought to be predictive of Phase Three DO (skill).

The posited initial model was fit to sample 1 resulting in a highly significant chi-square statistic, 491.30 (97), $p < .0001$, with a chi-square/degrees of freedom ratio of 5.07. The other fit indices confirmed that the initial model provides poor fit for the sample 1 data

(CFI = .86; NNFI = .83; SRMSR = .15; RMSEA = .12 [.11-.13]). The model was then fit to sample 2 resulting in a highly significant chi-square value, 514.09 (97), $p < .0001$, with a chi-square/degrees of freedom ratio of 5.30. The other fit indices confirmed that the initial model provides poor fit for sample 2 as well (CFI = .86; NNFI = .83; SRMSR = .16; RMSEA = .12 [.11-.14]). Table 17, section 3 provides the fit indices for the initial model for both samples. For both samples, all standardized parameter estimates were significant at $p < .05$ with the exception of two for each sample. For both samples, the path between Phase One KNOW and Phase Three DO (.15, -.13) was found to be non-significant. All other latent paths were significant for both models. For sample 1, Phase Three negative spot reports were found not to contribute significantly to the DO construct. For sample 2, Phase One negative spot reports were found not to contribute significantly to the DO construct. The Wald test (Hatcher, 1994) for sample 1 suggested removing the path from KNOW (Ph1) to DO (Ph3) and the loading for negative spot reports (Ph3) from DO. The Wald test (Hatcher, 1994) for sample 2 suggested removing the path from KNOW (Ph1) to DO (Ph3) as well and the loading for negative spot reports (Ph1) from DO. Figures 19 and 20 display the standardized parameter estimates for the initial model from samples 1 and 2 respectively.

Intermediate Model

The intermediate structural model eliminated the path between Phase One KNOW and Phase Three DO as well as both sets of negative spot reports from the model. At this point, the intermediate model contained only the indicators found in the final versions of the Phase One and Phase Three models. This model was fit to sample 1 and provided a highly significant chi-square statistic, 428.86 (71), $p < .0001$, with a chi-square/degrees of freedom ratio of 6.04. The other fit indices confirmed that the intermediate model provides bad fit for

the sample 1 data (CFI = .87; NNFI = .84; SRMSR = .16; RMSEA = .13 [.12-.15]). The intermediate model was fit to sample 2 resulting in a highly significant chi-square statistic, 449.02 (71), $p < .0001$, with a chi-square/degrees of freedom ratio of 6.32. The other fit indices confirmed that the intermediate model provides bad fit for sample 2 as well (CFI = .87; NNFI = .84; SRMSR = .17; RMSEA = .14 [.13-.15]). However, for both samples, the difference between the initial model chi-square and the intermediate model chi-square (62.46, 65.07 respectively) was significant at $p < .0001$ for 26 degrees of freedom, suggesting that the intermediate model was an improvement over the initial model. Table 17, section 3 presents the fit indices for the intermediate model from both samples. Although not identical across the two samples, the relationships between the latent variables are very consistent. Figures 21 and 22 present the intermediate model with standardized parameter estimates from sample 1 and sample 2 respectively. All parameter estimates are significant at $p < .05$ with most being significant at $p < .01$. As the modification indices for both samples were examined, a pattern of suggestions from the Lagrange Multipliers (Hatcher, 1994) emerged involving correlating error terms between the indicators of Phase 3 BE and DO. This pattern made sense because both factors utilized peer ratings by the same raters as indicators. Although Phase One BE and Phase Three BE both utilized peer ratings of the same constructs as indicators, the raters and situations were different at each phase. From the LaGrange Multipliers, it was obvious the shared, unaccounted for variance between Phase Three BE and DO was dramatically impacting model fit. Although correlating the error terms (residuals) of manifest indicators is generally not recommended, correlating the disturbance terms of latent variables presents an acceptable option producing a result similar to adding a second-order factor to the model or to cross-loading the manifest indicators on another factor

that had been added to the model. Millsap (2002) indicates correlating disturbance terms is acceptable when theoretically justified.

Final Model

The final structural model added a path correlating the disturbance terms of Phase Three BE and DO. This model was fit to sample 1 and produced a much less significant chi-square statistic, 143.86 (70), $p < .0001$, with a chi-square/degrees of freedom ratio of 2.06. The other fit indices confirmed that the final model provided good fit for the sample 1 data (CFI = .97; NNFI = .97; SRMSR = .06; RMSEA = .06 [.05-.08]). All indices were in their acceptable or good fit ranges. The final model was fit to sample 2 and produced a less significant chi-square as well, 112.40 (70), $p = .001$, with a chi-square/degrees of freedom ratio of 1.61. The other fit indices confirmed that the final model provided good fit for sample 2 as well (CFI = .99; NNFI = .98; SRMSR = .05; RMSEA = .05 [.03-.06]). For both samples, the difference between the initial model chi-square and the final model chi-square (347.44, 401.69 respectively) was highly significant at $p < .0001$ for 27 degrees of freedom, suggesting that the final model was an improvement over the initial model. Additionally, for both samples, the difference between the intermediate model chi-square and the final model chi-square (284.98, 336.62 respectively) was highly significant at $p < .0001$ for a single degree of freedom, indicating that the final model was also an improvement over the intermediate model. Table 17, section 3 presents the fit indices for the final model from both samples. Based on the fit indices, the final model was found to provide a good to excellent fit to the data and was accepted.

Figures 23 and 24 present the standardized parameter estimates for the final model as fit to samples 1 and 2 respectively. All parameter estimates for both models were found to be

significant at $p < .05$. First, adding the path between the disturbance terms did not change the latent variable relationships. The magnitude of several of the parameter estimates between latent variables either increased or decreased slightly but the pattern of the latent relationships remained the same. A number of post hoc analyses relevant to the second research question were conducted (see Appendix D).

Question Three: Does one general factor or several specific factors describe BE?

Question three addresses the content specificity issue related to modeling performance. How is the dimensionality of a single performance construct best modeled? Two versions of the BE factor were fit to the data—a one-construct version and a three-construct version. Both versions used the same manifest variables. The results are presented for the initial set of manifest indicators. Then, the results are presented for a second, reduced set of manifest indicators. Within each section, the results are presented by version of the BE model (one- and three-construct models). Finally, based on the results of the proposed analyses, several post hoc models were tested, and their results are presented in Appendix E. Tables B9 through B11 present the descriptive statistics and zero-order correlations for the question three dataset and samples.

Initial Models

One-Factor BE

Figure 10 presents the one-factor conceptualization of the BE factor with manifest indicators. This model was fit to both samples and resulted in very poor fit. The model yielded a very significant chi-square statistic both for sample 1, 2293.41 (90), $p < .0001$, and for sample 2, 2133.35 (90), $p < .0001$. The CFI and NNFI were well below the threshold for acceptable fit for both samples. The SRMSR and RMSEA were well above the threshold for acceptable fit for both samples. Table 20 presents the fit indices for the initial one-factor BE model from both samples. Table 21 presents the standardized parameter estimates from both samples.

Three-Factor BE

Figure 11 presents the three-factor conceptualization of the BE factor with manifest indicators. This model was fit to both samples and resulted in very poor fit as well. The model yielded a very significant chi-square statistic both for sample 1, 2266.40 (87), $p < .0001$, and for sample 2, 2109.63 (87), $p < .0001$. The CFI and NNFI were well below the threshold for acceptable fit for both samples. The SRMSR and RMSEA were well above the threshold for acceptable fit for both samples. Table 20 presents the fit indices for the initial three-factor BE model from both samples. Table 21 presents the standardized parameter estimates from both samples. The correlation between the team and peer facilitation and effort constructs was found to be over 1.00 (or out-of-bounds) for both samples.

Second Models

Based on the results of the initial model CFAs, a second set of models was created by eliminating manifest variables from the initial set of indicators. The indicators were deleted

in an attempt to eliminate the influence of unspecified factors that the modification indices suggested might be operating in the data. These unspecified factors were most likely consistent with the phase, content, time, and/or situation in which the measurement took place (i.e., Phase One, Isolation, and Robin Sage). Four of the initial items were dropped. A balance had to be maintained between deleting indicators and maintaining enough indicators to operationalize the three-construct version. The operationalization would be jeopardized if more indicators had been dropped. Figures 25 and 26 present the second versions of the one- and three-factor models, respectively.

One-Factor BE

The second one-factor model was fit to both samples and resulted in improved fit over the initial model. However, the model produced a very significant chi-square statistic both for sample 1, 906.69 (44), $p < .0001$, and for sample 2, 979.39 (44), $p < .0001$. Interestingly, only the chi-square improved substantially. The change in the chi-square statistics between the initial and the second model for both samples, 1386.72 (46) for sample 1 and 1153.96 (46) for sample 2, was significant at $p < .0001$. The other fit indices exhibited little change over the initial version. Table 20 presents the fit indices for the second one-factor BE model from both samples. Table 21 presents the standardized parameter estimates from both samples.

Three-Factor BE

The second three-factor model was fit to both samples and resulted in improved fit over the initial three-factor model. However, the model produced a very significant chi-square statistic both for sample 1, 841.53 (41), $p < .0001$, and for sample 2, 914.54 (41), $p < .0001$. Interestingly, only the chi-square improved substantially. The change in the chi-square

statistics between the initial and the second model for both samples, 1424.87 (46) for sample 1 and 1195.09 (46) for sample 2, was significant at $p < .0001$. The other fit indices exhibited little change over the initial version. Table 20 presents the fit indices for the second three-factor BE model from both samples. Table 21 presents the standardized parameter estimates from both samples. A number of post hoc analyses relevant to the third research question were conducted (see Appendix E).

Question Four: What is the impact of overfitting on cross-validation?

Table 22 presents the results of exploring the impact of overfitting the model to the initial sample on the cross-validation on other samples. The initial model found in Figure 3 was fit to each of the four samples. The fit indices for the initial model are presented first in each section of the table followed by the fit indices for the iterative modifications. The iterative modification process was employed for each sample, resulting in four modified models being yielded in between five to seven iterations depending on the sample. No two modified models were the same. For each sample, the final modified model was fit to each of the three cross-validation samples. In each section of the table, the fit indices for the cross-validations are presented at the bottom of the table. Overall, there were 12 cross-validations (4 initial samples x 3 cross-validation samples). 11 of the 12 cross-validations resulted in a substantially worse fit for the cross-validation data than for the initial sample data. When samples 1, 2, and 4 were the initial sample (see Table 22, sections 1, 2, and 4 respectively), the chi-square, the CFI, the NNFI, the SRMSR, and the RMSEA indicated worse fit for the cross-validation samples than for the initial sample. For the nine cases in these three sections, the difference between the initial sample chi-square statistic and the cross-validation sample chi-square statistic was negative, indicating the model provided a worse fit for the cross-

validation sample. When the sample 3 modified model was fit the three cross-validation samples (samples 1, 2, & 4), the model provided a worse fit for the sample 1 and 4 data but provided a better fit for the sample 2 data. Table 22, section 3 presents the fit indices for the cross-validation of the sample 3 model. As can be seen, the model fit the sample 2 data better than the initial sample for all indices with the exception of the NNFI, for which the model fit both samples equally well.

CHAPTER 5: DISCUSSION

In chapter 1, six research goals were set forth. The first two goals were accomplished in chapter 2. The third and fourth goals are addressed by the four research questions, which were operationalized in chapter 3, evaluated empirically by the results in chapter 4 and are interpreted in this chapter. The fifth and sixth goals are addressed in this section as well. This chapter discusses the results by topic, the implications of the study for practice and research, the study limitations, and the insights into the nature of performance gleaned.

Discussion of the Results

This study has extended our understanding of the nature of training performance by applying a version of the Kraiger et al. (1993) multidimensional learning framework to level of proficiency training criteria (Sackett & Mullen, 1993) and confirming the entire three-factor model using confirmatory factor analysis (CFA) with archival data from the Special Forces Qualifications Course (SFQC). Although Tracey and associates (2001) confirmed a partial model with CFA, the entire Kraiger et al. (1993) model had not been previously operationalized for certification training (i.e., where certifying proficiency to a standard is the focus) and tested using CFA. Additionally, several alternative models inspired by the job performance modeling literature were tested revealing that the BE KNOW DO model provided the best and most consistent fit for the data. The BE KNOW DO version of the Kraiger et al. (1993) model was successfully applied to data from two phases of the training as well, allowing the relationship of these constructs to be studied across time and situation for the first time. The training performance factors were found to predict their counterpart later in training. Additionally, two research issues related to performance modeling—the specificity of performance factors and overfitting performance models to the data—were

studied. Exploring the dimensionality of the BE construct suggested that the performance situation and the measurement of the manifest indicators impact the degree of specificity that a particular model might support. In other words, the unspecified variance related to the specific training phases (situation and content) and the measurement methods utilized might be responsible for the failure of the models in question three to provide adequate fit. Cross-validating a model that had been modified to capitalize on the idiosyncratic characteristics of the initial sample supported the recommendations of several authors (e.g., MacCallum & Austin, 2000; Black, 2001; Millsap, 2002) that cross-validating model modifications should be a standard practice. Taken together with the theoretical argument integrating training evaluation and job performance modeling research, the empirical results of this study make a contribution to studying and understanding the nature of performance. In the next two sections, the results are discussed more specifically by the general research topics—the nature of training performance and issues related to modeling performance.

The Nature of Training Performance

The primary research objective of this study was to explore the latent structure of training performance. A three-factor model similar to Kraiger et al. (1993), Campbell's direct determinants, the Motowidlo et al. (1997) determinants, and the performance factors of Wilson and Grant (1997) was proposed, tested, and confirmed. At the very least, the findings of this study represent a confirmation of the Kraiger et al. (1993) model with level-of-performance training criteria using CFA. The first two research questions addressed the issue of the latent structure of training performance. Question one explored the dimensionality of training performance for the entire SFQC program and tested several alternative models. Question two investigated the relationship between training performance constructs across

time and training situation. The three-factor BE KNOW DO model was successfully confirmed for Phases One and Three of training, and a predictive model was fit successfully to the data.

For question one, the results supported the three-factor BE KNOW DO conceptualization of training performance. The model achieved adequate to good fit across the four samples. In an attempt to combat confirmation bias, alternative models were tested as recommended (e.g., MacCallum & Austin, 2000). The support for the Kraiger et al. (1993) version of the BE KNOW DO model was enhanced by model comparison. The alternative models were not identified. Many of the models violated the rank condition of identification when applied to the four samples. The rank condition is necessary but not sufficient for identification (Kaplan, 2000). It should not be surprising that the Campbell version of the BE KNOW DO model and the BE KNOW DO model with a unitary content factor were not identified. Both models violate the rank condition from conception with all items loading on a single factor—creating a situation in which parameter estimates are linear combinations of each other. This could explain why McCloy and colleagues (1994) used that non-standard procedure to estimate the variance of the direct determinants. Their approach assumed the loadings of the manifest indicators on the latent constructs were known, constraining them to either 1 or 0. One of the indications of problems with the Campbell version of BE KNOW DO in this study was the occurrence of out-of-bounds parameter estimates. Did McCloy and his colleagues (1994) conduct CFA first, get out-of-bounds parameter estimates, and then adapt their approach? If this was the case, it suggests the authors knew the model was not identified but failed to mention that in the article. It appears that Campbell has proposed a model for the direct determinants of performance that cannot be confirmed as theorized with

CFA or SEM. McCloy et al. (1994) estimated variances for the three direct determinant factors in support of the model. Is it appropriate to estimate the variance of the latent variables for a model that is not identified? If not, then their variance estimates were not meaningful. Despite the problematic methodology, McCloy et al. (1994) makes an important contribution because of its thorough discussion of the direct determinants. Regardless of the appropriateness of their methodology, the results of this study suggest that using the independent clusters approach (McDonald & Ho, 2002) of the Kraiger et al. (1993) or Wilson and Grant (1997) models is better from a modeling standpoint. A major success factor for modeling appears to be how the constructs are operationalized with manifest indicators.

Although the BE KNOW DO model provided an adequate to good fit for all four samples, a difference in the latent variable correlations was found between two groups of samples. Samples 1 and 3 yielded very similar correlations between the BE and DO constructs (.42 and .47 respectively) and the KNOW and DO constructs (.91 for both). Whereas, for samples 2 and 4, the correlations between BE and DO (.32 and .36 respectively) and KNOW and DO (.46 and .26 respectively) were similar as well. As can be seen in Table 13, the loadings for the manifest indicators varied across the samples for the KNOW and DO constructs in a similar pattern, while the loading were fairly consistent across samples for the BE construct. It should be noted that the peer-rating instrument was the only study measure designed by psychologists. Additionally, the correlation between BE and KNOW was found to be consistent across all four samples (.57, .55, .56, and .56 respectively). This suggested there might be two distinct groups of samples with some difference in demographics, experience, or measurement impacting the KNOW and DO constructs. A series of post hoc analyses (see Appendix C) disconfirmed that the difference was related to year attended or

MOS. However, the post hoc analyses uncovered that the number of significant correlations between knowledge and skill measures was related to the differences. Fewer significant correlations in a sample resulted in lower fit indices and a lower correlation between the KNOW and DO constructs. Interestingly, the differences in the knowledge measures and correlations between samples did not appear to impact the relationship between BE and KNOW. When the sample size was increased, there were no problems, suggesting that the factor driving the difference in the knowledge measures averages or cancels out as the number of participants increases. Interestingly, the 18A model would not converge at all with the KNOW factor. Of the NCO MOSs, the model provided the worst fit for the 18D sample. This indicates that restriction of range in terms of cognitive ability might be the culprit. This hypothesis needs to be tested. Human attribute data from SFAS can be used to explore the differences. However, the difference might not have anything to do with cognitive ability. It could be another factor, such as prior MOS, whether the SFQC candidate was from a combat-arms or a non-combat-arms MOS might impact performance. Other human attribute and experiential variables should be explored as available. The differences in parameter estimates do not invalidate the fact that the BE KNOW DO model provided a consistently good fit for the data across the four samples. However, when interpreting or utilizing the performance models and parameter estimates, the limitations of sample size and composition should be investigated and considered. Although the chi-square will be highly significant, parameter estimates from larger samples might be more stable and interpretable.

The results of question two provide robust support for the three-factor model, confirming it for two different training situations separated in time. For Phase One, the relationships between the latent variables were found to be similar across both samples with

the relationship between KNOW and DO found to be the strongest among the latent variables in both cases. This makes sense considering that Phase One is a more traditional training situation with a strong linkage between the test (KNOW) and simulation (DO) content. Additionally, the majority of correlations between Phase One knowledge and skill indicators were found to be significant (see Tables B7 and B8). Although the model provided good fit for the data, the Phase Three model did have a few problems, which were most likely the result of the DO construct being operationalized with ratings.

The confirmation in both phases allowed the predictive model between the two phases to be tested. The initial predictive model was altered slightly producing a structural model that provided good fit for the data. The BE KNOW DO constructs at Phase One predicted their counterparts in Phase Three significantly with varying degrees of relationship. Based on the findings from the SEM analyses, a series of post hoc models were tested (see Appendix D). These post hoc models isolated the BE KNOW DO model constructs in a series of one- and two-construct models to more fully understand the constructs over time. The results of the two-construct SEM analyses provided evidence that the Phase Three DO construct—operationalized by peer ratings and not FTX results as in Phase One—might be responsible for lowering model fit. The results suggested that the single constructs are stable over time to a varying degree with the BE and KNOW constructs being the most consistent across samples and time. KNOW seemed to maintain the strongest relationship over time and situation. Since declarative knowledge tests are most likely influenced heavily by cognitive ability and educational experience, these measures would be expected to have the strongest relationship across time and situation. BE seemed to have a stable structure across time, but the relationship between BE across phases was lower in magnitude than initially expected

considering the identical operationalization at both times. This suggests that BE might be more heavily influenced by time and/or situation than the other constructs. Beaty and colleagues (2001) demonstrated that personality predicted performance in weak situations (i.e., the cues for performance are ambiguous or unclear) and cognitive ability predicted performance in strong situations (i.e., cues for performance are well-defined or salient). Their explanation was that situations with well-defined cues for successful performance do not allow personality to influence performance. Whereas, successful performance in situations without salient performance cues allows for the operation of personality. The BE construct is probably highly related to or influenced by personality given the construct content, whereas, KNOW is probably not. BE should be expected to vary more across situations than KNOW. In the Hattrup and Jackson (1996) framework, person and situation differences interact simultaneously to create individual behavior and other criterion responses. They believe that situations, individuals, and their responses are multidimensional and that behavior is dynamic and can only be understood in the context of the interaction between the situation and the individual. Criterion measures of performance behavior should be a function of the interaction between human attributes and situational attributes, and performance constructs and measurement methods are most likely influenced to varying degrees by different individual and situational characteristics. Therefore, measures of BE and KNOW should be functions of the interaction of human attributes and situational characteristics. Since BE is thought to be influenced by relatively stable human attributes like personality, differences in BE when measured using the same criteria should be related more to differences in the situation. Whereas, declarative knowledge (KNOW) as measured by tests should be related more to cognitive ability and impacted less by situational characteristics. However, the

relationship of declarative knowledge measures over time should be impacted by the degree of content similarity between the measures and the amount of learning that occurred between measurements.

Interestingly, the fit of the predictive model was not acceptable until the disturbance terms for the Phase Three DO and BE factors were correlated. This was justifiable because both factors were operationalized with peer ratings captured at the same time and in the same training context from the same raters. The unspecified factor represented by correlating the disturbances is probably a combination of using a common method and the same raters at the same time in the same training context. Since Hattrup and Jackson (1996) believe situations are multidimensional (i.e., characterized by many facets or attributes) and should be treated as whole entities, this unspecified factor could be considered to measure the influence of the situation. This supports the idea that situational differences might have influenced the relationship between the BE constructs across phases. Both BE constructs are operationalized using the same method and content. The only difference is time, raters, and situation, and the first two can be classified as situational characteristics. When you consider the relationship between DO and BE at Phase Three, the operationalization of the latent variables in the model, and the strong relationship of KNOW across phases, it suggests that differences in the two situations might account for the magnitude of relationship between BE at Phase One and BE Phase Three.

One weakness was not utilizing measurement invariance techniques (Vandenberg & Lance, 2000) to investigate the latent structure over time. Unfortunately, since all constructs were not measured using the same indicators at both times, this would not have been possible. However, demonstrating that the three-factor model holds for two different training

situations over time and that there are robust relationships between like constructs is an extension of the literature.

Issues Related to Modeling Performance

Questions three and four of the dissertation relate to performance modeling issues. Specifically, question three dealt with the issue of specificity in modeling performance content, and question four explored the issue of capitalizing on idiosyncratic characteristics to improve model fit and its impact on cross-validation. The results and interpretation of these two questions are straightforward.

Question three addressed the issue of content specificity in modeling performance constructs. Two alternative models were tested—a one-factor and a three-factor. Neither the one-factor general model nor the three-factor specific model provided an acceptable fit for the data of either sample. The post-hoc analyses (see Appendix E) did not provide any conclusive evidence to support confirmation or disconfirmation of either model. However, the post hoc analyses suggested that personal discipline may not be a component of BE and that team and peer facilitation and effort were highly similar constructs. The results of the proposed and post hoc analyses suggested that the issue was related to the use of archival data and unspecified variance. The measures of personal discipline (negative spot reports and negative peer nominations) were suspected to be problematic a priori, and the influence of unspecified constructs, like the training situation, seemed to lower model fit. The modification indices repeatedly suggested correlating the error terms of measures collected during the same training phase, and spot reports and peer nominations were not ideal measures of personal discipline. Therefore, the results for this question cannot be interpreted

as support for either model. More research, preferably with data collected utilizing designed measures (i.e., not archival), should be pursued.

Question four investigated the issue of overfitting a model on a sample and its impact on cross-validation. In 11 of the 12 cases, the over-modified model provided a worse fit upon cross-validation, suggesting that the modifications had taken advantage of idiosyncratic variance in the initial sample. This demonstrates the importance of always cross-validating modified models or models generated from empirical data. Authors (e.g., MacCallum & Austin, 2000; Black, 2001; Millsap, 2002) are justified in recommending that cross-validation be a standard practice for model modification. The results of this study suggest that modified models capitalizing on the idiosyncratic variance of a single sample would not fair well when cross-validated on another sample, even one from the exact same population at the exact same time. Modified models presented in the literature without cross-validation should be viewed with skepticism. Also, the results suggest the sample splitting methodology is sufficient to detect modifications that capitalize on idiosyncratic variance when enough cases are present to utilize the methodology. Using the sample splitting methodology or multiple samples becomes increasingly important when the modifications are not theoretically justified or the research is engaged in exploratory work.

Implications and Future Directions

Practice

The implications for practice are fairly clear. Organizations need to approach training with a model of performance. Training designers need to incorporate measurement of the three performance constructs into their courses with a variety of measurement methods. Additionally, these measures should take place across the entire course and its content

domains. Measures at the end of training are useful. However, these measures might provide a differential view of performance or differential prediction of future performance than measures captured during other phases of training. Measures at different points in training might allow for content differences to be assessed or offer the opportunity to measure knowledge and skill acquisition or integration into existing mental structures. In order to serve as a direct determinant of job performance, the recommendations in the integration section of chapter 2 should be put into practice, knowing that iterative tweaking will be necessary. Linking training performance to more distal performance or to performance at different levels (e.g., team) requires a guiding framework and should be viewed as a potentially powerful tool by organizations. This dissertation has suggested that using a three-factor model of performance across all levels and contexts might be such a framework. Organizations should consider the "big picture" when designing training, performance management and selection systems. Utilizing a standardized model of performance in an organization would ensure construct congruence for process evaluation and validation. A common or complementary performance framework used at multiple levels would facilitate the alignment of individual and team performance with organizational goal and strategy. Regardless of the model or purpose, improved performance measurement can only benefit organizations.

More attention should be given to designing and collecting training metrics. The results suggest that the poor quality of some measures and the restriction of range on others might have impacted model fit in this study. For example, negative spot reports were found to be poor measures of the DO construct in most cases. This probably was a result of the measurement characteristics and multiple uses of spot reports. In cases where restriction of

range on certain measures is expected, there are often alternative measures of related constructs that are more appropriate. For example, declarative knowledge tests might be expected to suffer from the effects of range restriction in certification training. However, is declarative knowledge (i.e., content) the real construct of interest? Maybe, the job involves working in a fast-paced, constantly changing regulatory environment, such as biotechnology, and requires the incumbent to quickly absorb and integrate the new information into their existing knowledge structures. In that case, training metrics designed to assess the proficiency of knowledge acquisition and integration into existing knowledge structures would be more appropriate and would fit with recommendations in the literature (e.g., Kraiger et al., 1993). Often, tests of declarative knowledge are easier to design and administer. Practical constraints must be considered. Although no clear recommendations can be made as to level of specificity of content measurement and performance modeling (i.e., question three), more specific measurement can usually serve multiple purposes and be aggregated into more general measures. Ultimately, the purpose of measurement and resource constraints will determine the metrics utilized as well as many other characteristics of the measurement process.

Finally, the distinction between level-of-performance and change-oriented training evaluation (Sackett & Mullen, 1993) should be revisited from a practical standpoint. It is clear that a distinction in terms of purpose can be made and that level-of-performance evaluation is more beneficial in terms of measuring and linking individual training performance to more distal performance measures and outcomes. However, level-of-performance measurement is more costly and involved (Sackett & Mullen, 1993) and may not allow for the assessment of the training program or comparison between training

methods. Organizations should explore the pros and cons of always incorporating level-of-performance measurement into training. However, level-of-performance measurement could be operationalized in such a way as to allow for change to be assessed. Given the advances in the measurement and analysis of change (Riordian, Richardson, Schaffer, and Vandenberg, 2001; Chan, 2002), measuring level-of-performance criteria as defined by the three constructs in the BE KNOW DO model at multiple points throughout training would provide for the assessment of change at the latent level. This future direction for practice fits in the next section as well, and the idea of using latent models of training performance (i.e., level-of-performance evaluation) for change-oriented training program evaluation should be explored as an avenue of methodological research. Using repeated measures throughout training and analyzing the data with latent growth modeling techniques (e.g, Chan, 2002) would be a good place to start.

Research

The implications and future directions for the field of I/O psychology are important when the empirical results are considered in relation to the conceptual arguments presented to integrate training evaluation and job performance modeling research and to define BE (i.e., the third factor) in terms of fit. The confirmation of a three-factor training performance model using level-of-proficiency measures supports the Kraiger et al. (1993) model and provides a foundation to address many questions, including the mediation question posed in chapter 2, in the future. The suggestion that a context-free, three-factor model of performance could be used to describe all performance might serve to stimulate much research in I/O. Finally, one of the most interesting observations of the study is the impact of the situation on performance modeling, which might prove to be a fruitful area of future research. Although

pages of research implications and future directions could be generated, only a few specific directions for future research suggested by this work are presented.

The Three-Factor Training Model

A three-factor model was confirmed for training performance utilizing archival data from SF training. Future research should investigate whether the model holds for the constructs when operationalized using other manifest indicators (i.e., different measurement methods) and for both archival indicators and ones designed a priori to measure the constructs. Additionally, does this model hold for different populations and work types. According to the framework of Hattrup and Jackson (1996), situational characteristics interact with individual characteristics to influence criterion measures. The nature of the work and the population are factors that could interact to impact performance measurement. In addition to cross-validating the model across different training contexts, determining the measures to best operationalize the constructs when the organization, population, or work exhibits certain characteristics would be another area of future research.

Situational Impacts on Performance

The findings suggest that situational variation might have been influencing model fit. The work of Hattrup and Jackson (1996) provide a framework for categorizing situations. They propose using characteristics in four categories—information, task, physical, and social—to classify situations in terms of the extent that situations provide uniform expectations for appropriate behavior or cues for successful performance. Mischel (1977) originally proposed the idea of strong and weak situations interacting with human attributes to constrain behavior. Weak situations provide little or inconsistent cues for appropriate behavior or successful performance. Strong situation provide cues or information that

prescribe what behaviors are appropriate or that lead to successful performance. Hattrup and Jackson (1996) note that the strength of the situation relies heavily on the consistency of perceptions and interpretations across individuals. The perceptions of people in the organization determine the appropriateness of actions or responses in various situations. This is particularly the case when ratings or other subjective measurement methods are utilized to operationalize performance. In fact, measures like ratings, spot reports or awards should be primarily measures of fit between situational expectations and individual characteristics and behavior. The impact of situational differences on the measurement and modeling of performance should be studied. Additionally, when using training performance to predict job performance, the possibility of differential prediction resulting from situational differences impacting the criterion measurement should be explored. Beatty and colleagues (2001) looked at the differential prediction issue from the other side and found that different human attributes are more predictive of performance in certain situations (i.e., strong versus weak). Mischel (1977) posited that strong situations would constrain personality. Beaty et al. (2001) demonstrated this result. Continued research in this vein is justified as well.

The Mediation Question

Does training performance when serving as direct determinants fully or partially mediate the relationship between indirect determinants (i.e., human attributes) and job performance? Although not addressed in this study, the findings provide a foundation to explore the mediation question. The training evaluation literature has provided evidence that the outcomes of training partially mediate the relationship between human and situational variables and job performance (e.g., Colquitt et al., 2000). Whereas, the two major job performance theories (Campbell, 1999; Motowidlo et al., 1997) posit full mediation by the

direct determinants. This dissertation presented a compelling argument that the three-factor model operationalized with level-of-performance criteria could serve as training performance and the direct determinants of future performance. Additionally, given the perspective of Hattrup and Jackson (1996), the interaction of indirect determinants (individual and situational characteristics) needs to be modeled as well. Research into the mediation question is definitely needed. Data from the SF pipeline could be used not only to operationalize an entire performance model (i.e., indirect determinants, direct determinants, and performance factors) but also to test the mediation question and to explore modeling the interaction of situational and individual characteristics. Research to address the mediation question should be undertaken in other situations as well. Unfortunately, few non-military organizations collect the extensive data needed to operationalize a complete model.

Context-Free Three-Factor Performance Model

Research into the purposed context-free, content-free three-factor model of performance should be undertaken. The same research design utilized to address the mediation question could be used to explore the viable of the three-factor model if performance throughout the model was operationalized successfully as three-factors. Of course, the factors would have to approximate cognitive, skill-based, and affective performance or relate to human attributes expected to underlie these constructs. Research should be conducted in a wide variety of settings utilizing a variety of measures at different levels of specificity to fully explore the model. Confirming the model across a number of situations would provide strong support for its viability.

Specificity of Modeling Performance Content

The question related to the specificity of performance measurement and modeling was not successful in determining if a general factor or several specific factors were most appropriate for modeling performance. Future research should investigate this issue utilizing the knowledge and skill factors as well as the affective factor with data from one performance situation as to eliminate the impact of situational variation. Additionally, the issue should be investigated in the context of all three constructs simultaneously. In other words, is the content-general, three-factor model or the content-specific, nine-factor model a better fit? Models that operationalize second-order factors might be explored as well. Regardless, the specificity of the construct depends on the manifest indicators available for modeling. By investigating this issue across numerous situations, a clear set of recommendations may emerge.

Impact of Imputation

If missing data had been a problem that restricted usable sample size and prevented testing a research question, then it would have been necessary to employ a missing data technique to increase sample size. This has been a common practice in performance research. For example, Project A employed two strategies (Campbell & Knapp, 2001). For variables with the data available for 90% or more of participants, a summary score (i.e., mean) was calculated and used. In cases where data for 90% or more of participants were not available, Proc Impute (Wise & McLaughlin, 1988, as cited in Campbell & Knapp, 2001) was used to predict the missing data based on the values of other variables. For example, in Project A's Batch A MOS CV1 sample, only 15% of the total sample (N = 5,268) had complete data. The post-imputation sample contained 4,039 complete cases, 77% of the total sample. The

impact on imputation needs to be studied to determine its effect on modeling performance. Addressing when and how it is appropriate to impute data for modeling would be an important contribution.

Limitations

The major limitation of this study was the archival nature of the data. The researcher had no control over the design of measures and the initial collection of the data. As a result, the best measures were selected from the available data. The manifest indicators might have suffered from some criterion contamination (Goldstein, 1993; Gatewood & Field, 1998), which Campbell (1999) indicates can cause serious problems. If the criterion measures were poor, this might have inaccurately lowered the validity coefficients or correlations with other indicators. The use of multiple indicators and CFA and SEM would have helped "correct" for reliability and contamination problems in some indicators. Criterion deficiency in some of the manifest indicators might have contributed to weak correlations as well. The silver lining is that despite the potential data issues the three-factor model was found not only to be tenable but robust across samples and training situations. This suggests that data designed to operationalize the three-factor model would produce exceptional results. An additional measurement issue is restriction of range. The results suggest that restriction of range might be impacting the model fit in terms of the relationships between knowledge and other indicators. The selection and assessment process for SF ensures a certain amount of range restriction as does training to mastery in order to certify an a priori level of proficiency. Given that cognitive ability is a major selection criterion for SF and for assignment to certain MOSs, the fact that declarative knowledge measures suffer from restriction of range is not surprising. One potential solution would be to include other measures of cognitive

performance in addition to declarative knowledge tests, like measures of knowledge structure or acquisition as suggested by Kraiger and his colleagues (1993). A related issue is differential attrition in that the folders of many soldiers were missing key data points and therefore were excluded from analysis. If these candidates differed significantly from the ones in the sample, the model of performance might be inaccurate. In terms of modeling, although alternative models were operationalized and tested, one weakness is that all the alternative models were versions of the three-factor model. Testing a two-factor model that combines knowledge and skill measures into a task construct would be justifiable given the support for the two-factor model of task and contextual performance. Finally, the nature of the population might impact the ability to generalize the findings to other populations. There are few jobs in other organizations that parallel the work and performance context of SF. However, similarities do exist with some civilian jobs in public and private sector organizations. Some law enforcement jobs are obvious candidates. However, the SF roles of team leader, medic, and communications sergeant have commonalities with civilian jobs as well. These findings should be replicated in other organizations utilizing both archival data and measures designed to operationalize the three-factor model.

Insights into the Nature of Performance

From the conceptual arguments presented, there are several important insights to discuss. First, the BE construct was defined in terms of fit, an idea that reflects the concept that certain behaviors and human characteristics are more appropriate in certain organizations (Kristoff, 1996) or situations (Hattrup & Jackson, 1996; Beaty et al., 2001) and that more subjective measures tap into these fit perceptions. This resembles the work habits determinant in the Motowidlo et al. (1997) model and the citizenship factor of the Wilson

and Grant (1997) model. Second, a persuasive argument for integrating training and job performance research was presented. This argument suggested that training performance as operationalized as level of performance on the Kraiger et al. (1993) constructs could serve as the direct determinants of job performance in the Campbell and/or Motowidlo et al. (1997) models. Given the empirical evidence and the theoretical argument, this seems justifiable. Third, a general performance model was posited at the end of the integration section. A three-factor model was proposed—potentially utilizing cognitive, skill-based, and affective performance constructs. This model using the terminology from the training evaluation literature resembles several models in the job performance literature, notably the Wilson and Grant (1997) work with the North Carolina Highway Patrol. The idea that all performance can be described in terms of three factors regardless of the context (training or job), the content (e.g., computer skills), the situation (e.g., Robin Sage versus Isolation), the measures (e.g., skill indicators as times to criterion or ratings), or the level (individual, team or organization) is a powerful concept that should be investigated. The integration and extension of the current training evaluation and job performance literatures is a necessary first step. Future research must build on the integration suggested here to explore the context-free model.

What is the most important finding? The most important finding from this research was confirming the entire three-factor training performance model utilizing CFA with level-of-performance criteria. Additionally, investigating the latent structure of performance across time and situation was a close second. The most important insight gained from the data involved the knowledge that the influence of the situation (Hattrup & Jackson, 1996) should be studied. Finally, the most important ideas generated in this dissertation are related to the

integration of training evaluation and job performance modeling research and the proposal of a three-factor context-free model of performance. Both these ideas are now possible, even plausible. Possibilities generate more research and deeper understanding. In the near future, the evidence to address several questions raised by this dissertation—including the mediation question—should be available. Hopefully, the work presented here will serve as a foundation for much of this future research.

References

Alliger, G. M., & Janak, E. A. (1989). Kirkpatrick's levels of training criteria: Thirty years later. *Personnel Psychology, 42*, 331-342.

Alliger, G. M., Tannenbaum, S. I., Bennett, W., Traver, H., & Shotland, A. (1997). A meta-analysis of the relations among training criteria. *Personnel Psychology, 50*, 341-358.

Astin, A. W. (1964). Criterion-centered research. *Educational and psychological measurement, 24*, 807-822.

Austin, J. T., & Villanova, P. (1992). The criterion problem: 1917-1992. *Journal of Applied Psychology, 77*, 836-874.

Avery, R. D., & Murphy, K. R. (1998). Performance evaluation in work settings. *Annual Review of Psychology, 49*, 141-168.

Beaty, J. C., Cleveland, J., & Murphy, K. R. (2001). The relationship between personality and contextual performance in "strong" versus "weak" situations. *Human Performance, 14*, 125-148.

Bentler, P. M., (1990). Comparative fix indexes in structural models. *Psychological Bulletin, 107*, 238-246.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*, 588-606.

Black, W. C. (2001). Invited reaction: The influence of individual characteristics and the work environment on varying levels of training outcomes. *Human Resources Development Quarterly, 12*, 25-31.

Borman, W. C., Hanson, M. A., & Hedge, J. W. (1997). Personnel selection. *Annual Review of Psychology, 48*, 299-337.

Borman, W. C., & Motowidlo, S. J. (1993). Expanding the criterion domain to include elements of contextual performance. In N. Schmit, W. C. Borman & Associates (Eds.), *Personnel selection in organizations* (pp. 71-98). San Francisco: Jossey-Bass.

Borman, W. C., White, L. A., & Dorsey, D. W. (1995). Effects of ratee task performance and interpersonal factors on supervisor and peer performance ratings. *Journal of Applied Psychology, 80*, 168-177.

Broad, M. L., & Newstrom, J. W. (1992). *Transfer of training: Action-packed strategies to ensure high payoff from training investments*. Reading, MA: Addison-Wesley.

Brooks, J. E. (1997). Introduction. In J. E. Brooks & M. M. Zazanis (Eds.), *Enhancing U.S. Army Special Forces: Research and applications* (pp. 3-6). Alexandra, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Browne, M. W., & Cudeck, R. (1993). Alternate ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Thousand Oaks, CA: Sage.

Caligiuri, P. M., & Day, D. V. (2000). Effects of self-monitoring on technical, contextual, assignment-specific performance. *Group & Organization Management, 25,* 154-174.

Campbell, J. P. (1990). Modeling the performance prediction problem in industrial and organizational psychology. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (pp. 687-732). Palo Alto, CA: Consulting Psychologists Press.

Campbell, J. P. (1999). The definition and measurement of performance in the new age. In D. R. Ilgen & E. D. Pulakos (Eds.), *The changing nature of performance: Implications for staffing, motivation, and development* (pp. 399-429). San Francisco: Jossey-Bass.

Campbell, J. P., Gasser, M. B., & Oswald, F. L. (1996). The substantive nature of job performance variability. In K. R. Murphy (Ed.), *Individual differences and behavior in organizations* (pp. 258-299). San Francisco: Jossey-Bass.

Campbell, J. P., & Knapp, D. J. (2001). *Exploring the limits of personnel selection and classification*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Campbell, J. P., McCloy, R. A., Oppler, S. H., & Sager, C. E. (1993). A theory of performance. In N. Schmit, W. C. Borman & Associates (Eds.), *Personnel selection in organizations* (pp. 35-70). San Francisco: Jossey-Bass.

Campbell, J. P., McHenry, J. J., & Wise, L. L. (1990). Modeling job performance in a population of jobs. *Personnel Psychology, 43,* 313-333.

Chan, D. (2002). Latent growth modeling. In F. Drasgow & N. Schmitt (Eds.), *Measuring and analyzing behavior in organizations: Advances in measurement and data analysis* (pp. 302-349). San Francisco: Jossey-Bass.

Colquitt, J. A., LaPine, J. A., & Noe, R. A. (2000). Toward a integrative theory of training motivation: A meta-analytic path analysis of 20 years of research. *Journal of Applied Psychology, 85,* 678-707.

Conway, J. M. (1999). Distinguishing contextual performance from task performance for managerial jobs. *Journal of Applied Psychology, 84,* 3-13.

Conway, J. M., Lombardo, K., & Sanders, K.C. (2001). A meta-analysis of incremental validity and nomological networks for subordinate and peer ratings. *Human Performance, 14,* 267-303.

Davenport, T. O. (1999). *Human capital: What it is and why people invest it.* San Francisco: Jossey-Bass.

DuBois, C. L., Sackett, P. R., Zedeck, S., & Fogli, L. (1993). Maximal versus typical performance. *Journal of Applied Psychology, 78*, 205-211.

Elder, G. H., Jr., Pavalko, E. K., & Clipp, E. C. (1993). *Working with archival data: Studying lives*. Newbury, CA: Sage.

Ford, J. K., Smith, E. M., Weissbein, D. A., Gully, S. M., & Salas, E. (1998). Relationships of goal orientation, metacognitive activity, and practice strategies with learning outcomes and transfer. *Journal of Applied Psychology, 83*, 218-233.

Gatewood, R. D., & Field, H. S. (1998). *Human resources selection*. Fort Worth, TX: The Dryden Press.

Goldstein, I. (1993). *Training in organizations*. Pacific Grove, CA: Brooks/Cole Publishing.

Gottfredson, L. S. (1991). *The evaluation of alternative measures of job performance*. Washington, DC: National Academy of Science.

Grant, L. (1996). A comprehensive examination of the latent structure of job performance (Doctoral Dissertation, North Carolina State University, 2000). *Dissertation Abstracts International, 57,* 6629.

Hatcher, L. (1994). *A step-by-step approach to using the sas system for factor analysis and structural equations modeling.* Cary, NC: SAS Institute, Inc.

Hattrup, K., & Jackson, S. E. (1996). Learning about individual difference by taking situations seriously. In K. R. Murphy (Ed.), *Individual differences and behavior in organizations* (pp. 507-547). San Francisco: Jossey-Bass.

Hough, L. M., & Oswald, F. L. (2000). Personnel selection: Looking toward the future--remembering the past. *Annual Review of Psychology, 51*, 631-664.

Howell, D.C. (1992). *Statistical methods for psychology*. Belmont, CA: Duxbury Press.

Hu, L., & Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods, 3*, 424-453.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1-55.

Hunt, S. T. (1996). Generic work behavior: An investigation into the dimensions of entry-level hourly job performance. *Personnel Psychology, 49,* 51-83.

Hunter, J. E. (1983). A causal analysis of cognitive ability, job knowledge, job performance, and supervisor ratings. In F. J. Landy, S. Zedeck and J. Cleveland (Eds.), *Performance measurement and theory* (pp. 257-266). Hillsdale, NJ, Lawrence Erlbaum Associates.

Jackson, D. L. (2001). Sample Size and Number of Parameter Estimates in Maximum Likelihood confirmatory factor analysis: A Monte Carlo investigation. *Structural Equation Modeling, 8*, 205-223.

Kaplan, D. (2000). *Structural equation modeling: Foundations and extensions*. Thousand Oaks, CA: Sage.

Kirkpatrick, D. L. (1959a). Techniques for evaluating training programs--part 1: Reactions. *Journal of American Sociate of Training Directors, 13*, 3-9.

Kirkpatrick, D. L. (1959b). Techniques for evaluating training programs--part 2: Learning. *Journal of American Sociate of Training Directors, 13*, 21-26.

Kirkpatrick, D. L. (1960a). Techniques for evaluating training programs--part 3: Behavior. *Journal of American Sociate of Training Directors, 14*, 13-18.

Kirkpatrick, D. L. (1960b). Techniques for evaluating training programs--part 4: Results. *Journal of American Sociate of Training Directors, 14*, 28-32.

Kirkpatrick, D. L. (1967). Evaluation of training. In R. L. Craig (Ed.), *Training and development handbook: A guide to human resources development*. New York: McGraw-Hill.

Kirkpatrick, D. L. (1979). Techniques for evaluating training programs. *Training and Development Journal, 33,* 78-92.

Kirkpatrick, D. L. (1996). *Evaluating training programs: The four levels*. San Francisco: Berrett-Koehler.

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* (pp. 157-210). San Francisco: Jossey-Bass.

Kozlowski, S.W.J., Gully, S.M., Brown, K.G., Salas, E., Smith, E.M., & Nason, E.R. (2001). Effects of training goals and goal orientation traits on multidimensional training outcomes and performance adaptability. *Organizational Behavior and Human Decision Processes, 85(1),* 1-31.

Kraiger, K. (1999). Performance and employee development. In D. R. Ilgen & E. D. Pulakos (Eds.), *The changing nature of performance: Implications for staffing, motivation, and development* (pp. 366-396). San Francisco: Jossey-Bass.

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.

Kristoff, A. L. (1996). Person-organization fit: An integrative review of its conceptualizations, measurement, and implications. *Personnel Psychology, 49,* 1-50.

Lance, & Vandenberg, (2002). Confirmatory factor analysis. In F. Drasgow & N. Schmitt (Eds.), *Measuring and analyzing behavior in organizations: Advances in measurement and data analysis* (pp. 221-254). San Francisco: Jossey-Bass.

Landy, F. J., & Farr, J. L. (1983). *The measurement of work performance: Methods, theory, and applications*. New York: Academic Press.

Loehlin, J. C. (1998). *Latent variable models: An introduction to factor, path, and structural analysis.* Mahwah, NJ: Lawrence Erlbaum.

Long, J. S. (1983). *Confirmatory factor analysis: A preface to LISREL*. Newbury Park, CA: Sage.

Loviscky, G. E., Rosenberg, A. S., Mathieu, J. E., & Mohammed, S. (1998). *Predicting task and contextual performance in a team setting.* Thirteenth Annual Conference of the Society for Industrial and Organizational Psychology, Dallas, TX.

MacCallum, R. C. (1995). Model specification: Procedures, strategies and related issues. In R. H. Hoyle (Ed.), *Structural equations modeling: Concepts, issues, and applications* (pp. 16-36). Thousand Oaks, CA: Sage.

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology, 51*, 201-226.

MacCallum, R. C., Roznowski, M., Mar, C. M., & Reith, J. V. (1994). Alternative strategies for cross-validation of covariance structure models. *Multivariate Behavioral Research, 29*, 1-32.

Marrs, R. W. (Fall 2000). SFAS redesign: An essential evolution. *Special Warfare, 13(4),* pp. 2-5.

McArdle, J. J. (1996). Current directions in structural factor analysis. *Current Directions in Psychological Science, 5(1),* 11-18.

McCloy, R. A., Campbell, J. P., & Cudeck, R. (1994). A confirmatory test of a model of performance determinants. *Journal of Applied Psychology, 79,* 493-505.

McCrae, R. R., & Costa, P. T. (1996). Toward a new generation of personality theories: Theoretical contexts for the five-factor model. In J. S. Wiggins, *The five-factor model of personality: Theoretical perspectives* (pp. 51-87). New York: The Guilford Press.

McDonald, R. P., & Ho, M. H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods, 7,* 64-82.

Millsap, R. E. (2002). Structural equation modeling: A user's guide. In F. Drasgow & N. Schmitt (Eds.), *Measuring and analyzing behavior in organizations: Advances in measurement and data analysis* (pp. 257-301). San Francisco: Jossey-Bass.

Mischel, W. (1977). The interaction of person and situation. In D. Magnusson & N. S. Endler (Eds.), *Personality at the crossroads: Current issues in interactional psychology* (pp. 333-352). Hillsdale, NJ: Lawrence Erlbaum Associates.

Motowidlo, S. J., Borman, W. C., & Schmit, M. J. (1997). A theory of individual differences in task and contextual performance. *Human Performance, 10,* 71-83.

Motowidlo, S. J., & Schmit, M. J. (1999). Performance assessment in unique jobs. In D. R. Ilgen & E. D. Pulakos (Eds.), *The changing nature of performance: Implications for staffing, motivation, and development* (pp. 56-86). San Francisco: Jossey-Bass.

Motowidlo, S. J., & Van Scotter, J. R. (1994). Evidence that task performance should be distinguished from contextual performance. *Journal of Applied Psychology, 79,* 475-480.

Nagle, B. F. (1953). Criterion development. *Personnel Psychology, 6,* 271-288.

Organ, D. W. (1988). *Organizational citizenship behavior: The good soldier syndrome*. Lexington, MA: Lexington.

Organ, D. W. (1997). Organizational citizenship behavior: It's construct clean-up time. *Human Performance, 10,* 85-97.

Ramirez, A. E. (2000). Individual, attitudinal, and organizational influences on training effectiveness: A test of Noe's model (Doctoral Dissertation, University of Tulsa, 2000). *Dissertation Abstracts International, 61,* 1122.

Riordan, C. M., Richardson, H. A., Schaffer, B. S., & Vandenberg, R. J. (2000). Alpha, beta, and gamma change: A review of past research with recommendations for new directions. In C. A. Schriesheim & L. L. Neider (Eds.), *Research in management: Vol. 1. Equivalence in measurement* (pp. 51-98). Greenwich, CT: Information Age.

Russell, T. L., Crafts, J. L., Tagliareni, F. A., McCloy, R. A., & Barkley, P. (1994). *Job analysis of Special Forces jobs* (ARI Technical Report). Alexandra, VA: Army Research Institute for the Behavioral and Social Sciences.

Sackett, P. R., & Mullen, E. J. (1993). Beyond formal experimental design: Towards an expanded view of the training evaluation process. *Personnel Psychology, 46,* 613-627.

Salas, E., & Cannon-Bowers, J. A. (2001). The science of training: A decade of progress. *Annual Review of Psychology, 52,* 471-499.

Schmitt, N., & Chan, D. (1998). *Personal selection: A theoretical approach*. Thousand Oaks, CA: Sage Publications.

Schwarzwald, J., Koslowsky, M., & Mager-Bibi, T. (1999). Peer rating versus peer nominations during training as predictors of actual performance criteria. *Journal of Applied Behavioral Science, 35*, 360-373.

Simon, S. J., & Werner, J. M. (1996). Computer training through behavior modeling, self-paced, and instructional approaches: A field experiment. *Journal of Applied Psychology, 81*, 648-659.

Surface, E. A. (2000). [Review of the book *Results: How to assess performance, learning and results in organizations*]. *Personnel Psychology, 53*, 236-240.

Swanson, R. A., & Holton, E. F. (1999). *Results: How to assess performance, learning, and perception in organizations*. San Francisco: Berrett-Koehler.

Talbot, C. (1992). Evaluation and validation: A mixed approach. *Journal of European Industrial Training, 16*, 26-32.

Tannenbaum, S. I., Cannon-Bowers, J. A., Salas, E., & Mathieu, J. E. (1993). *Factors that influence training effectiveness: A conceptual model and longitudinal analysis* (Technical Report 93-011). Orlando, FL: Naval Training Systems Center.

Taylor, P. J., & O'Driscoll, M. P. (1998). A new integrated framework for training needs analysis. *Human Resources Management Journal, 8*, 29-50.

Tesoro, F., & Tootson, J. (2000). *Implementing global performance measurement systems: A cookbook approach.* San Francisco: Jossey-Bass.

Tracey, J. B., Hinkin, T. R., Tannenbaum, S. I., & Mathieu, J. E. (2001). The influence of individual characteristics and the work environment on varying levels of training outcomes. *Human Resources Development Quarterly, 12(1),* 5-22.

Tracey, J. B., Tannenbaum, S. I., & Kavanagh, M. J. (1995). Applying trained skills on the job: The importance of the work environment. *Journal of Applied Psychology, 80,* 239-252.

U. S. Army (1999). *Army leadership* (Field Manual 22-100 (electronic)). Fort Leavenworth, KS: Center for Army Leadership.

Van Scotter, J. R., & Motowidlo, S. J. (1996). Interpersonal facilitation and job dedication as separate facets of contextual performance. *Journal of Applied Psychology, 81*, 525-531.

Viswesvaran, C., & Ones, D. S. (2000). Perspectives on models of job performance. *International Journal of Selection and Assessment, 8(4),* 216-226.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4-69.

Warr, P., Allan, C., & Birdi, K. (1999). Predicting three levels of training outcomes. *Journal of Occupational and Organizational Psychology, 72*, 351-375.

Werner, J. M. (1994). Dimensions that make a difference: Examining the impact of in-role and extrarole behavior on supervisory ratings. *Journal of Applied Psychology, 79,* 98-107.

Wilson, M. A., & Grant, L. (1997). *Validation of a trooper selection system: Project technical report* (Technical Report 96-1147 NCSU). Raleigh, NC: North Carolina State University.

Zazanis, M. M., Zaccaro, S. J., & Kilcullen, R. N. (2001). Identifying motivation and interpersonal performance using peer evaluations. *Military Psychology, 13(2),* 73-88.

Footnotes

[1] The order of the research questions changed from the original proposal. Questions two and three have switched positions. The new order reflects the acknowledgement of grouping by question topic. The first two research questions relate directly to the confirmation of a three-factor training performance model. Even though the current question two relates to the structure of training performance over time, confirming the three-factor BE KNOW DO model of training performance for both Phase One and Phase Three is a prerequisite for investigating the time issue. The second two research questions address issues related to modeling performance.

[2] In 2001, the SF pipeline was restructured slightly and relabeled. SFAS became Phase One, and the other phases moved one position back (e.g., Phase One, the first phase of SFQC, became Phase Two). In this document, since all the data were from participants who attended prior to 2001, the previous labels are used. Phase One refers to the first phase of SFQC, and Phase Three is the final phase of SFQC training.

[3] The "Data Collection Procedure" section replaced the "Data and Measurement Issues" section. The only deletion from the section was the potential data issues section. It was deleted because none of the potential problems were issues—for example, we had enough useable data to operationalize the models.

[4] The training performance measures (manifest indicators) are described in limited detail in some cases as to protect the "test security" of some measurement instruments that are still used in the same or similar form.

[5] The introduction of the research models was moved forward in the methods section to reflect when it occurred in the process. The models were not operationalized with the final

manifest indicators until the data collection was complete and the descriptive statistics computed and evaluated to allow for the selection of the variables. All model figures were removed from the body of the text and placed at the end of the document per APA style. All figures were changed from the proposal to reflect the selection of manifest indicators.

[6] The following changes were made to the analytic procedure section. The order of the sub-sections was changed. The discussion of evaluating model fit is now last, following the procedure utilized by research questions. Several sub-sections were integrated and expanded to become the modeling conventions and strategy section. However, there were no substantive changes in strategy—just more references and examples. A section elaborating on the model modification and post hoc strategies was added. A section related to the additional data processing needed to prepare the dataset for modeling was added; it discusses recoding, transforming, and drawing subsets for each question. The only deletion was one section that discussed potential issues that did not come to pass. One paragraph dealt with imputing data if missing data was problematic.

[7] The procedure for questions one, two, and three have been elaborated. In the proposal, the procedure was general. The new procedure is more specific, reflects the constraints placed on the procedure by the data collected, presents what was done, and does not differ in intent or operations from the proposal. One of the differences relates to the number of cross-validation samples. The data dictated the number of samples. Nor could the same indicators and samples be used for all questions as hoped in the proposal.

[8] The procedure for question four was changed substantially. The original procedure was found to be untenable and was modified to retain the intent of the question. As an added bonus, the new procedure resulted in more presentable and interpretable output.

Table 1.

A Comparison of Models of Job Performance

Model of Performance	Performance Factors	Performance Determinants
Campbell Model		
Campbell (1990) Campbell et al. (1990) Campbell et al. (1993) McCloy et al. (1994) Campbell et al. (1996) Campbell (1999)	Job-specific task proficiency Non-job-specific task proficiency Written and oral communication Demonstrating effort Maintaining personal discipline Facilitating peer & team performance Supervision & leadership Management & administration	Declarative knowledge (DK) Procedural knowledge and skill (PKS) Motivation (M)
Task-Contextual Performance Model		
Borman & Motowidlo (1993) Motowidlo & Van Scotter (1994) Borman et al. (1995) Van Scotter & Motowidlo (1996) Motowidlo et al. (1997)	Task performance Contextual performance	Task knowledge Task skill Task habits Contextual knowledge Contextual skill Contextual habits
In-Role & Extra-Role Model		
Werner (1994)	In-role Extra-role	Not Specified
Grant (1996)	Know in-role Do in-role Extra-role	Not Specified
Other Performance Models		
Hunter (1983) Meta-analysis	General Performance Factor (Supervisory Ratings)	Knowledge Skill
Viswesvaran (1993)	Ten factors: One Overall Performance Factor with nine others	Not Specified
Hunt (1996)	Nine Factors	Not Specified
Wilson & Grant (1997)	Know the job Do the job Citizenship	Various Individual Differences Variables
Caligiuri & Day (2000)	Technical Contextual Expatriate-specific	Not Specified

Note. The table is not meant to be a comprehensive listing of performance models.

Table 2.

A Comparison of Two Predominate Training Evaluation Models

Training Criteria Models	Levels/Factors	Examples from Research
Kirkpatrick Model		
Kirkpatrick (1959b) Kirkpatrick (1959a) Kirkpatrick (1960a) Kirkpatrick (1960b) Kirkpatrick (1967) Kirkpatrick (1979) Kirkpatrick (1996)	1. Reactions 2. Learning 3. Behavior 4. Outcomes	Warr, Allan & Birdi (1999) Alliger & Janak (1989) <u>Expanded or Adapted Model:</u> Talbot (1992) Tannenbaum, Cannon-Bowers, Salas & Mathieu (1993) Alliger, Tannenbaum, Bennett, Traver & Shotland (1997) Swanson & Holton (1999)
Kraiger, Ford and Salas (1993) Model		
Kraiger, Ford & Salas (1993)	1. Cognitive 2. Skill-based 3. Affective	Ford, Smith, Weissbein, Gully & Salas (1998) Colquitt, LaPine & Noe (2000) Kozlowski, Gully, Brown, Salas, Smith & Nason (2001) Tracey, Hinkin, Tannenbaum & Mathieu (2001) Simon & Werner (1996)

Note. The table is not meant to be a comprehensive listing of articles.

Table 3.

An Integration of Training Evaluation and Job Performance Research Using the Kirkpatrick Framework

Kirkpatrick Level	Training Performance	Job Performance
Level One: Reactions		
	Reactions are attitudes toward training and have been categorized as being utility or affective in nature (Alliger et al., 1997). Reactions are not an issue in this discussion.	
Level Two: Learning		
	Traditionally refers to change in knowledge and skill resulting from training (Goldstein, 1993).	The direct determinants in job performance models and research—if defined as knowledge, skill, and a third factor, like motivation—would fit here.
	Kraiger et al. (1993) extend it to changes in cognitive, skill, or affective components.	Hunter (1983)—job knowledge and skill
	Colquitt et al. (2000) offers empirical support for the framework—knowledge, skill, and self-efficacy (affective measure).	Campbell et al. (1993)—declarative knowledge, procedural knowledge and skill, and motivation
	Ford et al. (1998), Kozlowski et al. (2001), Tracey et al. (2001) use the Kraiger et al. (1993) training criteria.	Motowidlo et al. (1997)—task knowledge, task skill, task habits, contextual knowledge, contextual skill, and contextual habits
	Tracey et al. (2001) even employs a limited CFA model and Kozlowski et al. (2001) utilized a path analytic technique.	McCrea & Costa (1995)—although not a job performance theory, direct determinants are viewed as characteristic adaptations
Level Three: Behavior		
	Operationalized as transfer of training and job performance.	Job performance is behavior, not the outcome of behavior, and is multidimensional. All models of the content of performance fit here.
	Colquitt et al. (2000) uses a meta-analytic path analysis to demonstrate relationship between other constructs and transfer of training and job performance.	Campbell et al (1993) Borman & Motowidlo (1993) Grant (1996) Wilson & Grant (1997)
Level Four: Results		
	Organizationally desirable outcomes of learning and behavior. Although Campbell (1999) mentions that job performance can be a determinant of organizational outcomes, this level is not within the scope of this document.	

Note. Examples in the table are illustrative, not exhaustive.

Table 4.

Possible Manifest Indicators from Phase One and Phase Three of the Special Forces Qualifications Course Training

Variable	Format
Phase One	
Small Unit Tactics Exam	Test
Land Navigation Exam	Test
4 Land Navigation Practice Exercises	Ratio
4 Land Navigation Star Exercises	Ratio
Times to Pass Star	Count
Peer Rating: Technical Performance	Rating
Peer Rating: Effort and Persistence	Rating
Peer Rating: Physical Performance	Rating
Peer Rating: Social Interaction Skills	Rating
Peer Rating: Teamwork	Rating
Peer Rating: Leadership Performance	Rating
Positive Spot Reports	Count
Negative Spot Reports	Count
Common Tasks Test (# out of #)	Ratio
Times to Pass SUT Exercise	Count
Phase Three	
Comprehensive Exam	Test
Isolation Exam	Test
Peer Rating: Technical Performance	Rating
Peer Rating: Physical Performance	Rating
Peer Rating: Effort and Persistence	Rating
Peer Rating: Physical Performance	Rating
Peer Rating: Social Interaction Skills	Rating
Peer Rating: Teamwork	Rating
Peer Rating: Leadership Performance	Rating
Cadre Assessment: Technical/Tactical Proficiency	Rating
Cadre Assessment: Decision Making	Rating
Cadre Assessment: Planning	Rating
Cadre Assessment: Use of Available Systems	Rating
Cadre Assessment: Teaching/counseling	Rating
Cadre Assessment: Supervision	Rating
Cadre Assessment: Soldier Team development	Rating
Cadre Assessment: Professional Ethics	Rating
Positive Spot Reports	Count
Negative Spot Reports	Count
Peer Nominations (“Would not want” on team)	Count
Peer Nominations (“Would want” on team)	Count
Peer Evaluation (# out of #): After Isolation	Ranking
Peer Evaluation (# out of #): After Robin Sage	Ranking

Table 5.

Most Likely Special Forces Qualifications Course Phase One and Three Measures by Construct

Construct Measures	Format	Phase
BE Measures		
Phase 1 Peer Ratings:		
Effort and Persistence	Rating	1
Social Interaction Skills	Rating	1
Teamwork	Rating	1
Positive Spot Reports	Count	1
Phase 3 Peer Ratings:		
Effort and Persistence	Rating	3
Social Interaction Skills	Rating	3
Teamwork	Rating	3
Peer Evaluations: Isolation	Ranking	3
Peer Evaluations: Robin Sage	Ranking	3
KNOW Measures		
Small Unit Tactics Exam	Test Score	1
Land Navigation Exam	Test Score	1
Comprehensive Exam	Test Score	3
Isolation Exam	Test Score	3
DO Measures		
Common Tasks Test	Ratio	1
Times to Pass Land Navigation Exercise (STAR)	Count	1
Time to Pass SUT Exercise	Count	1
Peer Ratings: Technical (Phase 1)	Rating	1
Peer Ratings: Technical (Phase 3)	Rating	3
Negative Spot Reports (Phase 1)	Count	1
Negative Spot Reports (Phase 3)	Count	3

Table 6.

Manifest Indicators Utilized in Models by Research Question and Construct

Latent Variables	Manifest Indicators by Research Question		
	One and Four	Two	Three
BE	Peer Ratings: Effort (ISO) Peer Ratings: Team (Ph1) Peer Rankings (ISO) Peer Rankings (RS)	Peer Ratings: Effort (Ph1) Peer Ratings: Effort (RS) Peer Ratings: Team (Ph1) Peer Ratings: Team (RS) Peer Ratings: Social (Ph1) Peer Ratings: Social (RS)	Peer Ratings: Effort (Ph1) Peer Ratings: Effort (ISO) Peer Ratings: Effort (RS) Peer Ratings: Team (Ph1) Peer Ratings: Team (ISO) Peer Ratings: Team (RS) Peer Ratings: Social (Ph1) Peer Ratings: Social (ISO) Peer Ratings: Social (RS) Peer Rankings (ISO) Peer Rankings (RS) Negative Nominations (ISO) Negative Nominations (RS) Negative Spot Reports (Ph1) Negative Spot Reports (Ph3)
KNOW	Land Navigation Exam (Ph1) SUT Exam (Ph1) Isolation Exam (ISO)	Land Navigation Exam (Ph1) SUT Exam (Ph1) Isolation Exam (ISO) Comprehensive Exam (Ph3)	
DO	Negative Spot Reports (Ph1) Negative Spot Reports (Ph3) STAR FTX (Ph1) SUT FTX (Ph1)	Negative Spot Reports (Ph1) Negative Spot Reports (Ph3) STAR FTX (Ph1) SUT FTX (Ph1) Peer Ratings: Tactical (RS) Peer Ratings: Leadership (RS)	

Note. The variables in the table were selected from the SFQC database to operationalize the constructs based on definition, potential sample size, and data quality. Ph1 refers to data collected in Phase One of SFQC. ISO indicates data collected during the Isolation exercise. RS indicates data collected during the Robin Sage field training exercise (FTX). Ph3 refers to data collected in Phase Three. All FTX measures are times to criterion (i.e., times to successful complete exercise).

Table 7.

Descriptive Statistics for Manifest Indicators Used in the Research Models

Variables	*n*	Min	Max	*M*	*SD*	Skewness	
						Statistic	*SE*
Peer Ratings: Effort (Ph1)	1259	1.33	5.00	3.8738	.5636	-.804	.069
Peer Ratings: Effort (ISO)	1303	1.9	5.00	3.991	.412	-.850	.068
Peer Ratings: Effort (RS)	993	1.42	4.93	3.9516	.4732	-1.054	.078
Peer Ratings: Social (Ph1)	1259	1.6	5.00	3.695	.542	-.718	.069
Peer Ratings: Social (ISO)	1303	1.93	4.92	3.7852	.4382	-.853	.068
Peer Ratings: Social (RS)	993	1.58	5.00	3.7774	.4988	-.776	.078
Peer Ratings: Team (Ph1)	1259	1.57	5.00	4.0060	.5314	-1.049	.069
Peer Ratings: Team (ISO)	1303	2.00	4.92	4.0481	.4034	-1.110	.068
Peer Ratings: Team (RS)	993	1.25	5.00	4.0275	.4759	-1.276	.078
Peer Rankings (RS)	1392	1	18	8.08	4.14	.089	.066
Peer Rankings (ISO)	1306	1	18	8.35	4.22	.022	.068
Negative Nominations (RS)	1424	0	13	1.30	2.19	2.490	.065
Negative Nominations (ISO)	1408	0	13	.98	1.92	2.963	.065
Negative Spot Reports (Ph1)	1354	0	13	1.82	2.10	1.701	.066
Negative Spot Reports (Ph3)	1417	0	12	.98	1.42	2.096	.065
Land Navigation Exam (Ph1)	1221	30	100	77.04	14.53	-.605	.070
SUT Exam (Ph1)	1321	60	100	87.23	8.43	-.578	.067
Isolation Exam (Ph3)	1429	18	50	39.62	6.39	-.696	.065
Comprehensive Exam (Ph3)	1158	65	100	86.74	6.26	-.376	.072
STAR FTX (Ph1)	1365	1	5	1.58	.92	1.655	.066
SUT FTX (Ph1)	1363	1	5	1.79	1.02	1.271	.066
Peer Ratings: Tactical (RS)	993	1.50	5.00	3.9065	.4985	-.739	.078
Peer Ratings: Leadership (RS)	993	1.33	5.00	3.8044	.5320	-.650	.078

Note. Statistics are presented for the entire database prior to transformation and elimination of incomplete cases.

Table 8.

Recoded and/or Transformed Variables Used in the Study

Variable	Recoding	Transformation	Statistics		
			N	*M*	*SD*
STAR FTX (Ph1)	Reversed	-	1365	4.42	0.92
SUT FTX (Ph1)	Reversed	-	1363	4.21	1.02
Peer Rankings (ISO)	Reversed	Arcsine	1306	0.66	0.40
Peer Rankings (RS)	Reversed	Arcsine	1392	0.68	0.42
Negative Nominations (ISO)	Reversed	SQRT (X + 1)	1408	1.31	0.52
Negative Nominations (RS)	Reversed	SQRT (X + 1)	1424	1.40	0.57
Negative Spot Reports (Ph1)	Reversed	SQRT (X + 1)	1354	3.47	0.34
Negative Spot Reports (Ph3)	Reversed	SQRT (X + 1)	1417	3.60	0.21

Note. Reversed refers to reflecting the data so larger values indicate higher levels of training performance. Howell (1992) indicates the arcsine transformation is helpful for dealing with proportions (i.e., soldier's rank out of number on the team). Additionally, Howell (1992) recommends the square root (SQRT) as a transformation for data in the form of counts to stabilize the variance and decrease skewness. Descriptive statistics presented are after the transformation. Ph1 refers to data collected in Phase One of SFQC. ISO indicates data collected during the Isolation exercise. RS indicates data collected during the Robin Sage field training exercise. Ph3 refers to data collected in Phase Three.

Table 9.

Demographic Composition of Samples for Questions One and Four

Demographic Variables	Samples for Question One						
	Overall Question One	Sample 1	Sample 2	Sample 3	Sample 4	Sample A	Sample B
N	822	205	205	205	207	410	412
SF MOS							
18A	163 19.83%	50 24.39%	46 22.44%	35 17.07%	32 15.46%	96 23.41%	67 16.26%
18B	205 24.94%	52 25.37%	48 23.41%	50 24.39%	55 26.57%	100 24.39%	105 25.49%
18C	181 22.02%	36 17.56%	42 20.49%	50 24.39%	53 25.60%	78 19.02%	103 25.00%
18D	111 13.50%	22 10.73%	30 14.63%	30 14.63%	29 14.01%	52 12.68%	59 14.32%
18E	162 19.71%	45 21.95%	39 19.02%	40 19.51%	38 18.36%	84 20.49%	78 18.93%
Year Attended							
1997	62 7.54%	17 8.29%	14 6.83%	15 7.32%	16 7.73%	31 7.56%	31 7.52%
1998	346 42.09%	95 46.34%	77 37.56%	80 39.02%	94 45.41%	172 41.95%	174 42.23%
1999	367 44.65%	82 40.00%	101 49.27%	97 47.32%	87 42.03%	183 44.63%	184 44.66%
2000	47 5.72%	11 5.37%	13 6.34%	13 6.34%	10 4.83%	24 5.85%	23 5.58%
Class Attended							
1	203 24.70%	55 26.83%	59 28.78%	49 23.90%	40 19.32%	114 27.80%	89 21.60%
2	173 21.05%	31 15.12%	45 21.95%	41 20.00%	56 27.05%	76 18.54%	97 23.54%
3	192 23.36%	56 27.32%	39 19.02%	47 22.93%	50 24.15%	95 23.17%	97 23.54%
4	254 30.90%	63 30.73%	62 30.24%	68 33.17%	61 29.47%	125 30.49%	129 31.31%

Note. Percentages many not equal 100% due to rounding. Samples A and B are utilized in a Post Hoc analysis presented in Appendix C.

Table 10.

Demographic Composition of Samples for Question Two

Demographic Variables	Samples for Question Two		
	Overall Question Two	Sample 1	Sample 2
N	558	279	279
SF MOS			
18A	38 6.81%	23 8.24%	15 5.38%
18B	158 28.32%	80 28.67%	78 27.96%
18C	144 25.81%	74 26.52%	70 25.09%
18D	84 15.05%	40 14.34%	44 15.77%
18E	134 24.01%	62 22.22%	72 25.81%
Year Attended			
1997	60 10.75%	28 10.04%	32 11.47%
1998	184 32.97%	92 32.97%	92 32.97%
1999	274 49.10%	139 49.82%	135 48.39%
2000	40 7.17%	20 7.17%	20 7.17%
Class Attended			
1	154 27.60%	78 27.96%	76 27.24%
2	152 27.24%	78 27.96%	74 26.52%
3	188 33.69%	91 32.62%	97 34.77%
4	64 11.47%	32 11.47%	32 11.47%

Note. Percentages many not equal 100% due to rounding.

Table 11.

Demographic Composition of Samples for Question Three

Demographic Variables	Samples for Question Three		
	Overall Question Three	Sample 1	Sample 2
N	685	340	345
SF MOS			
18A	122 17.81%	54 15.88%	68 19.71%
18B	179 26.13%	96 28.24%	83 24.06%
18C	162 23.65%	89 26.18%	73 21.16%
18D	87 12.70%	37 10.88%	50 14.49%
18E	135 19.71%	64 18.82%	71 20.58%
Year Attended			
1997	75 10.95%	36 10.59%	39 11.30%
1998	275 40.15%	136 40.00%	139 40.29%
1999	297 43.36%	151 44.41%	146 42.32%
2000	38 5.55%	17 5.00%	21 6.09%
Class Attended			
1	218 31.82%	103 30.29%	115 33.33%
2	198 28.91%	109 32.06%	89 25.80%
3	194 28.32%	92 27.06%	102 29.57%
4	75 10.95%	36 10.59%	39 11.30%

Note. Percentages many not equal 100% due to rounding.

Table 12.

Fit Indices for Research Question One

Model	*N*	X^2	*df*	X^2/df	ΔX^2	GFI	CFI	NNFI	NFI	SRMSR	RMSEA
Section 1 Model A: BE, KNOW, DO											
Sample 1	205										
Initial Model		76.70 (.0006)	41	1.87		.94	.90	.86	.81	.06	.07 (.04-.09)
Final Model		60.16 (.0019)	32	1.88	16.54	.95	.92	.88	.85	.05	.07 (.04-.09)
Sample 2	205	47.73 (.036)	32	1.49		.96	.96	.94	.90	.05	.05 (.01-.08)
Sample 3	205	64.65 (.0006)	32	2.02		.94	.93	.90	.87	.06	.07 (.05-.10)
Sample 4	207	56.07 (.005)	32	1.75		.95	.93	.90	.86	.06	.06 (.03-.09)
Section 2 Model B: BE, KNOW, DO with Method Factors											
Sample 1	205										
Initial Model		61.79 (.0006)	30	2.06		.95	.91	.83	.85	.05	.07 (.05-.10)
Final Model		45.42 (.0035)	23	1.98	16.37	.96	.95	.87	.88	.05	.07 (.04-.10)
Sample 2	205	37.83 (.03)	23	1.65		.97	.96	.93	.92	.04	.06 (.02-.09)
Sample 3	205	49.27 (.0011)	23	2.14		.96	.94	.89	.90	.05	.08 (.05-.10)
Sample 4	207	45.18 (.0038)	23	1.96		.96	.94	.87	.88	.05	.07 (.04-.10)

Table 12 (continued).

Model	N	X^2	*df*	X^2/df	ΔX^2	GFI	CFI	NNFI	NFI	SRMSR	RMSEA
Section 3 Model C: Campbell Version of BE, KNOW, DO											
Sample 1	205										
Initial Model		43.50 (.09)	32	1.36		.96	.97	.94	.89	.05	.04 (.00-.07)
Final Model		33.57 (.09)	24	1.40	9.93	.97	.97	.95	.91	.04	.04 (.00-.08)
Sample 2	205	20.25 (.68)	24	.84		.98	1.00	1.02	.96	.03	.00 (.00-.03)
Sample 3	205	42.85 (.01)	24	1.79		.96	.96	.92	.91	.05	.06 (.03-.09)
Sample 4	207	31.64 (.14)	24	1.32		.97	.98	.96	.92	.05	.04 (.00-.07)
Section 4 Model D: BE, KNOW, DO with General Soldiering											
Sample 1	205										
Initial Model		46.93 (.03)	30	1.56		.96	.95	.91	.88	.04	.05 (.00-.06)
Final Model		35.01 (.04)	22	1.59	11.93	.97	.96	.92	.91	.04	.05 (.01-.09)
Sample 2	205	21.63 (.48)	22	.98		.98	1.00	1.00	.95	.03	.00 (.00-.06)
Sample 3	205	35.46 (.04)	22	1.61		.97	.97	.94	.93	.04	.06 (.02-.09)
Sample 4	207	27.66 (.19)	22	1.26		.98	.98	.97	.93	.04	.04 (.00-.07)

Note. GFI = goodness-of-fit index; CFI = comparative fit index; NNFI = non-normed fit index; NFI = normed fit index; SRMSR = standardized root mean square residual; RMSEA = root mean square error of approximation estimate 90% confidence interval shown in parentheses. The *p* values for the X^2 (but not the ΔX^2) are reported in parentheses below the statistic's value. The ΔX^2 were not significant ($p > .05$). Diagrams for the models in the table can be found in Figures 3 through 6 and Figures 12 through 16.

Table 13.

Standardized Parameter Estimates for the BE KNOW DO Model

Manifest Indicators by Latent Variable	Standardized Parameter Estimates for Each Sample				
	Sample 1 Initial Model	Sample 1 Final Model	Sample 2 Final Model	Sample 3 Final Model	Sample 4 Final Model
BE					
Effort (ISO)	.72	.72	.72	.69	.65
Teamwork (Ph1)	.54	.53	.58	.55	.56
Isolation Rankings	.89	.89	.91	.93	.88
Robin Sage Rankings	.68	.68	.75	.78	.67
KNOW					
Land Navigation Exam	.22	.23	.21	.33	.32
SUT Exam	.52	.52	.61	.73	.47
Isolation Exam	.37	.36	.34	.19	.36
DO					
Negative Spots (Ph1)	.46	.41	.38	.21	.25
Negative Spots (Ph3)	.07*	-	-	-	-
STAR FTX	.36	.35	.51	.47	.39
SUT FTX	.30	.31	.65	.36	.94
Correlations					
BE and KNOW	.57	.57	.55	.56	.56
BE and DO	.43	.42	.32	.47	.36
KNOW and DO	.92	.91	.46	.91	.26

Note. The standardized parameter estimates are for Model A. See Figures 3 and 12 for a diagram of the models.

*$p > .05$ (not significant); all other parameters are significant at $p < .05$.

Table 14.

Standardized Parameter Estimates for BE KNOW DO Model with Methods

Manifest Indicators	Standardized Parameter Estimates for Each Sample									
	Sample 1 Initial Model		Sample 1 Final Model		Sample 2 Final Model		Sample 3 Final Model		Sample 4 Final Model	
BE	Factor	Method	Factor	Method	Factor	Method	Factor	Method	Factor	Method
Effort (ISO)	.83	-.03* RT	.82	.04* RT	.84	.32 RT	.76	.05* RT	.75	-.27 RT
Teamwork (Ph1)	.60	.59 RT	.59	-.26 RT	.72	-.47 RT	.61	.67 RT	.69	.55 RT
Isolation Rankings	.76	-.40 RK	.77	.43* RK	.75	.47 RK	.83	.47 RK	.73	.33 RK
Robin Sage Rankings	.54	-.56 RK	.54	.53 RK	.61	.52 RK	.66	.42 RK	.54	.62 RK
KNOW										
Land Navigation Exam	.29	-.30 T	-.32	.78 T	.13*	.44 T	.20	.17 T	.30	.21* T
SUT Exam	.50	.12* T	-.54	-.08* T	.53	.20* T	.46	.88 T	.54	.26* T
Isolation Exam	.37	.25* T	-.38	-.11 T	.38	-.02* T	.15	.08* T	.62	-.69 T
DO										
Negative Spots (Ph1)	.42	-.04* SP	.43	-	.36	-	-.23	-	.29	-
Negative Spots (Ph3)	.08*	.91 SP	-	-	-	-	-	-	-	-
STAR FTX	.37	.91 TC	.36	.05 TC	.55	-.82 TC	-.60	-.23* TC	.35	.46 TC
SUT FTX	.33	-.04* TC	.33	-.83 TC	.72	.08* TC	-.52	.74 TC	.81	.18* TC
Correlations										
BE-KNOW	.60		-.56		.68		.99*		.43	
BE-DO	.49		.48		.32		-.43		.49	
KNOW-DO	.89		-.83		.47		-1.07*		.24	

Note. RT = Ratings; RK = Rankings; T = Tests; SP = Spot Reports; and TC = Times to Criterion. The standardized parameter estimates are for Model B (see Figures 4, 13, and 14). *$p > .05$ (not significant); all other parameters are significant at $p < .05$.

Table 15.

Standardized Parameter Estimates for Campbell Version of BE KNOW DO

Manifest Indicators by Latent Variable	Standardized Parameter Estimates for Each Sample														
	Sample 1 Initial Model			Sample 1 Final Model			Sample 2 Final Model			Sample 3 Final Model			Sample 4 Final Model		
BE	B	K	D	B	K	D	B	K	D	B	K	D	B	K	D
Effort (ISO)	.39*	.45	-.60*	.23*	.46	-.47	-.55	.47	-.02*	-.04*	.49	-.62	.55	.27	.22
Teamwork (Ph1)	.21*	.43	-.34*	9.4	.36	3.88*	-.30	.58	.05*	-.03*	.53	-.36	.32	.40	.33
Isolation Rankings	.30*	.47	-.51*	.33*	.48	-.72	-.76	.50	.006*	-.07	.55	-.60	.73	.48	.17*
Robin Sage Rankings	-7.32*	.17*	-5.36*	.29*	.19*	-.63	-69	.35	.007*	-8.3	.36	-.02*	.67	.18*	.16*
KNOW															
Land Navigation Exam	-	.26	-	-	.25	-	-	.20	-	-	.35	-	-	.32	-
SUT Exam	-	.52	-	-	.53	-	-	.57	-	-	.64	-	-	.52	-
Isolation Exam	-	.32	-	-	.35	-	-	.36	-	-	.20	-	-	.36	-
DO															
Negative Spots (Ph1)	-	.36	-.02*	-	.34	-.004	-	.20	.08*	-	.23	.13	-	.10*	.24
Negative Spots (Ph3)	-	.05*	-.03*	-	-	-	-	-	-	-	-	-	-	-	-
STAR FTX	-	.35	.01*	-	.27	.04*	-	.34	.09*	-	.46	.10*	-	.25	.36
SUT FTX	-	.30	.003*	-	.36	.009*	-	.27	2.69*	-	.38	.04*	-	.23	.86

Note. B = BE; K = KNOW; and D = DO. The standardized parameter estimates are for Model C. See Figures 5 and 15 for diagrams.
*$p > .05$ (not significant); all other parameter estimates are significant at $p < .05$.

Table 16.

Standardized Parameter Estimates for BE KNOW DO with Unitary Content Factor

Manifest Indicators	Standardized Parameter Estimates									
	Sample 1 Initial Model		Sample 1 Final Model		Sample 2 Final Model		Sample 3 Final Model		Sample 4 Final Model	
BE	Factor	Content	Factor	Content	Factor	Content	Factor	Content	Factor	Content
Effort (ISO)	-.56	.47	.54	.49	.22*	.69	.58	.38	.63	.14*
Teamwork (Ph1)	-.37	.48	.31	.55	.02*	.72	.37	.49	.47	.43
Isolation Rankings	-.77	.41	.77	.43	.24*	.88	.85	.28	.87	.17*
Robin Sage Rankings	-.77	.09*	.73	.17*	.23*	.72	.74	.29	.67	.07*
KNOW										
Land Navigation Exam	.004*	.32	-.02*	.30	.0004*	.24	-.13*	.43	.28	.16*
SUT Exam	-.17*	.45	-.22	.42	.0008*	.62	.23*	.59	.45	.20
Isolation Exam	-.98*	.11*	-.87	.05*	.0005*	.35	.18*	.14*	.32	.14*
DO										
Negative Spots (Ph1)	.09*	.33	.09*	.33	.24	.25	.35	.30	.23*	.15*
Negative Spots (Ph3)	.23*	-.08*	-	-	-	-	-	-	-	-
STAR FTX	.02*	.32	.23*	.27	.19*	.36	-.25*	.45	-.15*	.72
SUT FTX	.28*	.26	.04*	.33	.67	.54	.23*	.43	.62*	.63
Correlations										
BE-KNOW	.13*		-.19*		-1265		.74*		.52	
BE-DO	-.34*		.06*		-1.96		-.08		.32*	
KNOW-DO	-.76*		-1.17*		-331.03*		-1.12*		-.05*	

Note. The standardized parameter estimates are for Model D. See Figures 6 and 16 for diagrams.
*$p > .05$ (not significant); all other parameter estimates are significant at $p < .05$.

Table 17.

Fit Indices for Research Question Two

Model	*N*	X^2	*df*	X^2/df	ΔX^2	GFI	CFI	NNFI	NFI	SRMSR	RMSEA
Section 1 Phase 1 Models											
Sample 1	279										
Initial Model		52.79	17	3.11		.96	.97	.94	.95	.05	.09 (.06-.11)
Final Model		33.07 (.0005)	11	3.01	19.72**	.97	.98	.96	.97	.05	.09 (.05-.12)
Sample 2	279										
Initial Model		31.26 (.0186)	17	1.84		.97	.99	.98	.97	.05	.06 (.02-.09)
Final Model		9.37 (.5877)	11	.85	21.89**	.99	1.00	1.00	.99	.03	.00 (.00-.06)
Section 2 Phase 3 Models											
Sample 1	279										
Initial Model		46.98	17	2.76		.96	.98	.97	.97	.05	.08 (.05-.11)
Final Model		27.43 (.004)	11	2.49	19.55**	.97	.99	.98	.98	.04	.07 (.04-.11)
Sample 2	279										
Initial Model		52.09	17	3.06		.96	.98	.97	.97	.05	.09 (.06-.11)
Final Model		35.29 (.0002)	11	3.21	16.80*	.97	.99	.98	.98	.04	.09 (.06-.12)

Table 17 (continued).

Model	*N*	X^2	*df*	X^2/df	ΔX^2	GFI	CFI	NNFI	NFI	SRMSR	RMSEA
Section 3 Predictive Models											
Sample 1	279										
Initial Model		491.30	97	5.07		.86	.86	.83	.84	.15	.12 (.11-.13)
Intermediate Model		428.84	71	6.04	62.46	.86	.87	.84	.85	.16	.13 (.12-.15)
Final Model		143.84	70	2.06	347.44	.93	.97	.97	.95	.06	.06 (.05-.08)
Sample 2	279										
Initial Model		514.09	97	5.30		.86	.86	.83	.84	.16	.12 (.12-.14)
Intermediate Model		449.02	71	6.32	65.07	.87	.87	.84	.85	.17	.14 (.13-.15)
Final Model		112.40 (.001)	70	1.61	401.69	.95	.99	.98	.96	.05	.05 (.03-.06)

Note. GFI = goodness-of-fit index; CFI = comparative fit index; NNFI = non-normed fit index; NFI = normed fit index; SRMSR = standardized root mean square residual; RMSEA = root mean square error of approximation estimate 90% confidence interval shown in parentheses. See Figures 7 through 9 and 17 through 24 for conceptual diagrams. Unless otherwise noted in parentheses or with asterisks, all X^2 values and ΔX^2 values in the table are significant at $p < .0001$. Values for the ΔX^2 are based on the initial model for each sample.
*$p < .05$; **$p < .01$.

Table 18.

Standardized Parameter Estimates for the Phase One BE KNOW DO Model

Manifest Indicators	Standardized Parameter Estimates			
	Sample 1 Initial Model	Sample 1 Final Model	Sample 2 Initial Model	Sample 2 Final Model
BE				
Peer Ratings: Effort	.93	.93	.95	.95
Peer Ratings: Teamwork	.99	.99	.98	.98
Peer Ratings: Social	.85	.85	.86	.86
KNOW				
Land Navigation Exam	.42	.35	.32	.31
SUT Exam	.36	.43	.65	.67
DO				
Negative Spot Reports	.48	-	-.15	-
STAR FTX	.39	.40	-.57	.57
SUT FTX	.61	.56	-.50	.47
Correlations				
BE and KNOW	.40	.40	.43	.43
BE and DO	.54	.60	-.51	.54
KNOW and DO	.84	.84	-.67	.65

Note. Standardized parameter estimates for Phase One training performance are reported. All parameter estimates greater than .15 are significant at $p < .05$.

Table 19.

Standardized Parameter Estimates for the Phase Three BE KNOW DO Model

Manifest Indicators	Standardized Parameter Estimates			
	Sample 1 Initial Model	Sample 1 Final Model	Sample 2 Initial Model	Sample 2 Final Model
BE				
Peer Ratings: Effort	.95	.95	.95	.95
Peer Ratings: Teamwork	.97	.97	.97	.97
Peer Ratings: Social	.87	.87	.87	.88
KNOW				
Comprehensive Exam	.32	.32	.58	.58
Isolation Exam	.36	.36	.45	.45
DO				
Negative Spot Reports	.10*	-	.25	-
Peer Ratings: Tactical	.92	.91	.93	.93
Peer Ratings: Leadership	1.00	1.00	1.00	1.00
Correlations				
BE and KNOW	.23*	.23*	.16*	.16*
BE and DO	.83	.82	.87	.87
KNOW and DO	.53	.53	.33	.33

Note. Standardized parameter estimates for Phase Three training performance are reported. *$p > .05$ (not significant); all other parameter estimates are significant at $p < .05$.

Table 20.

Fit Indices for Research Question Three

Model	N	X^2	*df*	X^2/df	ΔX^2	GFI	CFI	NNFI	NFI	SRMSR	RMSEA
One Factor											
Sample 1	340										
Initial Model		2293.41	90	25.48		.52	.50	.42	.49	.13	.27 (.26-.28)
Second Model		906.69	44	20.61	1386.72	.64	.62	.53	.61	.17	.24 (.23-.25)
Sample 2	345										
Initial Model		2133.35	90	23.71		.51	.51	.41	.50	.13	.26 (.25-.27)
Second Model		979.39	44	22.26	1153.96	.61	.60	.50	.59	.19	.25 (.24-.26)
Three Factor											
Sample 1	340										
Initial Model		2266.40	87	26.05		.51	.51	.41	.50	.13	.27 (.26-.28)
Second Model		841.53	41	20.53	1424.87	.68	.65	.53	.64	.17	.24 (.23-.25)
Sample 2	345										
Initial Model		2109.63	87	24.25		.53	.52	.42	.51	.13	.26 (.25-.27)
Second Model		914.54	41	22.31	1195.09	.65	.62	.49	.61	.19	.25 (.24-.26)

Note. See Figures 10, 11, 25, and 26 for the conceptual models of BE. GFI = goodness-of-fit index; CFI = comparative fit index; NNFI = non-normed fit index; NFI = normed fit index; SRMSR = standardized root mean square residual; RMSEA = root mean square error of approximation estimate 90% confidence interval shown in parentheses. All X^2 and ΔX^2 values in the table are significant at $p < .0001$.

Table 21.

Standardized Parameter Estimates for BE Models

Indicators by Construct	BE Models							
	One-Factor BE Models				Three-Factor BE Models			
	Initial Model		Second Model		Initial Model		Second Model	
	Sample 1	Sample 2	Sample 1	Sample 2	Sample 1	Sample 2	Sample 1	Sample 2
Effort (F1)								
Effort (Ph1)	.59	.51	.95	.95	.57	.51	.96	.96
Effort (ISO)	.78	.83	-	-	.75	.83	-	-
Effort (RS)	.84	.75	.42	.36	.86	.76	.43	.35
RS Ranking	.70	.65	.46	.43	.67	.63	.46	.42
ISO Ranking	.65	.65	-	-	.60	.63	-	-
Personal Discipline (F2)								
Negative Spots (Ph1)	.13	.08*	.11	.08*	.12	.04*	.11*	.01*
Negative Spots (Ph3)	.15	.15	.05*	.04*	.21	.14	.23	.15
ISO Negative Nominations	.53	.63	.36	.33	.61	.75	.70	.70
RS Negative Nominations	.61	.51	.32	.30	.74	.62	.64	.67
Team Facilitation (F3)								
Team (Ph1)	.57	.49	.96	.96	.56	.51	.97	.96
Team (ISO)	.78	.88	.46	.43	.76	.87	.44	.43
Team (RS)	.84	.74	-	-	.86	.75	-	-
Social (Ph1)	.53	.50	.87	.88	.51	.51	.87	.88
Social (ISO)	.73	.84	.45	.41	.69	.82	.43	.41
Social (RS)	.83	.74	-	-	.83	.73	-	-
Correlations								
F1-F2	-	-	-	-	.84	.81	.51	.43
F1-F3	-	-	-	-	1.04	1.03	.98	.99
F2-F3					.84	.85	.47	.45

Note. Standardized parameter estimates for both conceptualizations of BE for both sets of indicators.
*$p > .05$ (not significant); all other parameter estimates are significant at $p < .05$.

Table 22.

Fit Indices for Question Four

Model	*N*	X^2	*df*	X^2/df	ΔX^2	GFI	CFI	NNFI	NFI	SRMSR	RMSEA
Section 1 Sample 1 as Initial Sample											
Sample 1 Models	205										
Initial Model		76.70 (.0006)	41	1.87		.94	.90	.86	.81	.06	.07 (.04-.09)
Iteration 1		44.85 (.04)	30	1.50		.96	.96	.94	.88	.05	.05 (.01-.08)
Iteration 2		30.03 (.18)	24	1.25		.97	.98	.97	.92	.04	.04 (.00-.07)
Iteration 3		35.47 (.19)	29	1.22		.97	.98	.97	.91	.04	.03 (.00-.07)
Iteration 4		28.90 (.42)	28	1.03		.97	1.00	1.00	.93	.04	.01 (.00-.06)
Iteration 5		18.02 (.86)	26	.69		.98	1.00	1.04	.95	.04	.00 (.00-.03)
Iteration 6		15.40 (.93)	25	.61		.99	1.00	1.05	.96	.03	.00 (.00-.02)
Iteration 7		12.28 (.98)	24	.51		.99	1.00	1.06	.97	.03	.00 (.00-.00)
Cross-Validation of Iteration 5											
Sample 2	205	28.88 (.32)	26	1.11	-10.86	.97	.99	.99	.94	.04	.02 (.00-.06)
Sample 3	205	42.96 (.02)	26	1.65	-24.94	.96	.96	.94	.91	.05	.06 (.02-.09)
Sample 4	207	45.64 (.01)	26	1.76	-27.62	.96	.94	.90	.88	.05	.06 (.03-.09)

Table 22 (continued).

Model	N	X^2	df	X^2/df	ΔX^2	GFI	CFI	NNFI	NFI	SRMSR	RMSEA
Section 2 Sample 2 as Initial Sample											
Sample 2 Models	205										
Initial Model		57.35 (.05)	41	1.40		.95	.96	.95	.88	.06	.04 (.01-.07)
Iteration 1		31.38 (.35)	29	1.08		.97	.99	.99	.93	.04	.02 (.00-.06)
Iteration 2		24.95 (.63)	28	.89		.98	1.00	1.01	.95	.04	.00 (.00-.05)
Iteration 3		16.77 (.91)	26	.64		.98	1.00	1.04	.96	.03	.00 (.00-.02)
Iteration 4		11.36 (.97)	24	.47		.99	1.00	1.06	.98	.02	.00 (.00-.00)
Iteration 5		13.52 (.96)	24	.56		.99	1.00	1.05	.97	.03	.00 (.00-.00)
Cross-Validation of Iteration 3											
Sample 1	205	31.75 (.20)	26	1.22	-14.98	.97	.98	.97	.92	.04	.03 (.00-.07)
Sample 3	205	37.35 (.07)	26	1.44	-20.58	.97	.98	.96	.93	.05	.05 (.00-.08)
Sample 4	207	46.10 (.01)	26	1.77	-29.33	.96	.94	.90	.88	.05	.06 (.03-.09)

Table 22 (continued).

Model	N	X^2	df	X^2/df	ΔX^2	GFI	CFI	NNFI	NFI	SRMSR	RMSEA
Section 3 Sample 3 as Initial Sample											
Sample 3 Models	205										
Initial Model		78.80 (.0004)	41	1.92		.93	.92	.89	.85	.06	.07 (.04-.09)
Iteration 1		64.79 (.006)	39	1.66		.94	.94	.92	.88	.06	.06 (.03-.08)
Iteration 2		52.76 (.06)	38	1.39		.96	.97	.95	.90	.05	.04 (.00-.07)
Iteration 3		32.07 (.27)	28	1.15		.97	.99	.99	.94	.05	.03 (.00-.06)
Iteration 4		33.11 (.19)	27	1.23		.97	.99	.98	.93	.04	.03 (.00-.07)
Iteration 5		28.32 (.40)	27	1.05		.97	1.00	1.00	.94	.04	.01 (.00-.06)
Iteration 6		41.48 (.04)	27	1.54		.96	.97	.95	.92	.05	.05 (.01-.08)
Cross-Validation of Iteration 3											
Sample 1	205	44.73 (.02)	28	1.60	-12.66	.96	.95	.92	.88	.05	.05 (.02-.08)
Sample 2	205	29.72 (.38)	28	1.06	2.35	.97	1.00	.99	.93	.04	.02 (.00-.06)
Sample 4	207	41.63 (.05)	28	1.49	-9.56	.96	.96	.93	.89	.05	.05 (.01-.08)

Table 22 (continued).

Model	N	X^2	*df*	X^2/df	ΔX^2	GFI	CFI	NNFI	NFI	SRMSR	RMSEA
Section 4 Sample 4 as Initial Sample											
Sample 4 Models	207										
Initial Model		62.66 (.02)	41	1.53		.95	.94	.92	.84	.06	.05 (.02-.08)
Iteration 1		41.72 (.08)	30	1.39		.96	.97	.95	.89	.05	.04 (.00-.07)
Iteration 2		27.16 (.51)	28	.97		.97	1.00	1.00	.93	.04	.00 (.00-.05)
Iteration 3		37.56 (.16)	30	1.25		.97	.98	.97	.90	.05	.04 (.00-.07)
Iteration 4		30.52 (.34)	28	1.09		.97	.99	.99	.92	.04	.02 (.00-.06)
Iteration 5		34.25 (.23)	29	1.18		.97	.98	.98	.91	.05	.03 (.00-.06)
Cross-Validation of Iteration 3											
Sample 1	205	71.79 (<.0001)	30	2.39	-34.23	.94	.88	.82	.82	.08	.08 (.06-.11)
Sample 2	205	50.37 (.0113)	30	1.68	-12.81	.95	.95	.93	.89	.06	.06 (.03-.09)
Sample 3	205	69.90 (<.0001)	30	2.33	-32.34	.94	.91	.87	.86	.08	.08 (.06-.11)

Note. GFI = goodness-of-fit index; CFI = comparative fit index; NNFI = non-normed fit index; NFI = normed fit index; SRMSR = standardized root mean square residual; RMSEA = root mean square error of approximation estimate 90% confidence interval shown in parentheses. Negative ΔX^2 values suggest that the model did not cross-validate well. The *p* values for the X^2 (but not the ΔX^2) are reported in parentheses below the statistic's value.

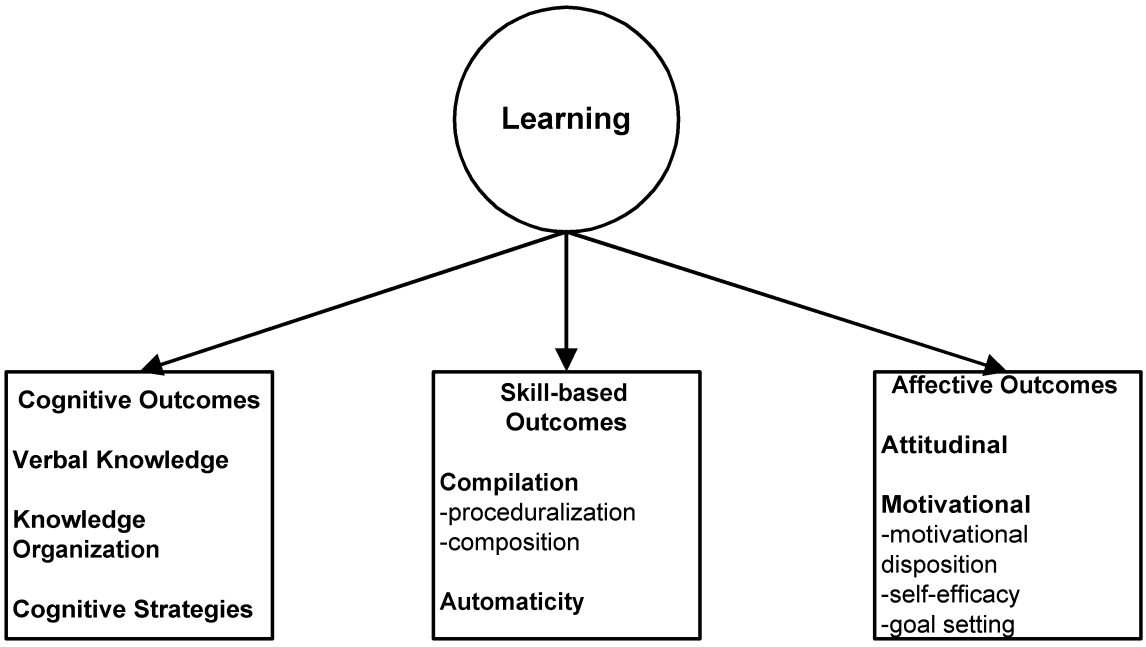


Figure 1. Kraiger, Ford, & Salas (1993) conceptualized three categories of learning criteria for training.

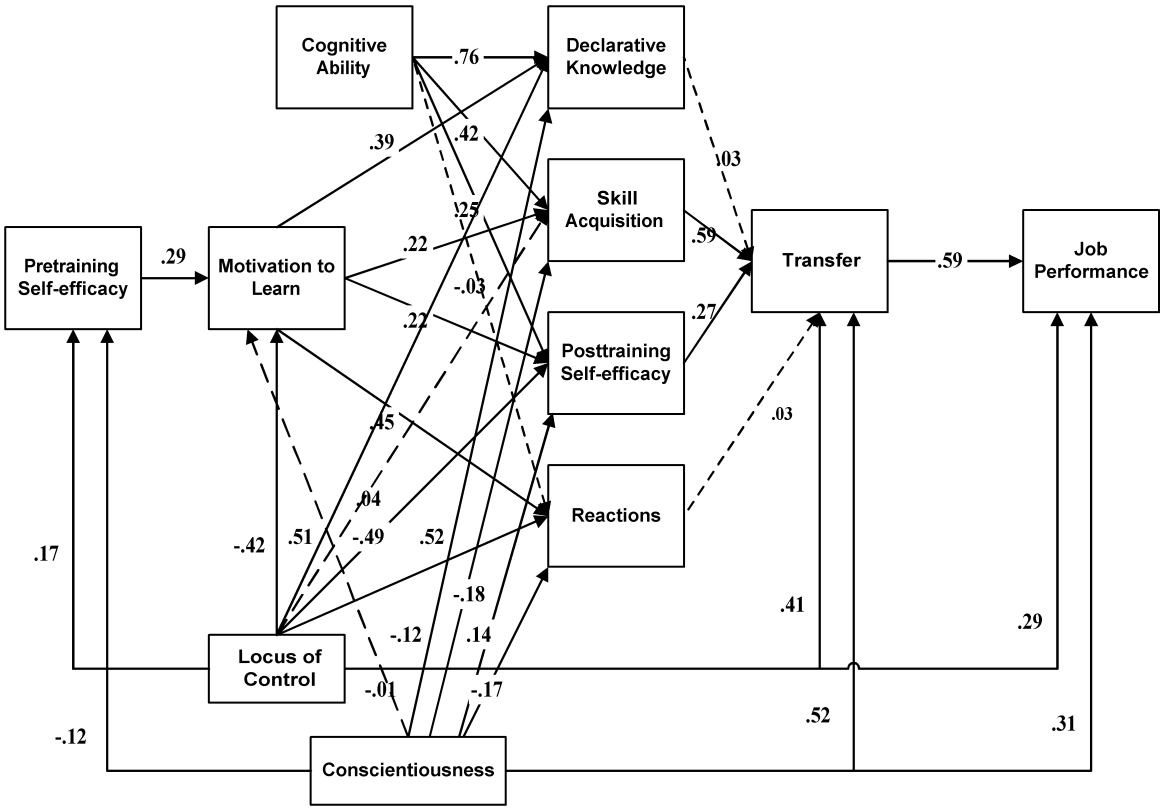


Figure 2. Colquitt, LaPine, and Noe (2000) partially mediated meta-analytic path model. Colquitt et al. (2000) conducted a meta-analytic path analysis that demonstrated that learning outcomes partially mediate human attributes and more distal dependent variables like job performance. The diagram provides a partial view of the partially mediated model presented in Colquitt et al. (2000)—showing just the main paths related to the training outcomes. Reproduced by permission of the first author.

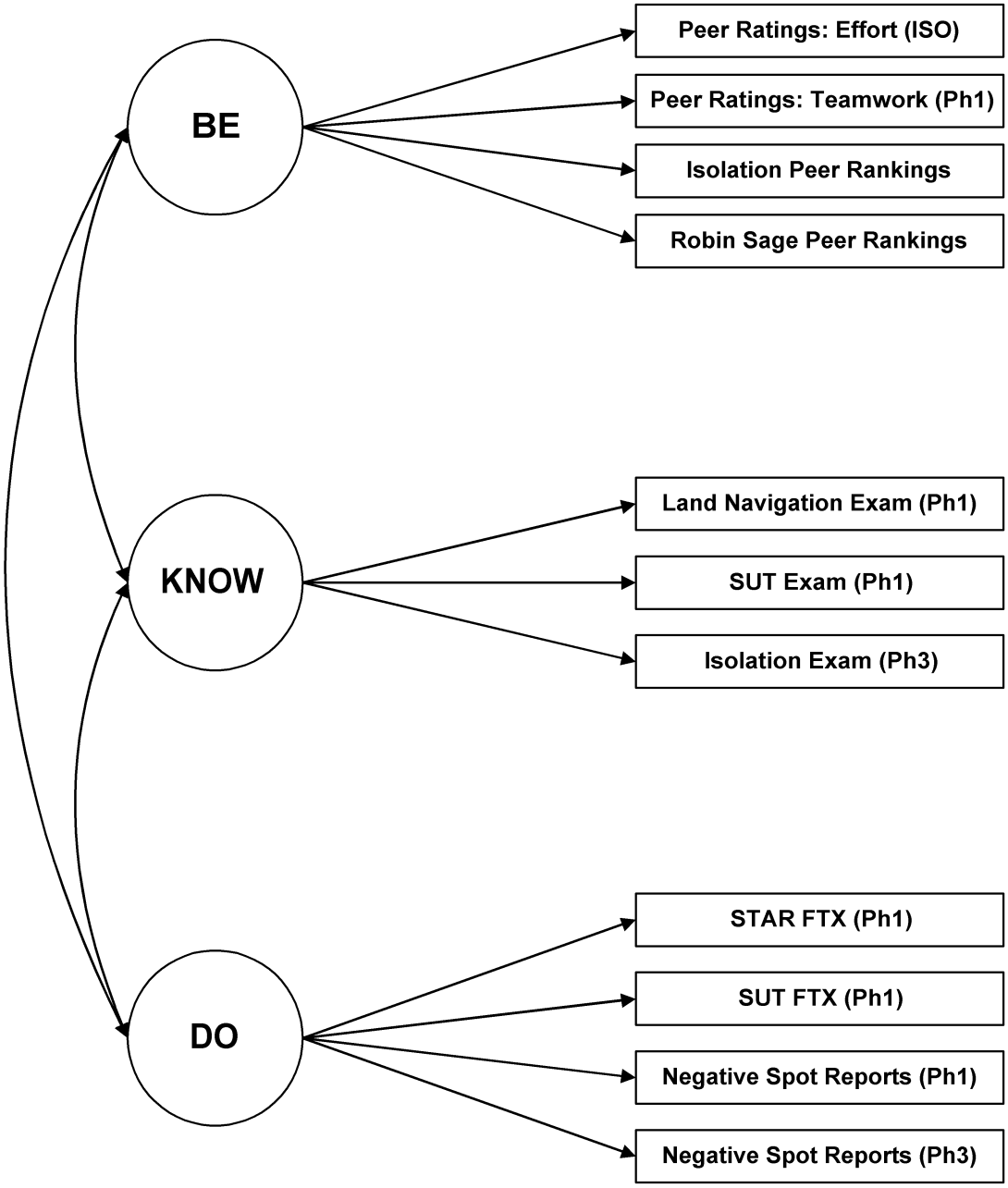


Figure 3. The BE KNOW DO model of training performance operationalized with manifest indicators. Ph1 refers to Phase One of SFQC; Ph3 refers to Phase Three of SFQC; ISO refers to the isolation exercise during Phase Three; SUT refers to small unit tactics; and FTX refers to field training exercise. STAR is the land navigation FTX. Latent constructs are shown in ovals, and manifest variables are shown in rectangles.

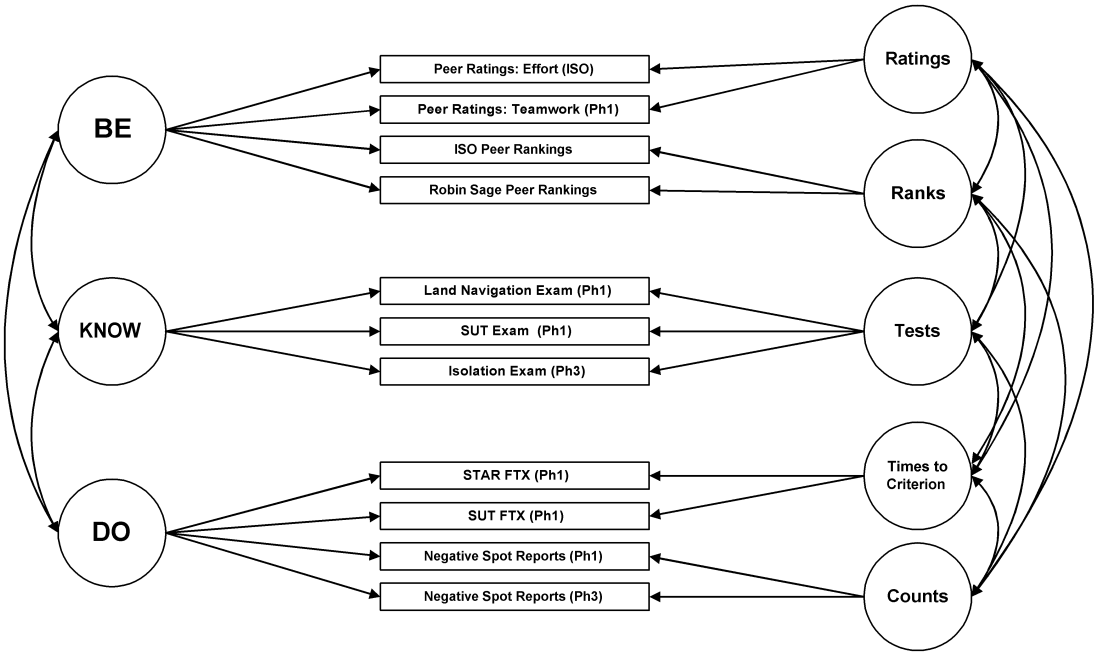


Figure 4. The BE KNOW DO model of training performance with correlated method factors operationalized with manifest indicators. Ph1 refers to Phase One of SFQC; Ph3 refers to Phase Three of SFQC; ISO refers to the isolation exercise during Phase Three; SUT refers to small unit tactics; and FTX refers to field training exercise. STAR is the land navigation FTX. Latent constructs are shown in ovals, and manifest variables are shown in rectangles.

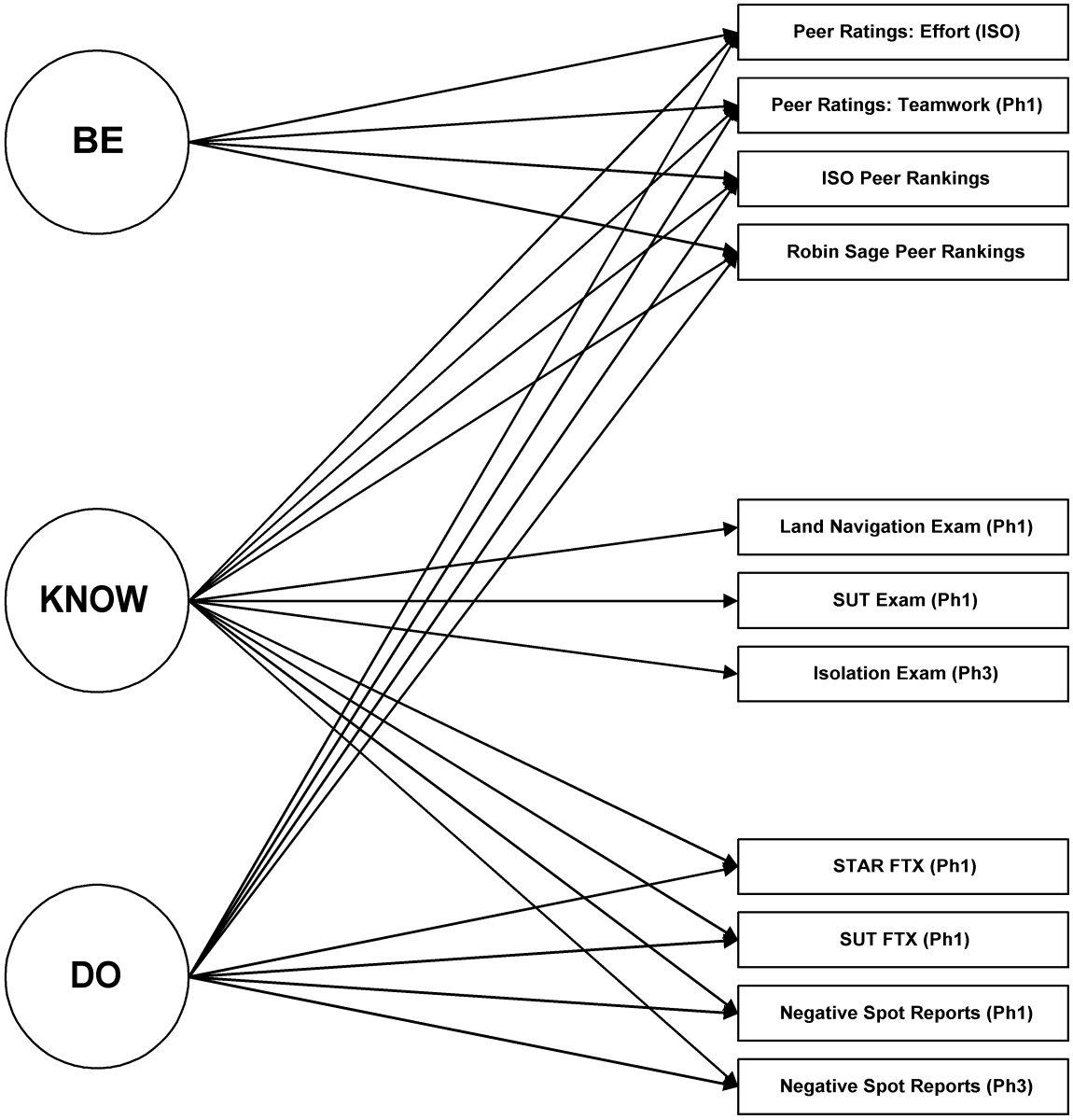


Figure 5. Campbell version of the BE KNOW DO model of training performance operationalized with manifest indicators. Ph1 refers to Phase One of SFQC; Ph3 refers to Phase Three of SFQC; ISO refers to the isolation exercise during Phase Three; SUT refers to small unit tactics; and FTX refers to field training exercise. STAR is the land navigation FTX. Latent constructs are shown in ovals, and manifest variables are shown in rectangles.

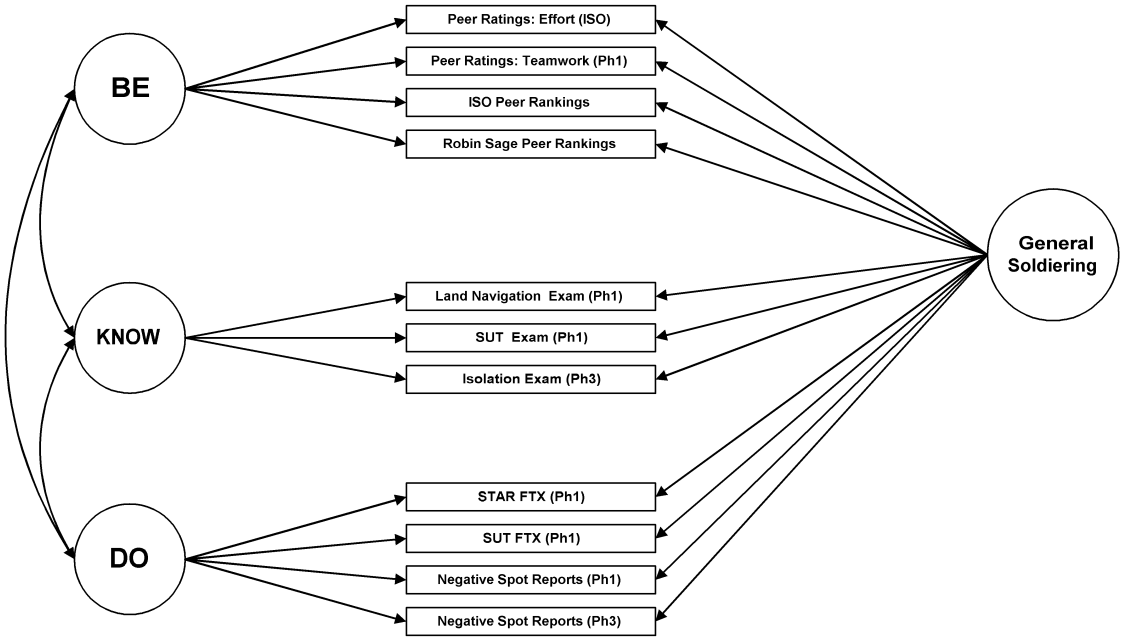


Figure 6. BE KNOW DO model of training performance with a general soldiering factor operationalized with manifest indicators. Ph1 refers to Phase One of SFQC; Ph3 refers to Phase Three of SFQC; ISO refers to the isolation exercise during phase three; SUT refers to small unit tactics; and FTX refers to field training exercise. STAR is the land navigation FTX. Latent constructs are shown in ovals, and manifest variables are shown in rectangles.

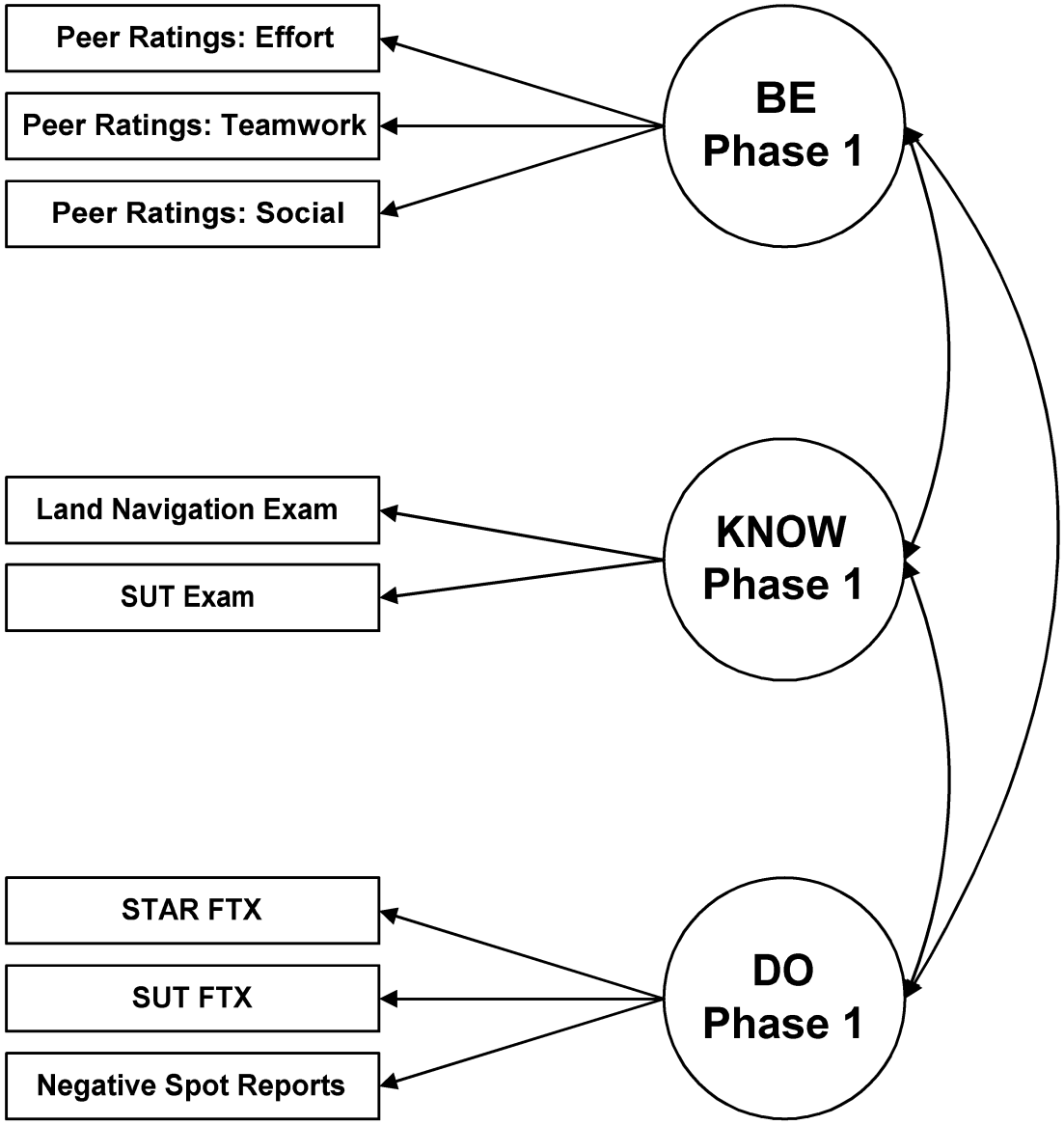


Figure 7. Phase One SFQC training performance model presented with manifest indicators. All measures were captured in Phase One. Latent constructs are shown in ovals, and manifest variables are shown in rectangles.

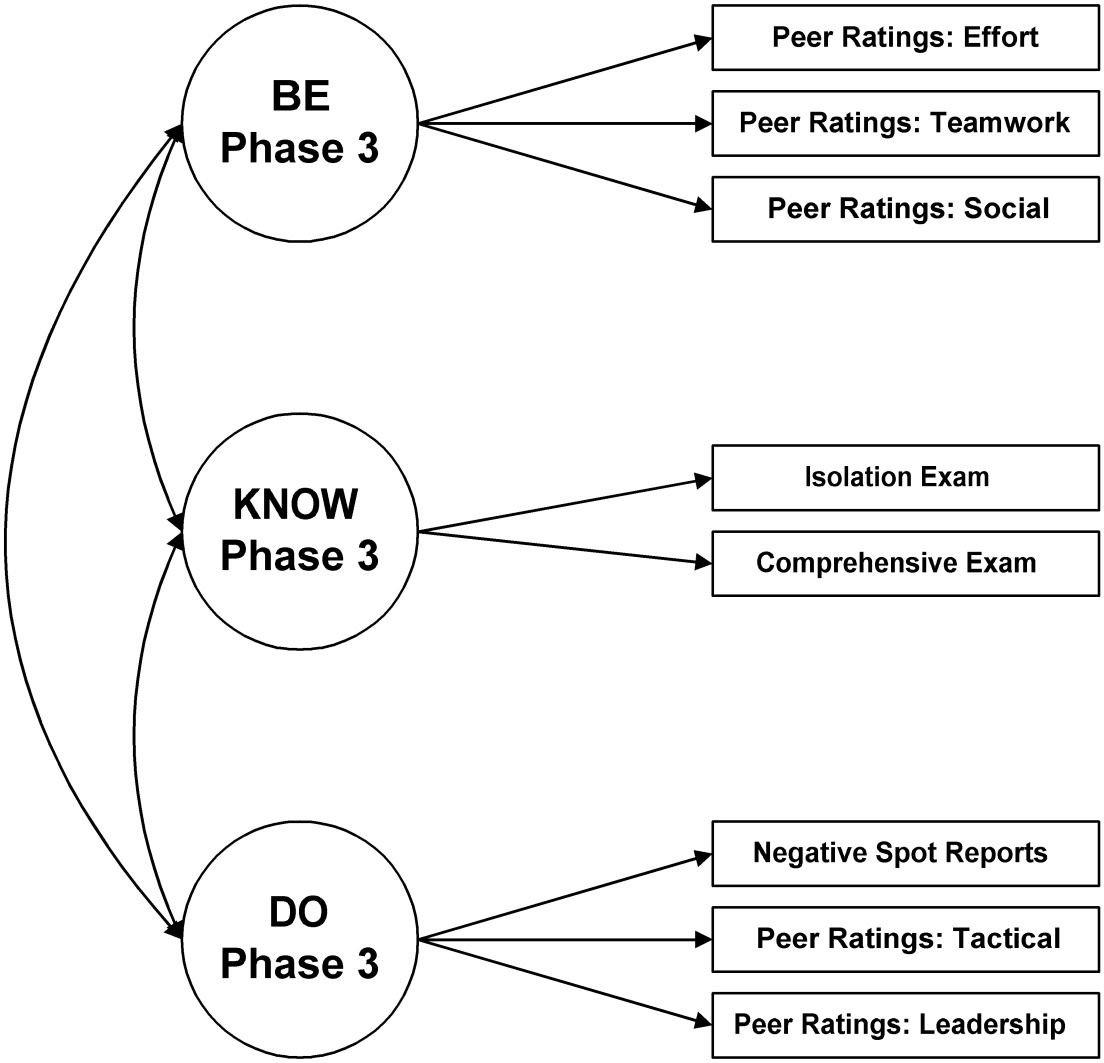


Figure 8. Phase Three SFQC training performance model presented with manifest indicators. All measures were captured in Phase Three. Latent constructs are shown in ovals, and manifest variables are shown in rectangles.

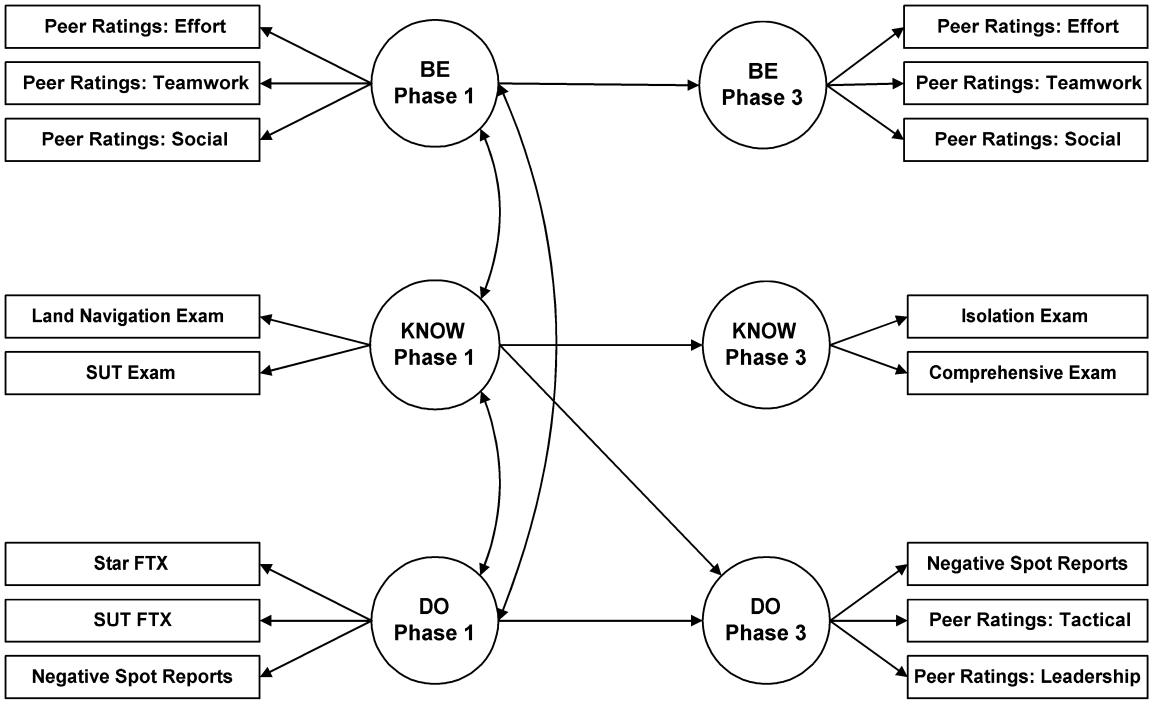


Figure 9. Posited relationship between Phase One and Phase Three training performance presented with manifest indicators. Latent constructs are shown in ovals, and manifest variables are shown in rectangles. Single-headed arrows between latent variables (ovals) indicate predictive relationships; double-headed arrows between latent constructs represent correlation.

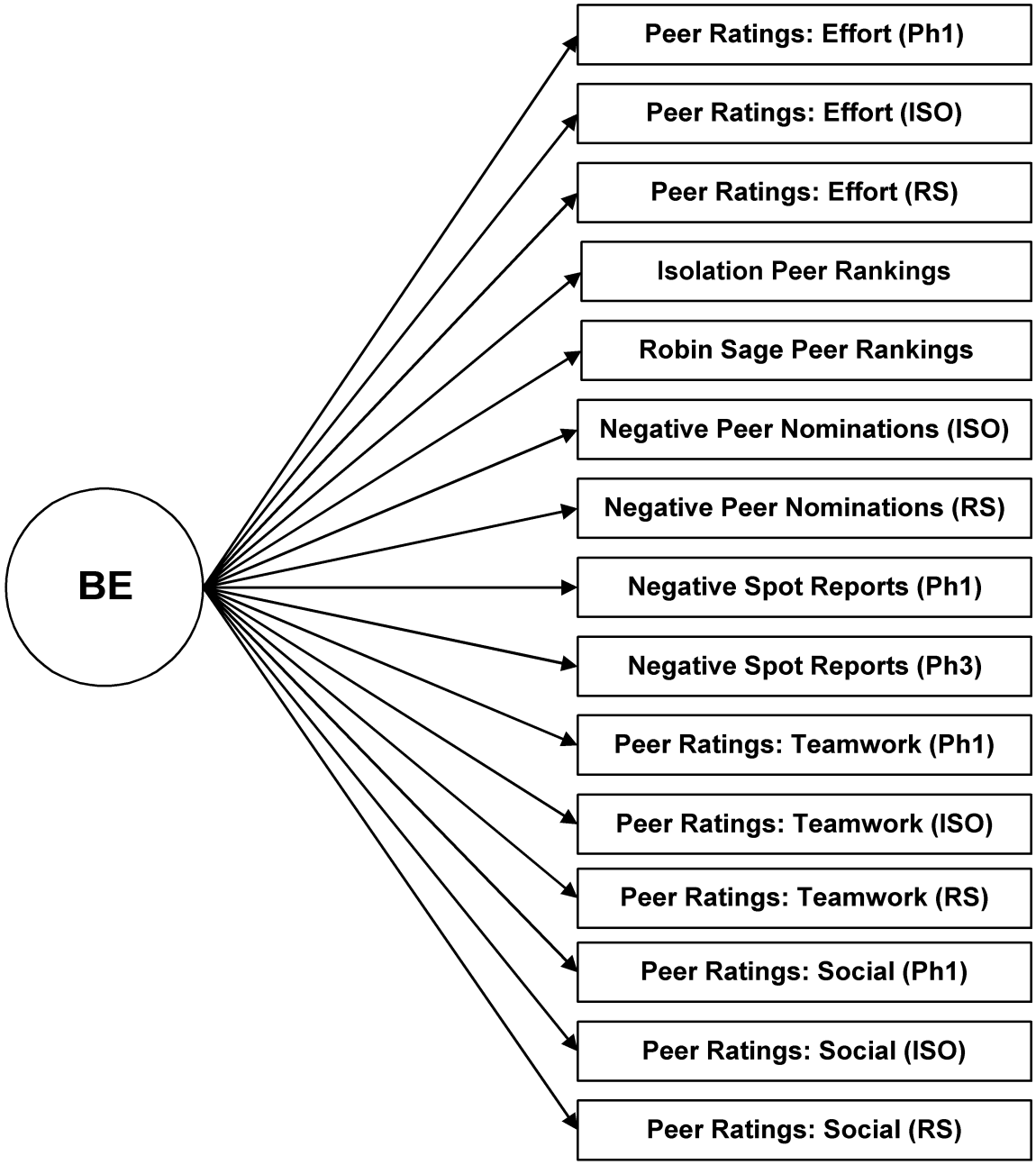


Figure 10. One-factor conceptualization of the BE construct presented with manifest indicators. Ph1 refers to Phase One of SFQC; Ph3 refers to Phase Three of SFQC; ISO refers to the isolation exercise during Phase Three; and RS refers to the Robin Sage field training exercise (FTX) in Phase Three. Latent constructs are shown in ovals, and manifest variables are shown in rectangles.

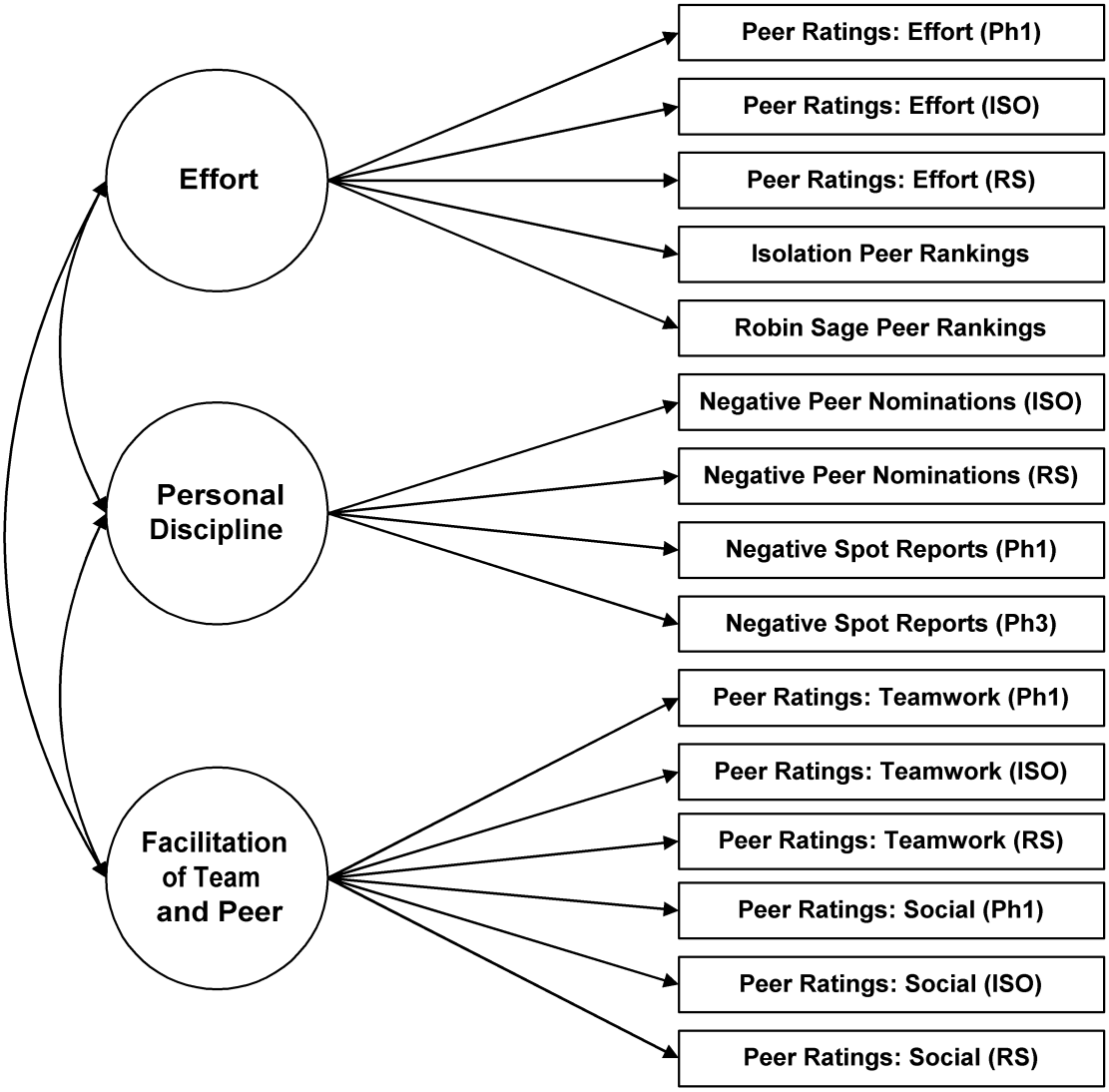


Figure 11. Three-factor conceptualization of the BE construct presented with manifest indicators. Ph1 refers to Phase One of SFQC; Ph3 refers to Phase Three of SFQC; ISO refers to the isolation exercise during Phase Three; and RS refers to the Robin Sage field training exercise (FTX) in Phase Three. Latent constructs are shown in ovals, and manifest variables are shown in rectangles.

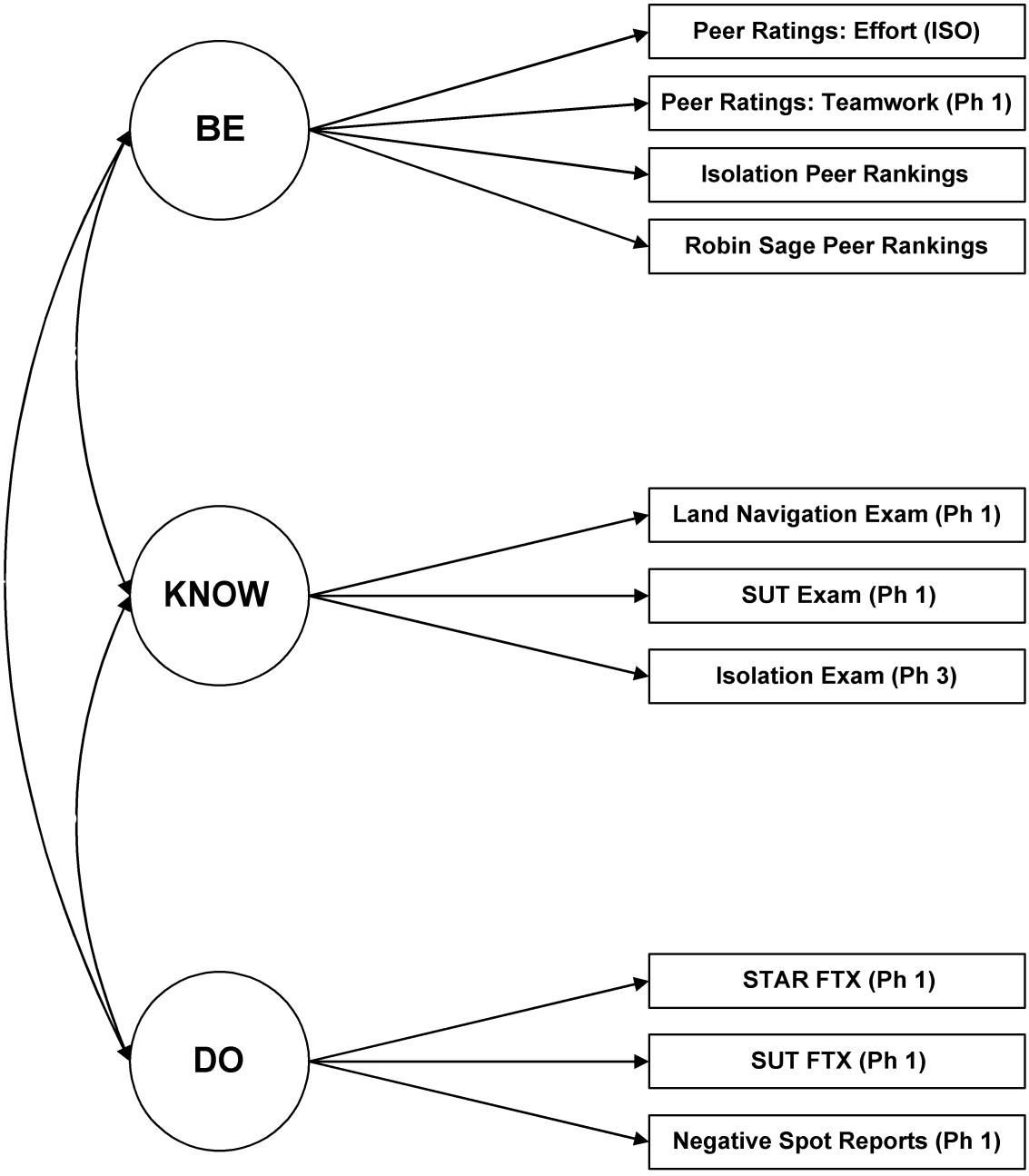


Figure 12. Final BE KNOW DO model presented with manifest indicators. See Figure 3 for the initial BE KNOW DO model. Ph1 refers to Phase One of SFQC; Ph3 refers to Phase Three of SFQC; ISO refers to the isolation exercise during Phase Three; SUT refers to small unit tactics; and FTX refers to field training exercise. STAR is the land navigation FTX. Latent constructs are shown in ovals, and manifest variables are shown in rectangles.

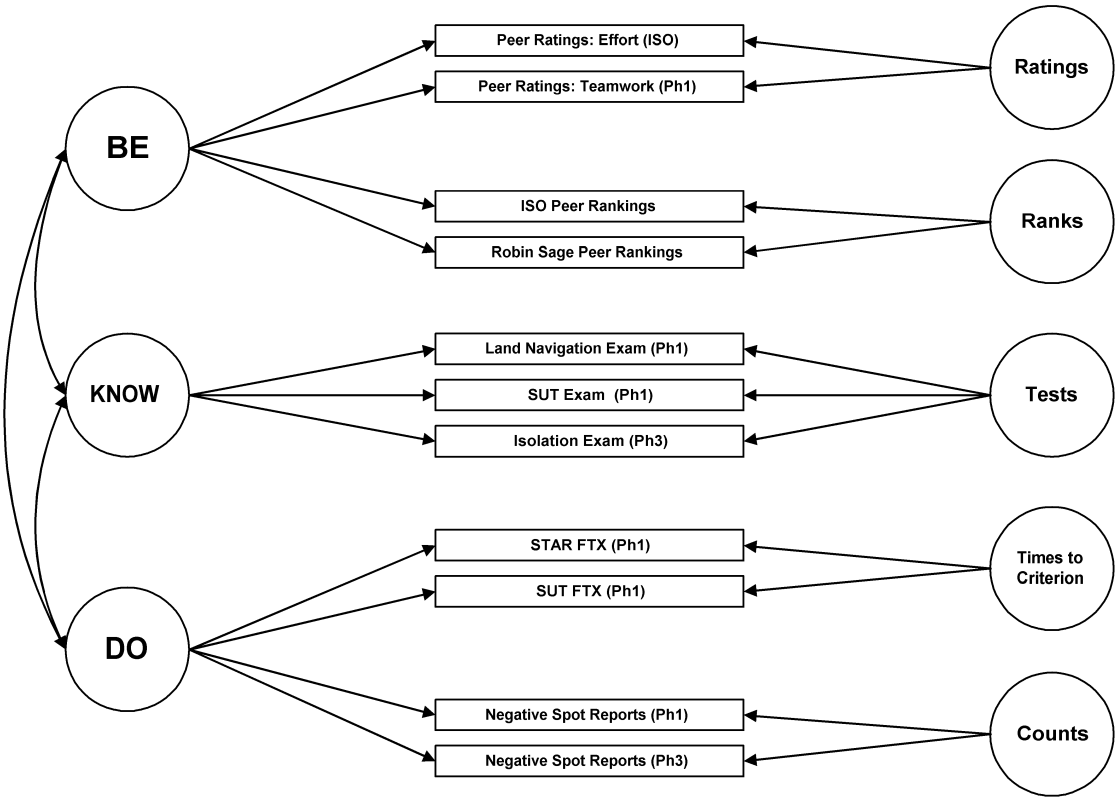


Figure 13. BE KNOW DO model with uncorrelated method factors presented with the initial set of manifest indicators. This model was originally proposed with correlated method factors (see Figure 4) but failed to converge in that form. Latent constructs are shown in ovals, and manifest variables are shown in rectangles.

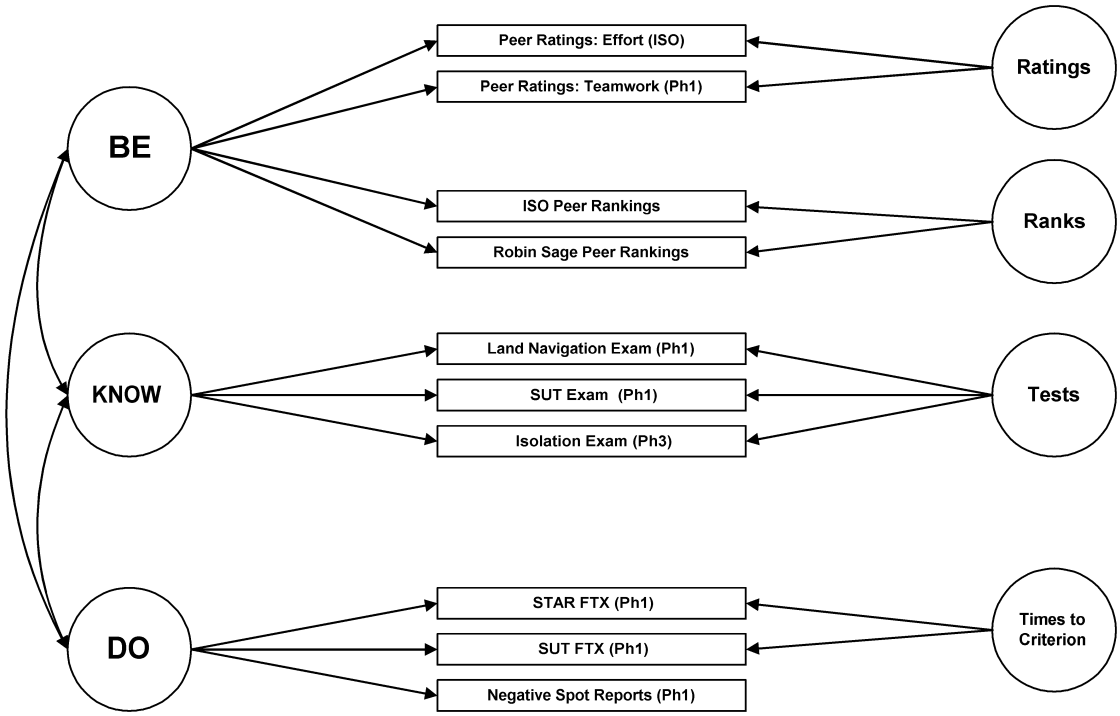


Figure 14. BE KNOW DO model with uncorrelated method factors presented with the final set of manifest indicators. Latent constructs are shown in ovals, and manifest variables are shown in rectangles.

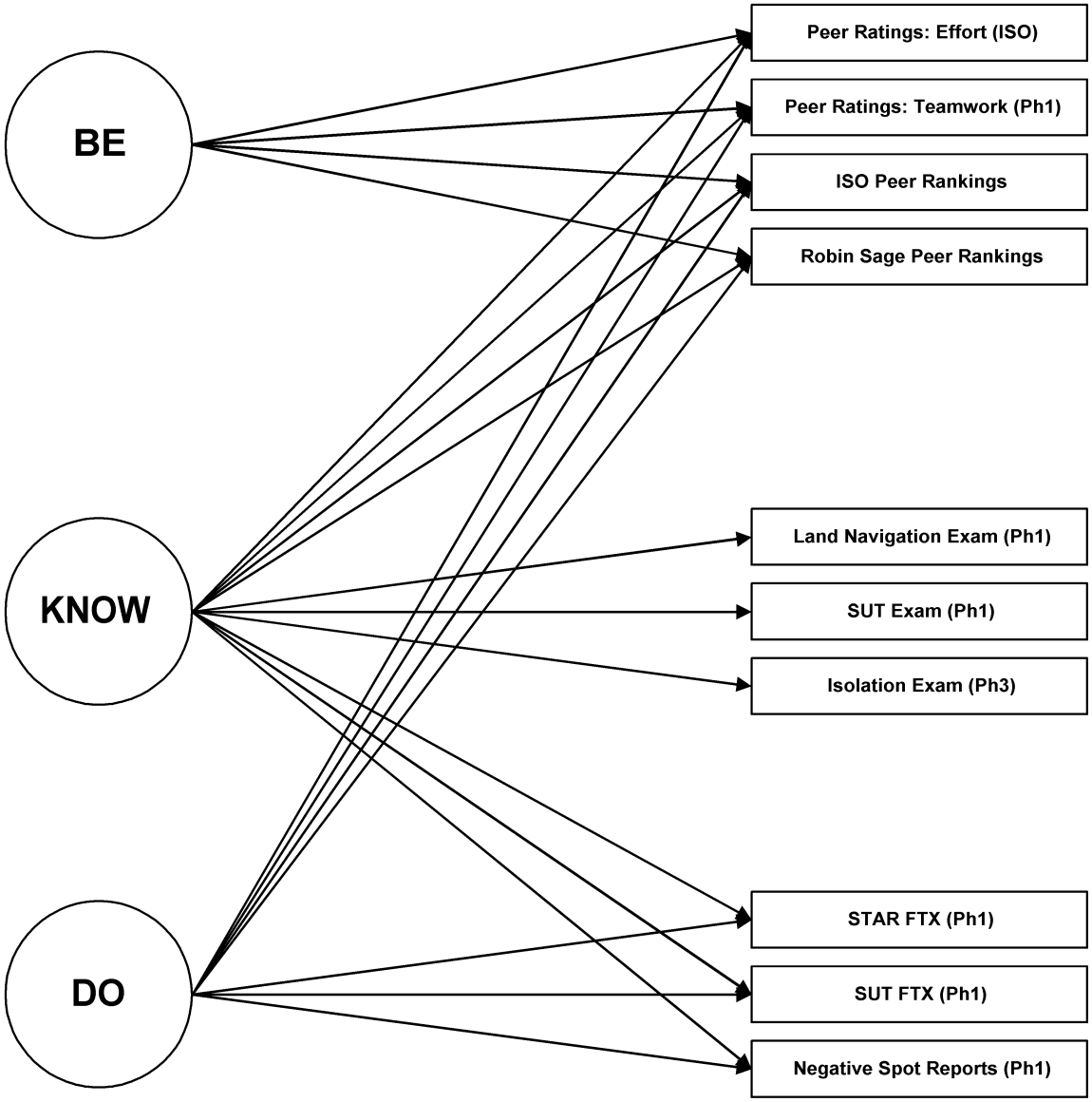


Figure 15. Campbell version of the BE KNOW DO model presented with the final set of manifest indicators. See Figure 5 for the model with the initial set of indicators. Latent constructs are shown in ovals, and manifest variables are shown in rectangles.

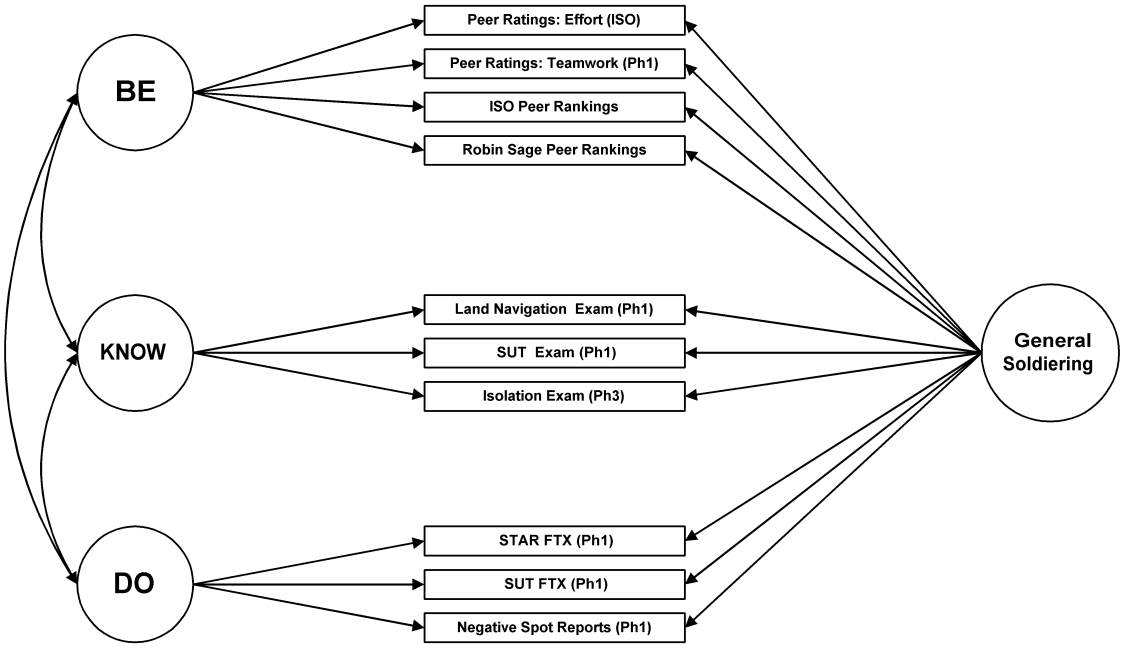


Figure 16. BE KNOW DO model with a general soldiering content factor presented with the final set of manifest indicators. See Figure 6 for the model with the initial set of indicators. Latent constructs are shown in ovals, and manifest variables are shown in rectangles.

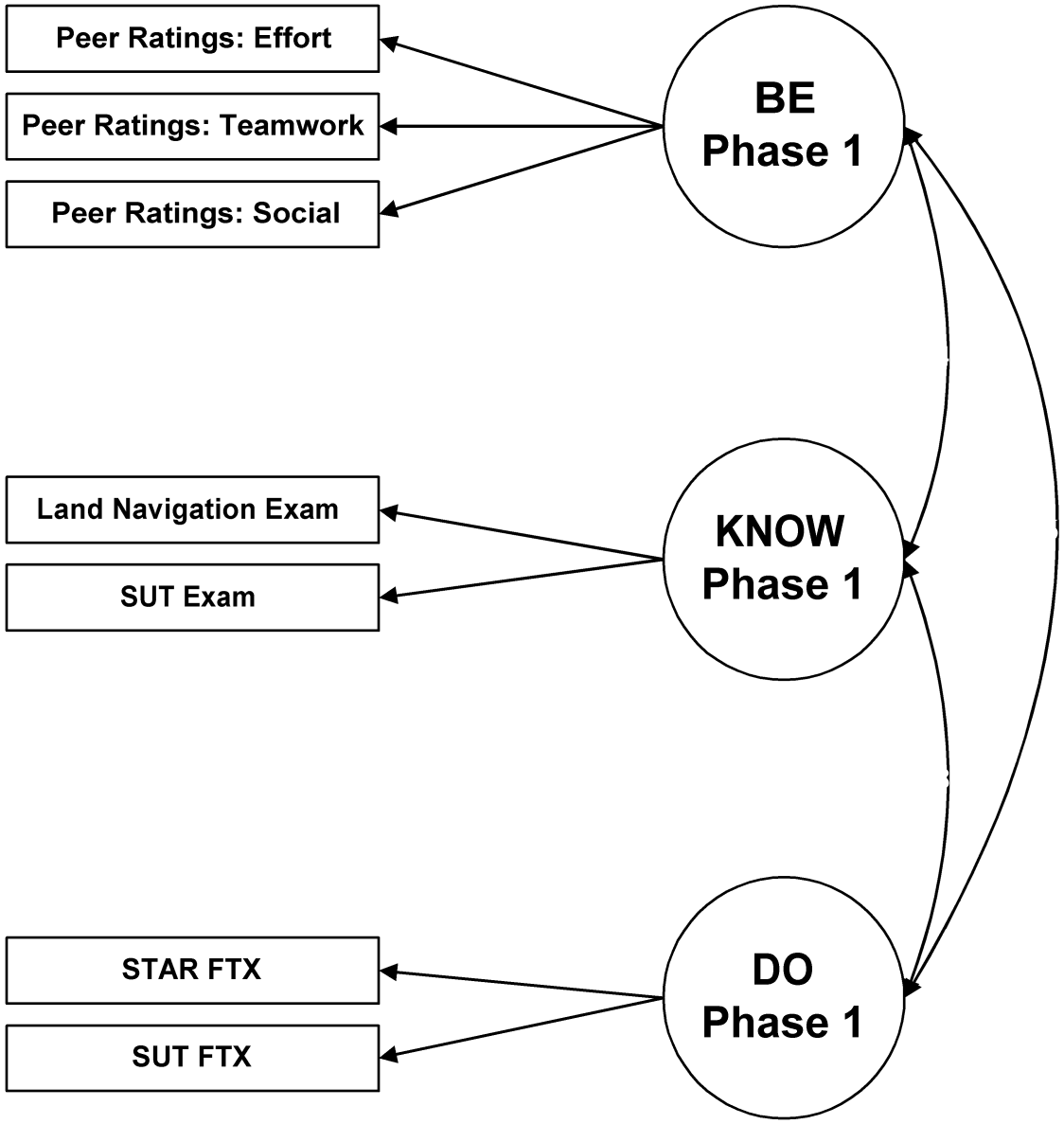


Figure 17. Final Phase One training performance model presented with manifest indicators. See Figure 7 for the model with the initial set of indicators. Latent constructs are shown in ovals, and manifest variables are shown in rectangles.

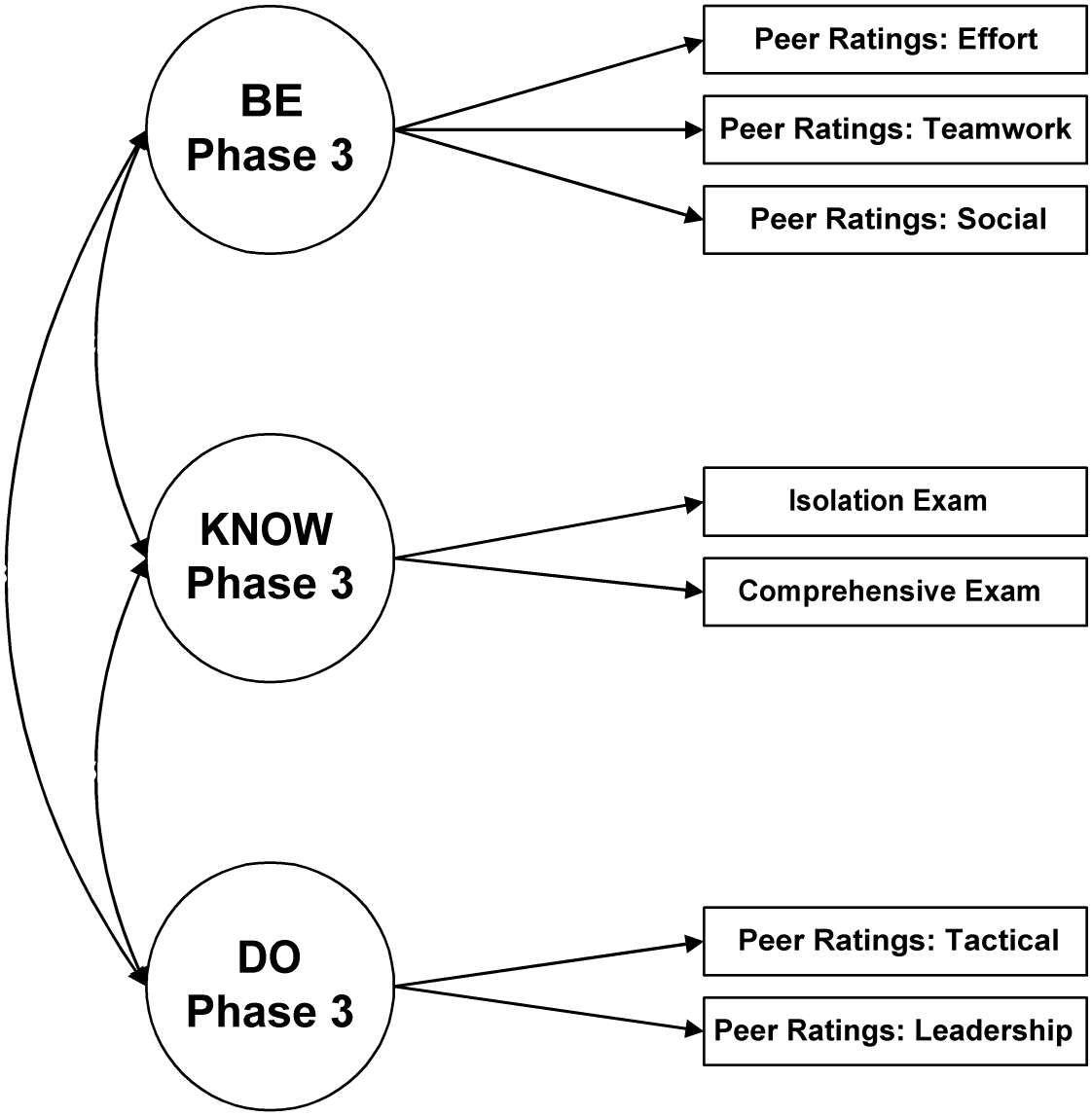


Figure 18. Final Phase Three training performance model presented with manifest indicators. See Figure 8 for the model with the initial set of indicators. Latent constructs are shown in ovals, and manifest variables are shown in rectangles.

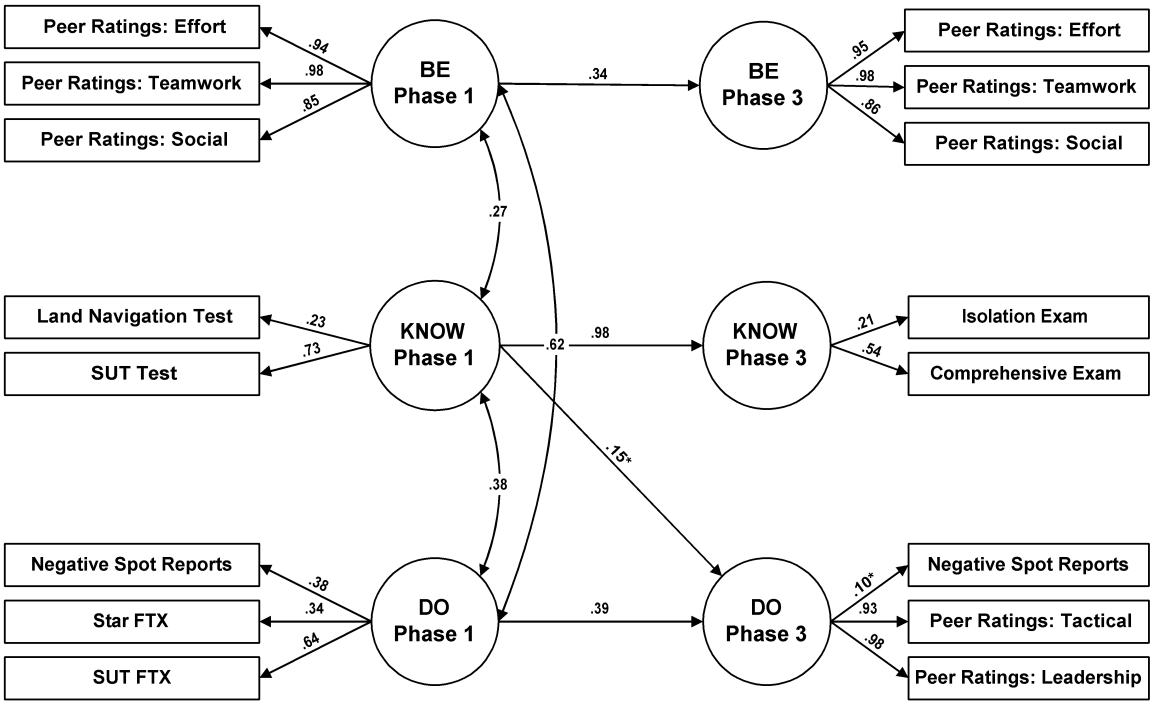


Figure 19. Initial model positing the relationship between Phase One and Phase Three training performance presented with standardized parameter estimates from sample one. Parameter estimates greater than .15 are significant at $p < .05$. Non-significant parameters are marked with an *. This figure represents a structural model with single-headed arrows between latent constructs (ovals) representing a predictive relationship.

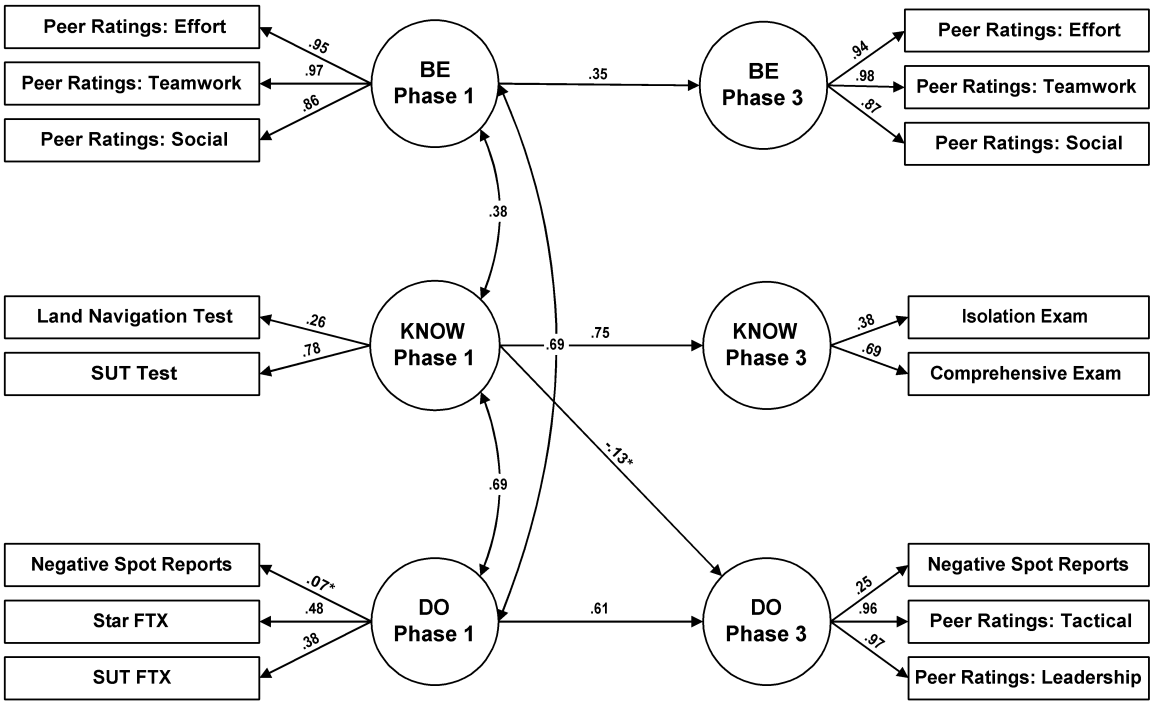


Figure 20. Initial model testing the posited relationship between Phase One and Phase Three training performance presented with standardized parameter estimates from sample two. Parameter estimates greater than .13 are significant at $p < .05$. Non-significant parameters are marked with an *. This figure represents a structural model with single-headed arrows between latent constructs (ovals) representing a predictive relationship.

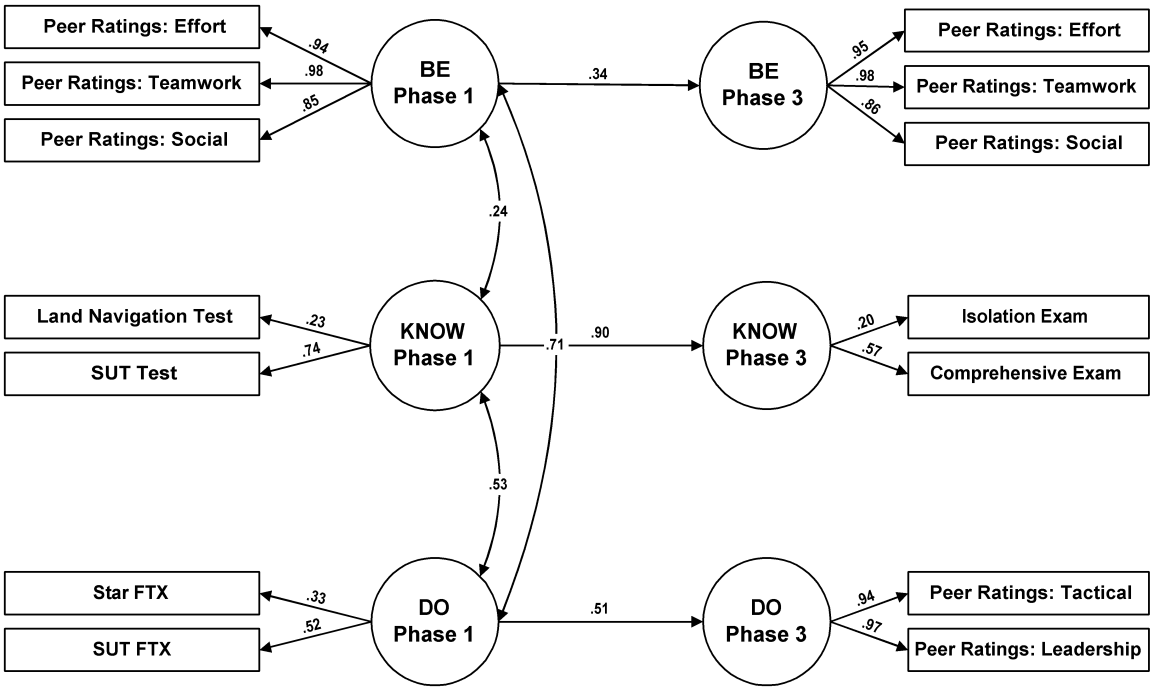


Figure 21. Intermediate model of the relationship between Phase One and Phase Three training performance presented with standardized parameter estimates from sample one. All parameter estimates are significant at $p < .05$. This figure represents a structural model with single-headed arrows between latent constructs (ovals) representing a predictive relationship.

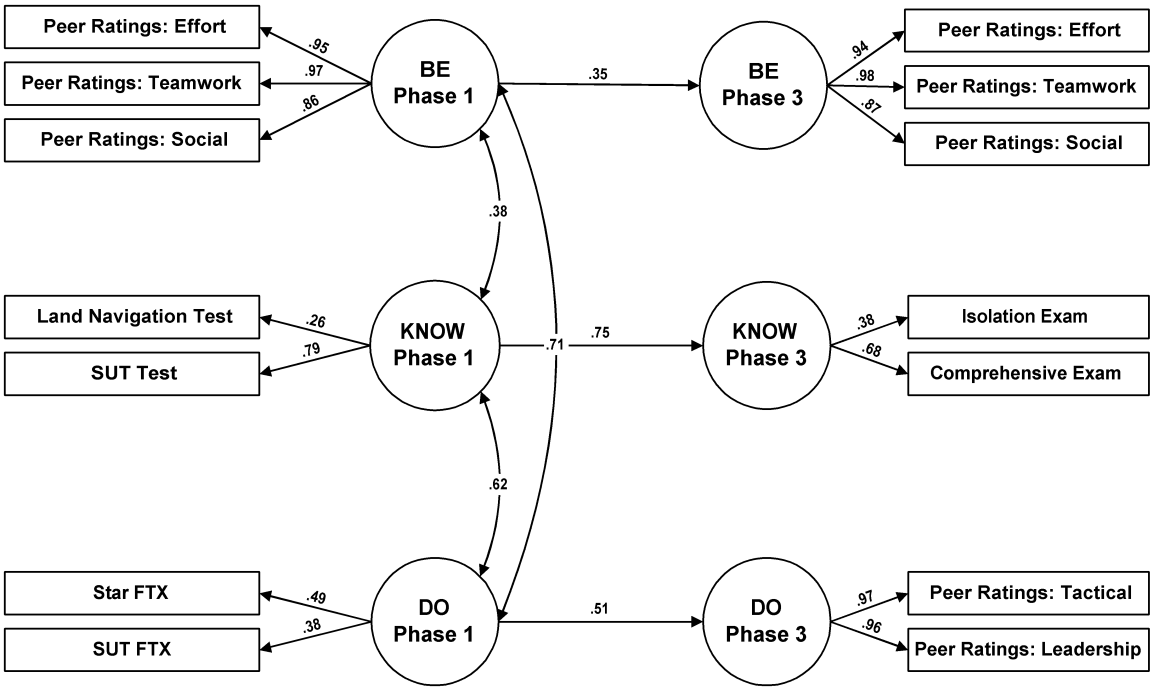


Figure 22. Intermediate model of the relationship between Phase One and Phase Three training performance presented with standardized parameter estimates from sample two. All parameter estimates are significant at $p < .05$. This figure represents a structural model with single-headed arrows between latent constructs (ovals) representing a predictive relationship.

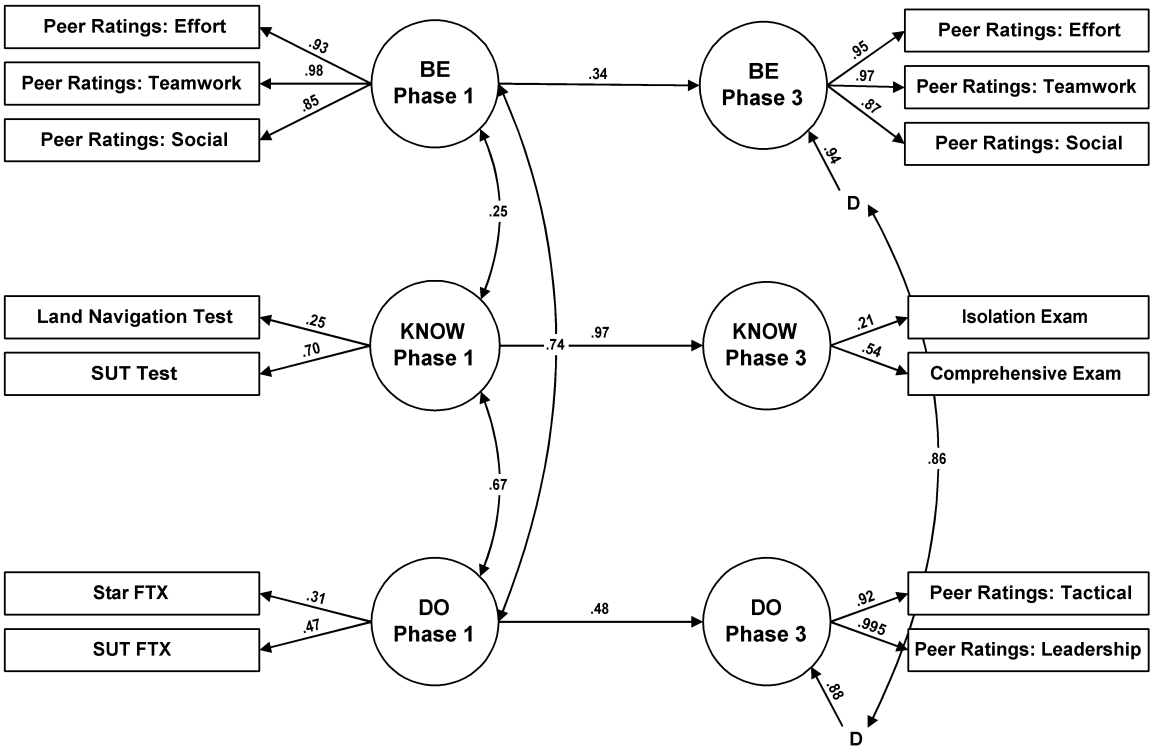


Figure 23. Final model of the relationship between Phase One and Phase Three training performance presented with standardized parameter estimates from sample one. All parameter estimates are significant at $p < .05$. This figure represents a structural model with single-headed arrows between latent constructs (ovals) representing a predictive relationship. D represents the disturbance term.

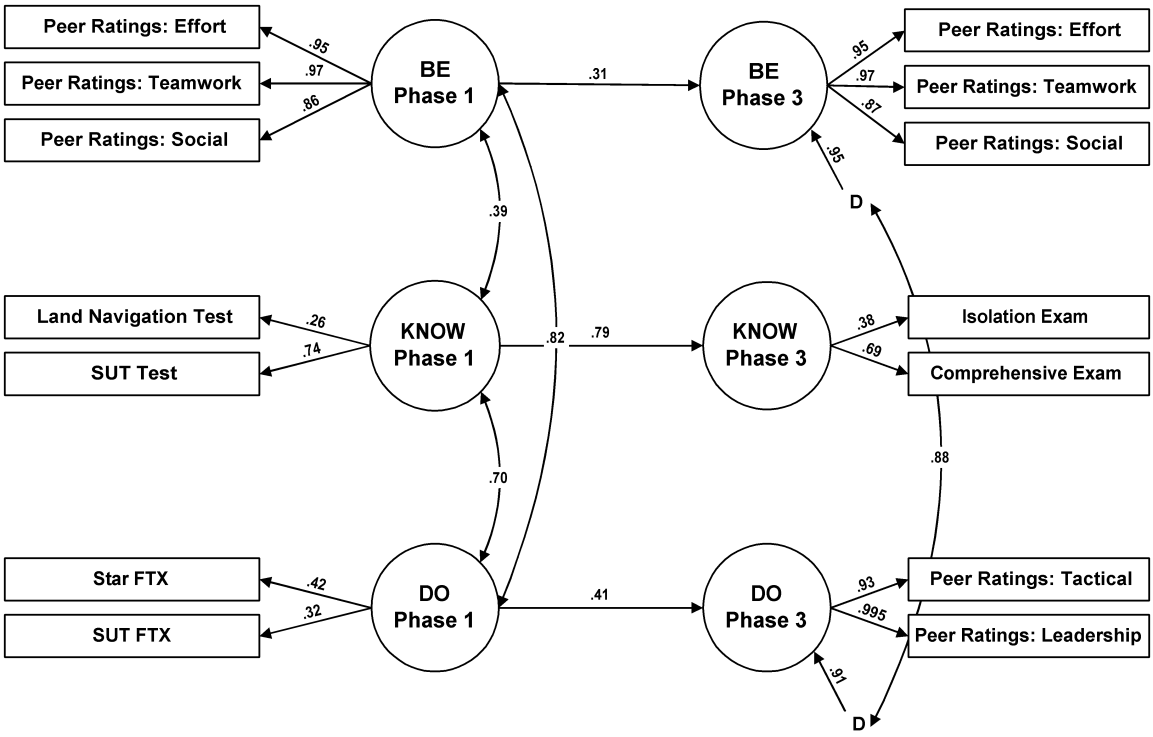


Figure 24. Final model of the relationship between Phase One and Phase Three training performance presented with standardized parameter estimates from sample two. All parameter estimates are significant at $p < .05$. This figure represents a structural model with single-headed arrows between latent constructs (ovals) representing a predictive relationship. D represents the disturbance term.

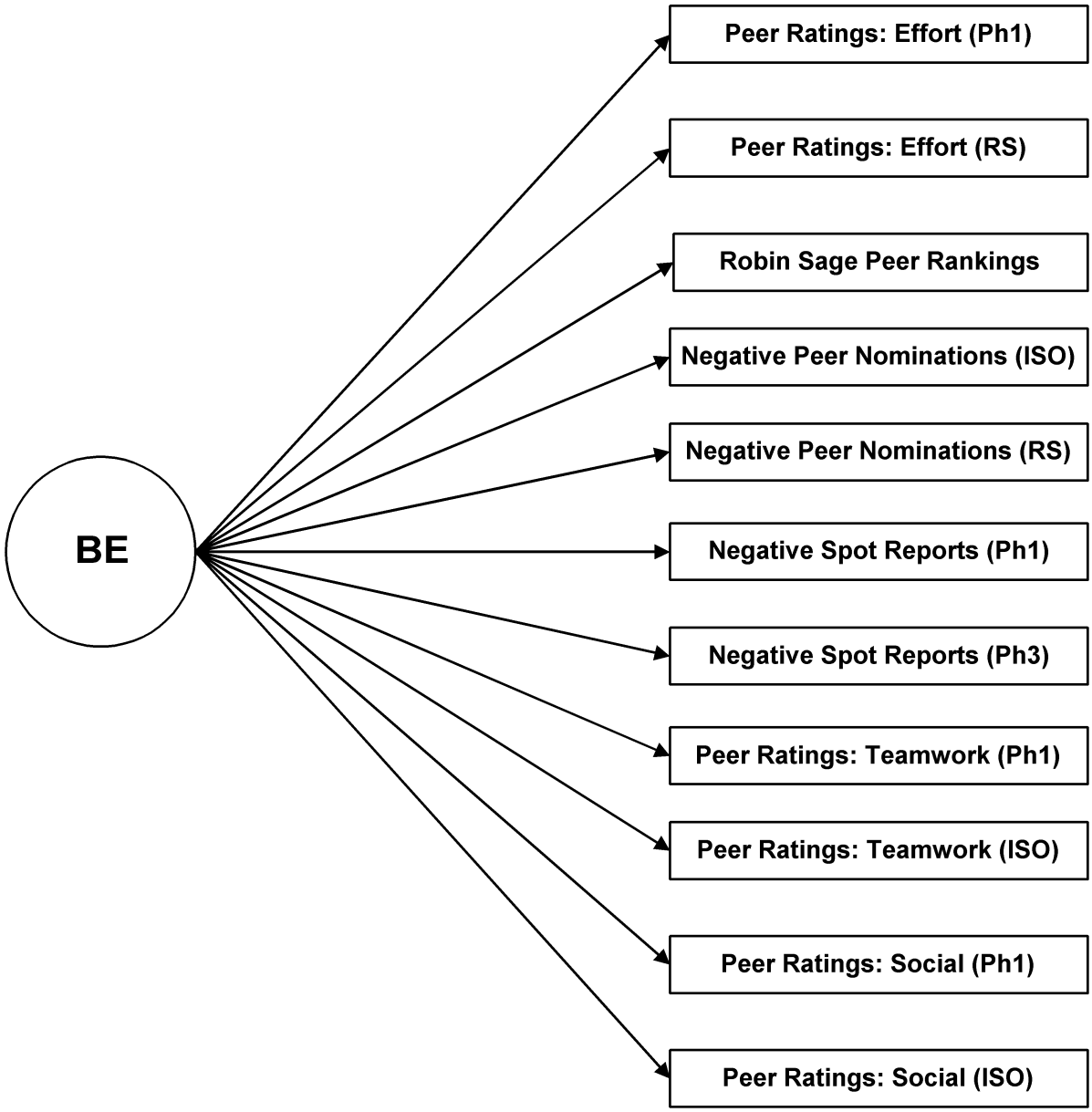


Figure 25. A second version of the one-factor conceptualization of the BE construct presented with manifest indicators. See Figure 10 for the initial version of the model. Latent constructs are shown in ovals, and manifest variables are shown in rectangles.

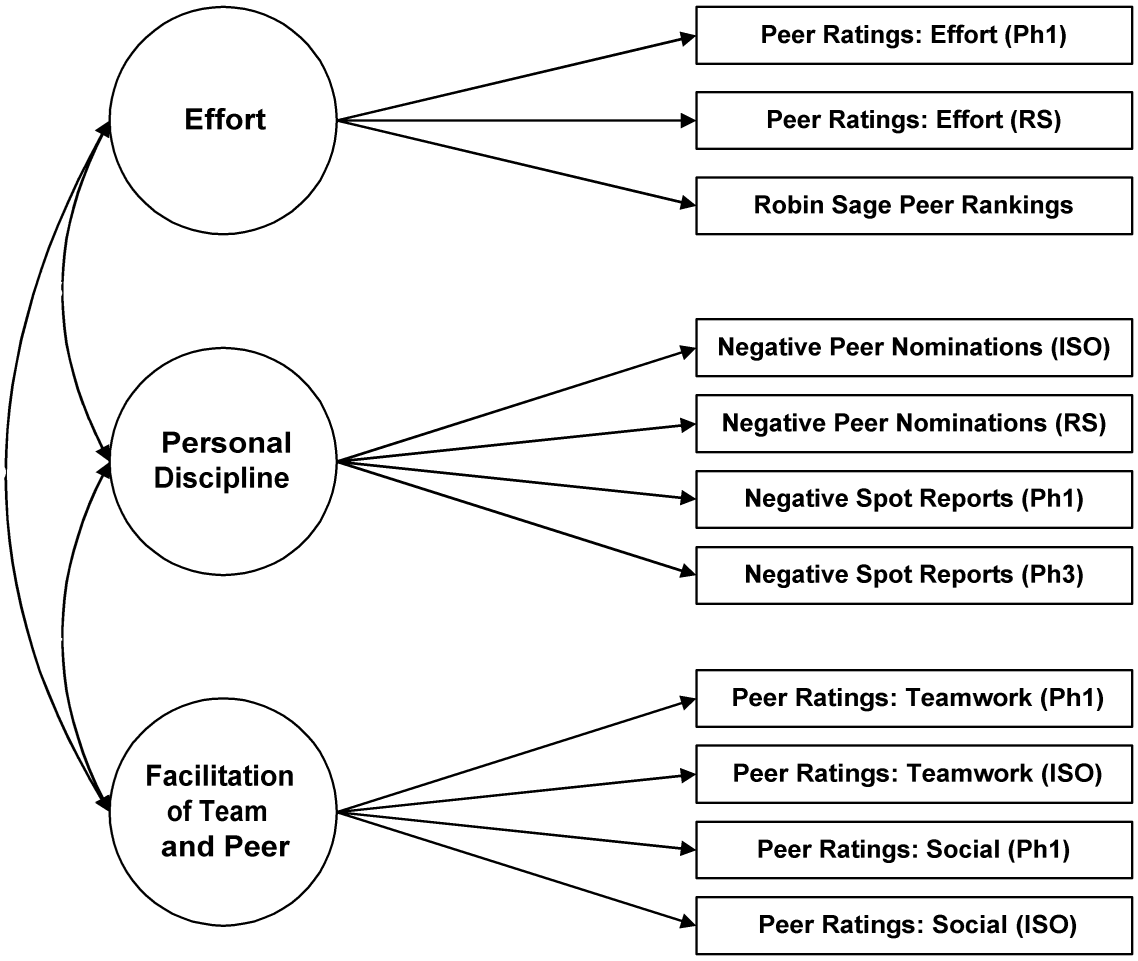


Figure 26. A second version of the three-factor conceptualization of the BE construct presented with manifest indicators. See Figure 11 for the initial version of the model. Latent constructs are shown in ovals, and manifest variables are shown in rectangles.

Appendix A

SFQC Data Collection Form

SFQC Data Collection Form: Phase I

MARKING INSTRUCTIONS

- Use a No. 2 pencil or a blue or black ink pen only.
- Do not use pens with ink that soaks through the paper.
- Make solid marks that fill the response completely.
- Make no stray marks on this form.

CORRECT: ● INCORRECT: ✓ ✗ ◒ ◔

SOCIAL SECURITY NO.

Prior MOS

MOS: 18 A ○ 18 B ○ 18 C ○ 18 D ○ 18 E ○

Class: ① ② ③ ④

Year: 96 ○ 97 ○ 98 ○ 99 ○ 00 ○

SUT Exam	Retest?	SUT Retest
	Y N	

LN Exam	Retest?	LN Retest
	Y N	

Land Navigation Practice Exams:

PE1: ⓪ ① ② ③ ④ ⑤ out of: ② ③ ④ ⑤
PE2: ⓪ ① ② ③ ④ ⑤ out of: ② ③ ④ ⑤
PE3: ⓪ ① ② ③ ④ ⑤ out of: ② ③ ④ ⑤
PE4: ⓪ ① ② ③ ④ ⑤ out of: ② ③ ④ ⑤

Star Exercises: Out of: ① ② ③ points
Star1: ⓪ ① ② ③ ④ ⑤ ⑥
Star2: ⓪ ① ② ③ ④ ⑤ ⑥
Star3: ⓪ ① ② ③ ④ ⑤ ⑥
Star4: ⓪ ① ② ③ ④ ⑤ ⑥
Star5: ⓪ ① ② ③ ④ ⑤ ⑥
Times to pass the Star: ① ② ③ ④ ⑤

# of Positive Spot Reports: ⓪ ① ② ③ ④ ⑤ ⑥ ⑦ ⑧ ⑨ ⑩ ⑪ ⑫ ⑬
# of Negative Spot Reports: ⓪ ① ② ③ ④ ⑤ ⑥ ⑦ ⑧ ⑨ ⑩ ⑪ ⑫ ⑬

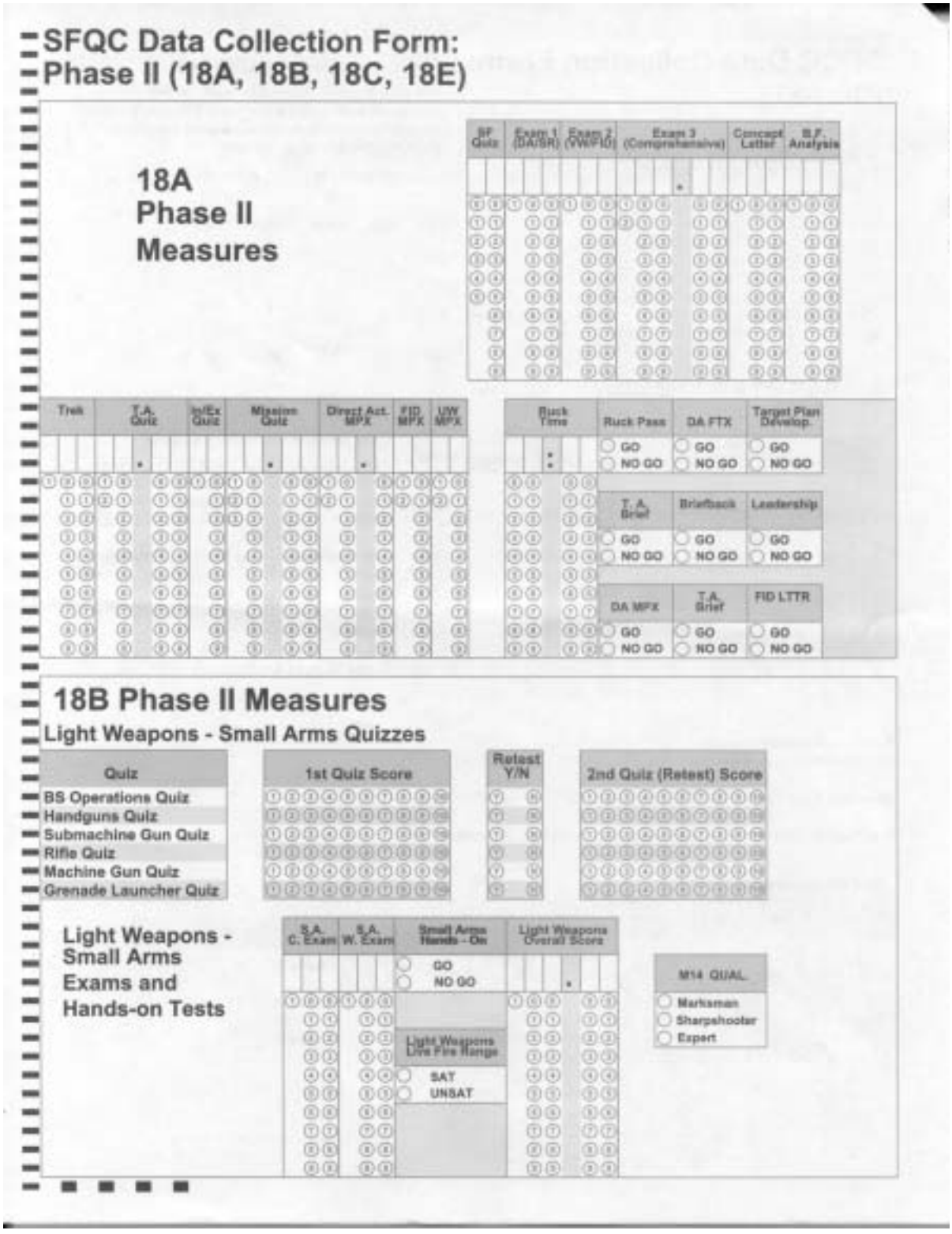
Common Task Test: ⓪ ① ② ③ ④ ⑤ ⑥ ⑦ ⑧ ⑨ ⑩ out of: ② ③ ⑩

# of Student Observation and Evaluation Exercises to receive "go": ① ② ③ ④ ⑤

Peer Evaluations:

Effort	Physical	Social	Teamwork	Tactical	Leadership
•	•	•	•	•	•

Recycles	
Phase 1:	⓪ ① ② ③ ④
Phase 2:	⓪ ① ② ③ ④
Phase 3:	⓪ ① ② ③ ④



SFQC Data Collection Form: Phase II (18A, 18B, 18C, 18E)

18A Phase II Measures

SF Quiz	Exam 1 (DA/SR)	Exam 2 (PW/PI/O)	Exam 3 (Comprehensive)	Concept Letter	S.F. Analysis

Trek	T.A. Quiz	In/Ex Quiz	Mission Quiz	Direct Act. MPX	FID MPX	UW MPX

Buck Time	Ruck Pass	DA FTX	Target Plan Develop.
	GO / NO GO	GO / NO GO	GO / NO GO

T.A. Brief	Briefback	Leadership
GO / NO GO	GO / NO GO	GO / NO GO

DA MPX	T.A. Brief	FID LTTR
GO / NO GO	GO / NO GO	GO / NO GO

18B Phase II Measures

Light Weapons - Small Arms Quizzes

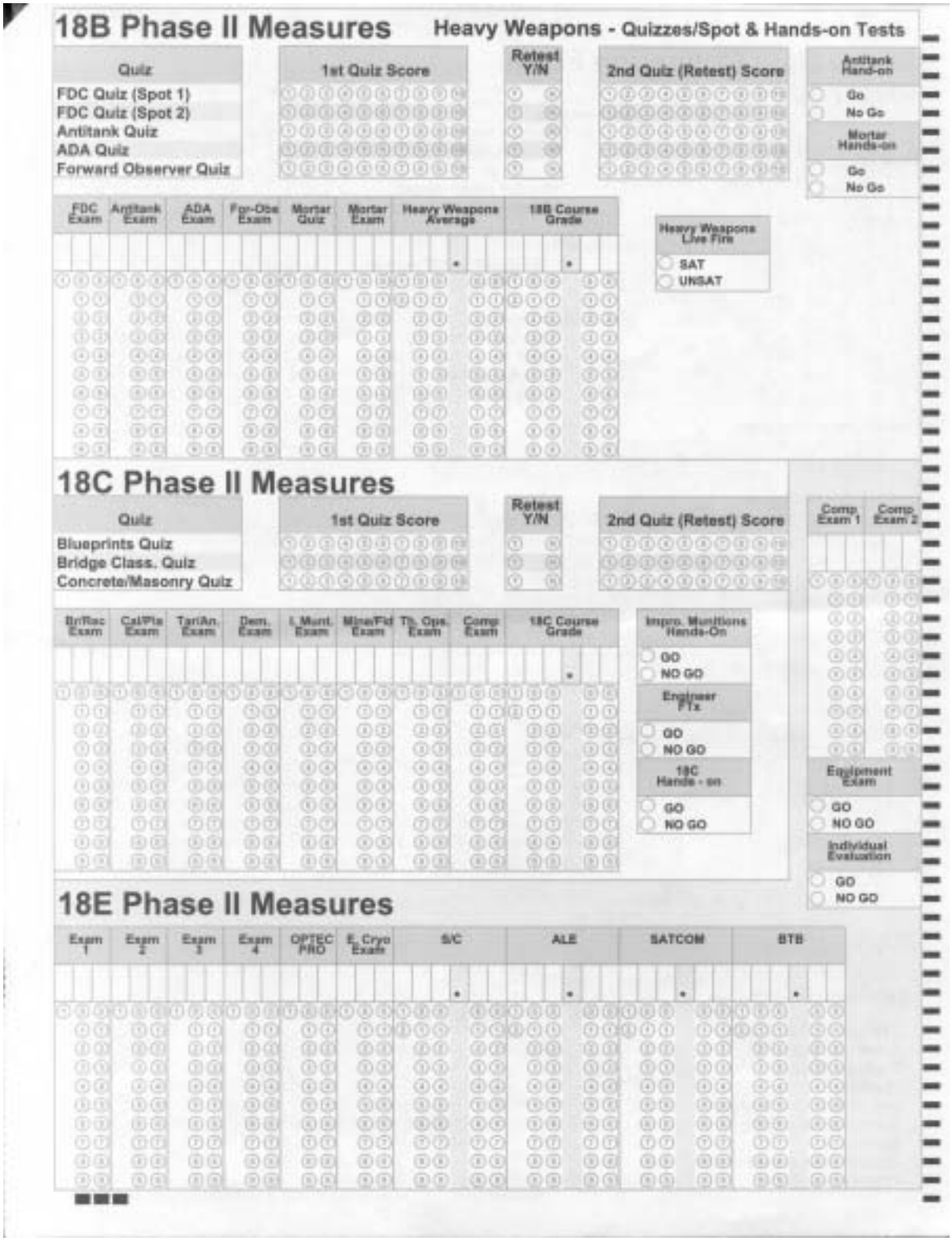
Quiz	1st Quiz Score	Retest Y/N	2nd Quiz (Retest) Score
BS Operations Quiz			
Handguns Quiz			
Submachine Gun Quiz			
Rifle Quiz			
Machine Gun Quiz			
Grenade Launcher Quiz			

Light Weapons - Small Arms Exams and Hands-on Tests

S.A. C. Exam	S.A. W. Exam	Small Arms Hands - On	Light Weapons Overall Score
		GO / NO GO	

Light Weapons Live Fire Range
SAT / UNSAT

M14 QUAL
Marksman
Sharpshooter
Expert


18B Phase II Measures
Heavy Weapons - Quizzes/Spot & Hands-on Tests
Quiz
FDC Quiz (Spot 1)
FDC Quiz (Spot 2)
Antitank Quiz
ADA Quiz
Forward Observer Quiz
1st Quiz Score
Retest Y/N
2nd Quiz (Retest) Score
Antitank Hand-on
Go
No Go
Mortar Hands-on
Go
No Go
FDC Exam
Antitank Exam
ADA Exam
For-Obs Exam
Mortar Quiz
Mortar Exam
Heavy Weapons Average
18B Course Grade
Heavy Weapons Live Fire
SAT
UNSAT
18C Phase II Measures
Quiz
Blueprints Quiz
Bridge Class. Quiz
Concrete/Masonry Quiz
1st Quiz Score
Retest Y/N
2nd Quiz (Retest) Score
Comp Exam 1
Comp Exam 2
Br/Rec Exam
Cal/Pla Exam
Tar/An. Exam
Dem. Exam
L Munt. Exam
Mine/Fld Exam
Th. Ops. Exam
Comp Exam
18C Course Grade
Impro. Munitions Hands-On
GO
NO GO
Engineer FTx
GO
NO GO
18C Hands - on
GO
NO GO
Equipment Exam
GO
NO GO
Individual Evaluation
GO
NO GO
18E Phase II Measures
Exam 1
Exam 2
Exam 3
Exam 4
OPTEC PRO
E. Cryo Exam
S/C
ALE
SATCOM
BTB

SFQC Data Collection Form: Phase III

Comp. Exam

Iso. Exam

Spot Reports
- Positive (Isolation):
- Negative (Isolation):
- Positive (Robin Sage):
- Negative (Robin Sage):

Deploy with or not
- # "Would not want" on team:
- # "Would want" on team:

Peer Evaluations
- After Isolation:
- out of
- After Robin Sage:
- out of

Peer Ratings (Isolation)

Effort	Physical	Social	Teamwork	Tactical	Leadership

Peer Ratings (Robin Sage)

Effort	Physical	Social	Teamwork	Tactical	Leadership

Cadre Assessment Ratings

	Isolation	Robin Sage	Final
Technical & Tactical			
Communications			
Decision Making			
Planning			
Use of Available Systems			
Teaching/Counseling			
Supervision			
Soldier Team Development			
Professional Ethics			

Appendix B

Correlation Matrices for the Research Questions

Table B1.

Descriptive Statistics and Zero-Order Correlations for Question One Dataset

Variables		v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11
M		3.96	3.99	.67	.65	3.46	3.59	38.82	77.32	86.84	4.26	4.43
SD		.41	.54	.42	.41	.36	.22	6.55	14.35	8.46	1.01	.93
N		822	822	822	822	822	822	822	822	822	822	822
Effort (ISO)	v1	--										
Team (Ph1)	v2	.45	--									
RS Ranking	v3	.48	.38	--								
ISO Ranking	v4	.62	.48	.67	--							
Negative Spots (Ph1)	v5	.11	.12	.06	.07	--						
Negative Spots (Ph3)	v6	.09	.04	.13	.06	-.02	--					
ISO Exam	v7	.20	.15	.10	.17	.04	.10	--				
Land Nav Exam	v8	.07	.14	.05	.11	.16	-.02	.05	--			
SUT Exam	v9	.23	.26	.16	.31	.07	.02	.18	.18	--		
SUT FTX	v10	.13	.28	.13	.19	.23	.01	.06	.09	.14	--	
STAR FTX	v11	.15	.20	.07	.15	.09	.06	.13	.10	.21	.22	--

Note. Correlations greater than .07 are significant at $p < .05$. Correlations have been rounded to the nearest hundredth for display purposes. All CFA models were run using the correlation matrix prior to rounding.

Table B2.

Descriptive Statistics and Zero-Order Correlations for Question One, Sample 1

Variables		v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11
M		4.01	4.05	.67	.67	3.46	3.55	39.55	78.14	87.27	4.24	4.44
SD		.42	.57	.41	.41	.33	.28	6.48	13.43	8.60	1.05	.96
N		205	205	205	205	205	205	205	205	205	205	205
Effort (ISO)	v1	--										
Team (Ph1)	v2	.48	--									
RS Ranking	v3	.44	.36	--								
ISO Ranking	v4	.63	.44	.63	--							
Negative Spots (Ph1)	v5	.18	.12	.09	.10	--						
Negative Spots (Ph3)	v6	.06	.09	.18	.03	.00	--					
ISO Exam	v7	.18	.01	.10	.14	.11	.14	--				
Land Nav Exam	v8	.12	.16	-.01	.09	.18	-.06	.03	--			
SUT Exam	v9	.23	.23	.12	.31	.18	-.03	.21	.11	--		
SUT FTX	v10	.09	.23	.03	.10	.19	.01	.05	.09	.13	--	
STAR FTX	v11	.15	.15	.00	.17	.11	.04	.25	.06	.13	.08	--

Note. Correlations greater than .13 are significant at $p < .05$. Correlations have been rounded to the nearest hundredth for display purposes. All CFA models were run using the correlation matrix prior to rounding.

Table B3.

Descriptive Statistics and Zero-Order Correlations for Question One, Sample Two

Variables		v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11
M		3.98	3.98	0.67	0.66	3.46	3.59	39.07	77.75	86.56	4.19	4.48
SD		.41	.56	.40	.39	.36	.20	7.17	14.30	8.68	1.02	.83
N		205	205	205	205	205	205	205	205	205	205	205
Effort (ISO)	v1	--										
Team (Ph1)	v2	.46	--									
RS Ranking	v3	.52	.43	--								
ISO Ranking	v4	.66	.50	.70	--							
Negative Spots (Ph1)	v5	.09	.16	.06	.11	--						
Negative Spots (Ph3)	v6	.13	.01	.13	.09	-.03	--					
ISO Exam	v7	.20	.26	.16	.16	.01	.05	--				
Land Nav Exam	v8	.00	.14	.07	.10	.11	.10	.04	--			
SUT Exam	v9	.24	.32	.18	.29	.06	.10	.20	.16	--		
SUT FTX	v10	.08	.28	.11	.15	.29	.01	.10	.05	.15	--	
STAR FTX	v11	.19	.14	.10	.20	.13	.02	.09	.03	.24	.33	--

Note. Correlations greater than .13 are significant at $p < .05$. Correlations have been rounded to the nearest hundredth for display purposes. All CFA models were run using the correlation matrix prior to rounding.

Table B4.

Descriptive Statistics and Zero-Order Correlations for Question One, Sample Three

Variables		v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11
M		3.93	3.95	.67	.63	3.44	3.60	38.90	77.96	86.85	4.27	4.40
SD		.45	.52	.44	.42	.39	.20	6.46	14.26	8.14	.96	1.02
N		205	205	205	205	205	205	205	205	205	205	205
Effort (ISO)	v1	--										
Team (Ph1)	v2	.50	--									
RS Ranking	v3	.52	.40	--								
ISO Ranking	v4	.64	.49	.74	--							
Negative Spots (Ph1)	v5	.08	.08	.04	.00	--						
Negative Spots (Ph3)	v6	.17	.05	.20	.15	-.04	--					
ISO Exam	v7	.22	.10	.07	.16	-.06	.17	--				
Land Nav Exam	v8	.06	.21	.04	.10	.21	-.02	.04	--			
SUT Exam	v9	.31	.28	.27	.40	.12	.15	.14	.24	--		
SUT FTX	v10	.12	.26	.19	.24	.22	-.02	.00	.17	.19	--	
STAR FTX	v11	.19	.22	.13	.16	.05	.15	.08	.15	.34	.14	--

Note. Correlations greater than .13 are significant at $p < .05$. Correlations have been rounded to the nearest hundredth for display purposes. All CFA models were run using the correlation matrix prior to rounding.

Table B5.

Descriptive Statistics and Zero-Order Correlations for Question One, Sample Four

Variables		v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11
M		3.93	3.98	.65	.66	3.47	3.60	37.78	75.46	86.66	4.32	4.38
SD		.38	.52	.42	.41	.35	.19	5.95	15.28	8.47	1.01	.92
N		207	207	207	207	207	207	207	207	207	207	207
Effort (ISO)	v1	--										
Team (Ph1)	v2	.37	--									
RS Ranking	v3	.44	.35	--								
ISO Ranking	v4	.57	.47	.60	--							
Negative Spots (Ph1)	v5	.11	.10	.03	.09	--						
Negative Spots (Ph3)	v6	.02	-.01	.00	-.01	-.03	--					
ISO Exam	v7	.17	.21	.10	.23	.13	.06	--				
Land Nav Exam	v8	.08	.05	.09	.16	.16	-.07	.04	--			
SUT Exam	v9	.12	.20	.07	.24	-.06	-.13	.15	.22	--		
SUT FTX	v10	.26	.36	.18	.26	.23	.04	.10	.08	.10	--	
STAR FTX	v11	.04	.29	.02	.08	.06	.02	.07	.14	.14	.37	--

Note. Correlations greater than .13 are significant at $p < .05$. Correlations have been rounded to the nearest hundredth for display purposes. All CFA models were run using the correlation matrix prior to rounding.

Table B6.

Descriptive Statistics and Zero-Order Correlations for Question Two Dataset

	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	v12	v13	v14	v15	v16
M	3.80	3.61	3.95	3.92	3.74	4.00	3.87	3.75	1.62	1.38	4.43	4.22	85.5	76.6	38.5	86.5
SD	.57	.55	.53	.50	.51	.51	.53	.56	.58	.42	.95	1.05	8.32	14.2	6.19	6.13
N	558	558	558	558	558	558	558	558	558	558	558	558	558	558	558	558
v1	--															
v2	.80	--														
v3	.92	.84	--													
v4	.33	.25	.30	--												
v5	.31	.34	.30	.81	--											
v6	.32	.28	.30	.93	.85	--										
v7	.38	.30	.34	.76	.71	.75	--									
v8	.38	.31	.33	.81	.78	.81	.92	--								
v9	-.06	-.04	-.10	-.08	-.01	-.06	-.06	-.06	--							
v10	-.10	-.12	-.09	-.19	-.25	-.18	-.15	-.17	.02	--						
v11	.26	.19	.21	.11	.10	.11	.16	.17	-.10	-.05	--					
v12	.33	.26	.32	.13	.14	.13	.20	.19	-.27	.00	.25	--				
v13	.24	.19	.22	.08	.05	.06	.20	.20	-.06	-.02	.26	.15	--			
v14	.14	.09	.11	-.01	.01	-.03	.04	.04	-.20	.01	.16	.14	.18	--		
v15	.13	.12	.11	.16	.18	.13	.20	.21	-.05	-.15	.09	.05	.16	.04	--	
v16	.14	.16	.14	.00	.07	.02	.10	.14	.00	-.02	.16	.05	.40	.11	.19	--

Note. Correlations greater than .08 are significant at $p < .05$. v1 = Peer Rating: Effort (Ph1); v2 = Peer Rating: Social (Ph1); v3 = Peer Rating: Team (Ph1); v4 = Peer Rating: Effort (RS); v5 = Peer Rating: Social (RS); v6 = Peer Rating: team (RS); v7 = Peer Rating: Tactical (RS); v8 = Peer Rating: Leadership (RS); v9 = Negative Spot Reports (Ph1); v10 = Negative Spot Reports (Ph3); v11 = STAR Field Training Exercise (Ph1); v12 = SUT Field Training Exercise (Ph1); v13 = SUT Exam (Ph1); v14 = Land Navigation Exam (Ph1); v15 = Isolation Exam (ISO); and v16 = Comprehensive Exam (Ph3). Correlations have been rounded to the nearest hundredth for display purposes. All CFA models were run using the correlation matrix prior to rounding.

Table B7.

Descriptive Statistics and Zero-Order Correlations for Question Two, Sample One

	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	v12	v13	v14	v15	v16
M	3.81	3.62	3.96	3.90	3.74	3.97	3.87	3.76	3.42	3.59	4.41	4.15	85.7	76.6	37.9	86.2
SD	.58	.55	.54	.50	.51	.50	.51	.54	.37	.20	.97	1.14	8.09	13.7	6.53	6.12
N	279	279	279	279	279	279	279	279	279	279	279	279	279	279	279	279
v1	--															
v2	.79	--														
v3	.92	.84	--													
v4	.33	.28	.33	--												
v5	.29	.36	.32	.81	--											
v6	.30	.29	.31	.93	.85	--										
v7	.37	.31	.33	.74	.70	.72	--									
v8	.36	.32	.33	.80	.75	.78	.91	--								
v9	.18	.17	.21	.06	.03	.04	.09	.07	--							
v10	.11	.15	.11	.12	.21	.11	.07	.10	-.01	--						
v11	.20	.12	.14	.07	.05	.04	.12	.12	.15	.04	--					
v12	.39	.30	.38	.14	.15	.15	.25	.24	.31	.06	.23	--				
v13	.19	.12	.14	.04	-.02	.01	.16	.19	.08	-.03	.27	.14	--			
v14	.19	.11	.15	.02	.03	-.01	.07	.09	.30	.05	.20	.11	.15	--		
v15	.17	.15	.14	.15	.19	.13	.21	.23	.06	.13	.06	.06	.11	.02	--	
v16	.09	.13	.11	-.01	.06	-.01	.07	.13	-.01	.01	.12	.03	.39	.11	.11	--

Note. Correlations greater than .11 are significant at $p < .05$. v1 = Peer Rating: Effort (Ph1); v2 = Peer Rating: Social (Ph1); v3 = Peer Rating: Team (Ph1); v4 = Peer Rating: Effort (RS); v5 = Peer Rating: Social (RS); v6 = Peer Rating: team (RS); v7 = Peer Rating: Tactical (RS); v8 = Peer Rating: Leadership (RS); v9 = Negative Spot Reports (Ph1); v10 = Negative Spot Reports (Ph3); v11 = STAR Field Training Exercise (Ph1); v12 = SUT Field Training Exercise (Ph1); v13 = SUT Exam (Ph1); v14 = Land Navigation Exam (Ph1); v15 = Isolation Exam (ISO); and v16 = Comprehensive Exam (Ph3). Correlations have been rounded to the nearest hundredth for display purposes. All CFA models were run using the correlation matrix prior to rounding.

Table B8.

Descriptive Statistics and Zero-Order Correlations for Question Two, Sample Two

	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	v12	v13	v14	v15	v16
M	3.79	3.60	3.93	3.94	3.75	4.02	3.86	3.75	3.48	3.59	4.44	4.28	85.4	76.7	39.1	86.8
SD	.56	.54	.52	.51	.51	.52	.54	.57	.34	.20	.93	.95	8.56	14.8	5.78	6.15
N	279	279	279	279	279	279	279	279	279	279	279	279	279	279	279	279
v1	--															
v2	.81	--														
v3	.93	.84	--													
v4	.34	.23	.27	--												
v5	.34	.32	.29	.82	--											
v6	.35	.27	.30	.92	.85	--										
v7	.40	.30	.35	.78	.73	.78	--									
v8	.40	.30	.34	.83	.81	.83	.93	--								
v9	-.08	-.09	-.04	.09	-.02	.05	.02	.05	--							
v10	.10	.07	.07	.29	.30	.26	.23	.25	.05	--						
v11	.32	.28	.29	.16	.16	.17	.20	.22	.05	.06	--					
v12	.26	.22	.24	.12	.13	.11	.16	.15	.20	-.07	.27	--				
v13	.28	.26	.29	.12	.11	.11	.23	.21	.09	.06	.25	.17	--			
v14	.08	.07	.08	-.04	-.01	-.04	.01	-.01	.12	-.08	.12	.19	.21	--		
v15	.10	.10	.08	.17	.17	.13	.19	.20	.00	.18	.12	.02	.22	.07	--	
v16	.19	.19	.18	.01	.09	.04	.13	.16	.01	.02	.20	.06	.41	.11	.26	--

Note. Correlations greater than .11 are significant at $p < .05$. v1 = Peer Rating: Effort (Ph1); v2 = Peer Rating: Social (Ph1); v3 = Peer Rating: Team (Ph1); v4 = Peer Rating: Effort (RS); v5 = Peer Rating: Social (RS); v6 = Peer Rating: team (RS); v7 = Peer Rating: Tactical (RS); v8 = Peer Rating: Leadership (RS); v9 = Negative Spot Reports (Ph1); v10 = Negative Spot Reports (Ph3); v11 = STAR Field Training Exercise (Ph1); v12 = SUT Field Training Exercise (Ph1); v13 = SUT Exam (Ph1); v14 = Land Navigation Exam (Ph1); v15 = Isolation Exam (ISO); and v16 = Comprehensive Exam (Ph3). Correlations have been rounded to the nearest hundredth for display purposes. All CFA models were run using the correlation matrix prior to rounding.

Table B9.

Descriptive Statistics and Zero-Order Correlations for Question Three Dataset

	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	v12	v13	v14	v15
M	3.84	3.65	3.98	3.97	3.77	4.03	3.91	3.98	3.74	3.49	3.57	3.57	3.54	.64	.66
SD	.57	.55	.54	.42	.45	.41	.49	.50	.51	.33	.23	.34	.37	.40	.41
N	685	685	685	685	685	685	685	685	685	685	685	685	685	685	685
v1	--														
v2	.82	--													
v3	.92	.85	--												
v4	.46	.38	.43	--											
v5	.37	.42	.36	.76	--										
v6	.41	.36	.38	.88	.82	--									
v7	.36	.28	.33	.56	.49	.58	--								
v8	.33	.29	.31	.52	.51	.57	.92	--							
v9	.32	.33	.31	.50	.61	.56	.81	.84	--						
v10	.08	.07	.10	.12	.08	.10	.09	.06	.03	--					
v11	.03	.05	.02	.07	.09	.08	.16	.16	.20	-.05	--				
v12	.30	.32	.29	.54	.57	.55	.33	.33	.37	.04	.12	--			
v13	.27	.26	.26	.31	.33	.34	.60	.61	.56	.03	.17	.45	--		
v14	.49	.45	.45	.61	.55	.57	.40	.37	.40	.10	.03	.46	.27	--	
v15	.42	.36	.38	.47	.45	.47	.61	.58	.57	.07	.12	.38	.45	.65	--

Note. Correlations greater than .07 are significant at $p < .05$. v1 = Peer Rating: Effort (Ph1); v2 = Peer Rating: Social (Ph1); v3 = Peer Rating: Team (Ph1); v4 = Peer Rating: Effort (ISO); v5 = Peer Rating: Social (ISO); v6 = Peer Rating: team (ISO); v7 = Peer Rating: Effort (RS); v8 = Peer Rating: Team (RS); v9 = Peer Rating: Social (RS); v10 = Negative Spot Reports (Ph1); v11 = Negative Spot Reports (Ph3); v12 = Negative Peer Nominations (ISO); v13 = Negative Peer Nominations (RS); v14 = ISO Peer Rankings; and v15 = RS Peer Rankings. Correlations have been rounded to the nearest hundredth for display purposes. All CFA models were run using the correlation matrix prior to rounding.

Table B10.

Descriptive Statistics and Zero-Order Correlations for Question Three, Sample One

	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	v12	v13	v14	v15
M	3.83	3.64	3.98	3.96	3.76	4.02	3.90	3.97	3.73	3.47	3.56	3.57	3.52	.64	.65
SD	.58	.55	.55	.44	.46	.42	.52	.52	.52	.34	.24	.34	.41	.43	.41
N	340	340	340	340	340	340	340	340	340	340	340	340	340	340	340
v1	--														
v2	.81	--													
v3	.92	.85	--												
v4	.48	.39	.44	--											
v5	.39	.44	.38	.76	--										
v6	.44	.37	.39	.88	.81	--									
v7	.39	.28	.37	.57	.45	.57	--								
v8	.37	.31	.36	.54	.48	.55	.93	--							
v9	.37	.39	.38	.52	.61	.55	.81	.85	--						
v10	.08	.11	.13	.11	.10	.10	.11	.09	.05	--					
v11	.03	.05	.03	.11	.12	.11	.11	.13	.20	-.04	--				
v12	.33	.34	.30	.54	.56	.52	.30	.30	.35	.07	.18	--			
v13	.29	.23	.28	.36	.31	.35	.64	.63	.58	.07	.17	.44	--		
v14	.52	.47	.49	.62	.55	.56	.41	.40	.44	.13	.01	.46	.28	--	
v15	.44	.34	.40	.49	.44	.46	.61	.59	.60	.09	.08	.38	.44	.69	--

Note. Correlations greater than .10 are significant at $p < .05$. v1 = Peer Rating: Effort (Ph1); v2 = Peer Rating: Social (Ph1); v3 = Peer Rating: Team (Ph1); v4 = Peer Rating: Effort (ISO); v5 = Peer Rating: Social (ISO); v6 = Peer Rating: team (ISO); v7 = Peer Rating: Effort (RS); v8 = Peer Rating: Team (RS); v9 = Peer Rating: Social (RS); v10 = Negative Spot Reports (Ph1); v11 = Negative Spot Reports (Ph3); v12 = Negative Peer Nominations (ISO); v13 = Negative Peer Nominations (RS); v14 = ISO Peer Rankings; and v15 = RS Peer Rankings. Correlations have been rounded to the nearest hundredth for display purposes. All CFA models were run using the correlation matrix prior to rounding.

Table B11.

Descriptive Statistics and Zero-Order Correlations for Question Three, Sample Two

	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	v12	v13	v14	v15
M	3.85	3.66	3.98	3.97	3.78	4.03	3.92	4.00	3.75	3.50	3.57	3.57	3.55	.64	.66
SD	.56	.56	.53	.41	.43	.40	.47	.48	.49	.33	.22	.33	.33	.38	.42
N	345	345	345	345	345	345	345	345	345	345	345	345	345	345	345
v1	--														
v2	.83	--													
v3	.92	.85	--												
v4	.43	.38	.41	--											
v5	.35	.40	.35	.76	--										
v6	.39	.36	.37	.87	.84	--									
v7	.33	.28	.29	.55	.52	.60	--								
v8	.28	.27	.26	.50	.53	.59	.90	--							
v9	.25	.27	.24	.49	.62	.57	.81	.84	--						
v10	.08	.02	.08	.13	.06	.09	.06	.02	.01	--					
v11	.04	.05	.02	.01	.05	.05	.21	.19	.21	-.05	--				
v12	.27	.30	.29	.55	.59	.59	.37	.37	.40	.01	.05	--			
v13	.25	.28	.24	.26	.36	.34	.55	.58	.54	-.02	.17	.46	--		
v14	.46	.44	.41	.59	.55	.58	.37	.34	.37	.06	.06	.46	.24	--	
v15	.40	.38	.36	.45	.46	.48	.61	.57	.54	.06	.16	.38	.46	.62	--

Note. Correlations greater than .10 are significant at $p < .05$. v1 = Peer Rating: Effort (Ph1); v2 = Peer Rating: Social (Ph1); v3 = Peer Rating: Team (Ph1); v4 = Peer Rating: Effort (ISO); v5 = Peer Rating: Social (ISO); v6 = Peer Rating: team (ISO); v7 = Peer Rating: Effort (RS); v8 = Peer Rating: Team (RS); v9 = Peer Rating: Social (RS); v10 = Negative Spot Reports (Ph1); v11 = Negative Spot Reports (Ph3); v12 = Negative Peer Nominations (ISO); v13 = Negative Peer Nominations (RS); v14 = ISO Peer Rankings; and v15 = RS Peer Rankings. Correlations have been rounded to the nearest hundredth for display purposes. All CFA models were run using the correlation matrix prior to rounding.

Appendix C

Question One Post Hoc Analyses

Differential results for the basic BE KNOW DO model across the four samples suggested that samples 1 and 3 and samples 2 and 4 might differ on some key demographic, experiential, or measurement characteristic. Examination of Table 9 suggested two possibilities for variables that might have impacted the results across the four samples—the year the participant attended SFQC and the participant's SF MOS. Both possibilities can be argued as plausible. Each year brings in new leaders and cadre with different backgrounds and priorities. Changes in training and/or standards might have been made. These changes would in turn impact measurement and the latent structure of performance. Even if no salient changes were made to SFQC, the differences in assessors (or in the candidates) from year to year may impact the measurement and/or structure of performance (e.g., the frame of reference of the cadre could change). As for MOS, the candidates are allowed to request their MOS of choice. However, the assignments are made based on force requirements (e.g., how many medics are needed in SF), data from the candidate folders, and their MOS preferences. Several factors might be driving MOS-based differences (if they exist). If the candidates assigned to each MOS consistently differ on a key human attribute, interest, or experiential variable, then performance measurement could be impacted. For example, members of the 18D MOS (medics) are assigned on the basis of need, intelligence, and previous training performance because the medical training is very academically rigorous. Therefore, tests of declarative knowledge might not be useful performance indicators for 18D candidates because of restriction of range in that sub-population. Self-selection—to the extent it is allowed to operate in the system—might be a contributing factor to MOS homogeneity. Another example is that officers and noncommissioned officers (NCOs) differ in background and, therefore, the findings could be explained by differences between those two groups.

Finally, the roles performed by members of the various MOS might create quantitatively or qualitatively different opportunities to be observed or evaluated. Also, some measurement instruments might be differentially effective for members of different MOS.

To gain insights into this issue, several post hoc analyses were conducted. Tables C1 through C10 present the descriptive statistics and zero-order correlations for all samples used in the post hoc analyses. First, the final BE KNOW DO model (see Figure 12) was fit to the question one subset (N = 822) and resulted in a highly significant chi-square statistic, 139.63 (32), $p < .0001$, with a chi-square/degrees of freedom ratio of 4.36. The other fit indices suggested the final BE KNOW DO model provided an acceptable fit (CFI = .93; NNFI = .90; SRMSR = .05; RMSEA = .06 [.05-.08]). Table C11 displays the fit indices for the post hoc analyses. Second, two new samples (A and B) were drawn from the question one subset. These samples (n = 410, 412) were drawn to be approximately representative of the overall data set in terms of the "year attended" variable. Table 9 presents the demographic composition of the question one subset and the two samples. Tables C1 and C2 present the descriptive statistics and zero-order correlations for samples A and B. When the model was fit to samples A and B, the fit indices were approximately identical to those for the overall subset (see Table C11). The correlations between the latent variables were similar across the two samples and the overall subset as well. For the overall subset, the BE-KNOW correlation was .56, the BE-DO correlation was .39, and the KNOW-DO correlation was .59. The same correlations for sample A were .55, .36, and .61, respectively. The same correlations for sample B were .55, .42, and .51, respectively. The correlations definitely did not vary as much as the correlations did between the two groups of samples (see Table 13).

Third, to compare the years attended more directly, the question one subset was split into four samples, one for each year in the question one subset. As can be seen in Table C11, only two of the samples (1998, 1999) had enough participants to conduct a CFA. Tables C3 and C4 present the descriptive statistics and zero-order correlations for both samples. The final BE KNOW DO model (Figure 12) was fit to the 1998 (n = 346) and 1999 (n = 367) samples. Fitting the model to the 1998 sample produced a significant chi-square statistic, 59.35 (32), p = .002, with a chi-square/degrees of freedom ratio of 1.85. The other fit indices suggested a good fitting model (CFI = .96; NNFI = .94; SRMSR = .04; RMSEA = .05 [.03-.07]). Fitting the model to the 1999 sample resulted in a very significant chi-square value, 81.39 (32), $p < .0001$, with a chi-square/degrees of freedom ratio of 2.54. The other fit indices suggested an acceptable fit for the model (CFI = .93; NNFI = .90; SRMSR = .06; RMSEA = .07 [.05-.08]). The model provided a better fit for the 1998 sample than for the 1999 sample. Examining the correlations between the latent variable for both samples revealed some differences. The correlations for BE and DO (.36 for 1998; .43 for 1999) were fairly consistent between the samples and with the findings from samples A and B. However, the correlations between BE and KNOW (.65 for 1998; .49 for 1999) and KNOW and DO (.45 for 1998; .71 for 1999) differed between the 1998 and 1999 samples. Interestingly, the difference between the 1998 and 1999 values for the KNOW-DO correlation are reminiscent of the difference between samples 1 and 3 and samples 2 and 4 values for the same parameter. The manifest correlations between the measures of KNOW and DO found in Tables C3 and C4 were examined to gain more insights. For the 1998 sample (Table C3), only three of the nine correlations between the knowledge and skill indicators (e.g., STAR FTX correlated with land navigation exam) were statistically significant (with three

additional ones just below the threshold). For the 1999 sample (Table C4), seven of the nine correlations between knowledge and skill indicators were significant. This finding related to the higher KNOW-DO correlation for the 1999 sample. Tables B2 though B5 were consulted and the correlations between the indicators of knowledge and skill were examined for samples 1 through 4. As suspected, the group consisting of samples 1 and 3 with a very high KNOW-DO correlation (.91, .91) had more significant correlations between knowledge and skill measures. The group consisting of samples 2 and 4 with the lower KNOW-DO correlation (.46, .26) had fewer significant correlations. The composition by "year attended" of the four samples was found not to support the idea that samples 1 and 3 were composed of more members from 1999 and that samples 2 and 4 were composed of more members from 1998.

Fourth, the MOS issue was investigated. The question one subset was divided into five samples, one for each MOS. Tables C5 through C9 present the descriptive statistics and zero-order correlations for the five samples. Table C11 displays the fit indices for each MOS model. When the BE KNOW DO model was fit to the 18A (team leader) sample, it failed to converge. The model converged with no problems for the other MOS resulting in various degrees of fit. Fitting the model to the 18B (weapons sergeant) sample resulted in a non-significant chi-square, 45.55 (32), p = .06, with a chi-square/degrees of freedom ratio of 1.42. The other indices confirmed the model provided a good fit for the 18B sample (CFI = .96; NNFI = .95; SRMSR = .06; RMSEA = .05 [.00-.07]). Fitting the model to the 18C (engineering sergeant), 18D (medical sergeant), and 18E (communications sergeant) samples resulted in indices hovering around the threshold for acceptable fit, usually just below it. The model provided the least acceptable fit for the 18D sample. Interestingly, the correlations

between the latent variables varied greatly between MOS samples, with the correlations from the 18C sample resembling the correlations from the question 1 subset, sample A, and sample B the most. Due to smaller sample sizes, the results for the 18A, 18D, and 18E samples should be viewed with caution. As mentioned above, the model failed to converge for the 18A sample. Based on examining the manifest correlations for the 18A sample (Table C5), the KNOW factor was dropped from the model. The knowledge measures failed to correlate significantly with any other measures. The two-factor BE DO model was fit to the 18A sample resulting in a non-significant chi-square, 19.20 (13), $p = .12$, with a chi-square/degrees of freedom ratio of 1.48. The other fit indices suggested the model provided a good fit for the 18A sample (CFI = .97; NNFI = .95; SRMSR = .04; RMSEA = .05 [.00-.10]). The KNOW factor was confirmed to be problematic for the 18A sample.

Finally, the MOS findings suggested two additional courses of action. First, the model was fit to a sample consisting of only NCOs resulting in a very significant chi-square value, 112.07 (32), $p < .0001$, with chi-square/degrees of freedom ratio of 3.50. Table C10 displays the fit indices for the NCO model. Table C10 presents the descriptive statistics and zero-order correlations for the NCO sample. The model was found to fit the NCO sample very similarly to the way it fit the overall subset and samples A and B. The latent correlations were virtually identical as well (see Table C11). Second, the correlation matrices for the five MOS samples were examined (see Tables C5-C9) to determine the number of significant correlations between the three knowledge variables and the other variables in the model. The MOS as listed from having the most significant correlations to the least were found to be 18C (12), 18B (10), 18E (5), 18D (4), and 18A (0). To recap the post hoc findings, regardless of whether it was a MOS, year, or mixed sample, the significance and consistency of the

correlations between the knowledge measures and other measures (especially knowledge and skill) in the model seemed to impact the model fit and the parameter estimates.

Table C1.

Descriptive Statistics and Zero-Order Correlations for Question One, Sample A

Variables		v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11
M		4.02	4.00	.66	.67	39.31	77.94	86.92	3.46	3.57	4.22	4.46
SD		.56	.41	.40	.41	6.83	13.85	8.64	.34	.25	1.03	.90
N		410	410	410	410	410	410	410	410	410	410	410
Team (Ph1)	v1	--										
Effort (ISO)	v2	.47	--									
ISO Ranking	v3	.47	.64	--								
RS Ranking	v4	.39	.48	.66	--							
ISO Exam	v5	.14	.19	.15	.13	--						
Land Nav Exam	v6	.15	.06	.09	.03	.04	--					
SUT Exam	v7	.28	.24	.30	.15	.21	.13	--				
Negative Spots (Ph1)	v8	.14	.13	.10	.08	.06	.15	.12	--			
Negative Spots (Ph3)	v9	.05	.09	.05	.16	.10	.00	.02	-.02	--		
SUT FTX	v10	.25	.08	.13	.07	.08	.07	.14	.25	.01	--	
STAR FTX	v11	.14	.17	.19	.05	.17	.04	.18	.12	.04	.19	--

Note. Correlations greater than .09 are significant at $p < .05$. This sample was created for a Post Hoc analysis. Correlations have been rounded to the nearest hundredth for display purposes. All CFA models were run using the correlation matrix prior to rounding.

Table C2.

Descriptive Statistics and Zero-Order Correlations for Question One, Sample B

Variables		v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11
M		3.97	3.93	.64	.66	38.33	76.70	86.75	3.46	3.60	4.29	4.39
SD		.52	.41	.41	.43	6.23	14.82	8.29	.37	.19	.98	.97
N		412	412	412	412	412	412	412	412	412	412	412
Team (Ph1)	v1	--										
Effort (ISO)	v2	.44	--									
ISO Ranking	v3	.48	.61	--								
RS Ranking	v4	.37	.48	.67	--							
ISO Exam	v5	.15	.20	.19	.08	--						
Land Nav Exam	v6	.12	.07	.13	.07	.05	--					
SUT Exam	v7	.24	.22	.32	.17	.15	.23	--				
Negative Spots (Ph1)	v8	.09	.09	.04	.04	.02	.18	.03	--			
Negative Spots (Ph3)	v9	.02	.10	.07	.10	.12	-.05	.01	-.03	--		
SUT FTX	v10	.32	.18	.25	.19	.05	.12	.14	.22	.01	--	
STAR FTX	v11	.25	.13	.12	.08	.08	.14	.25	.06	.09	.25	--

Note. Correlations greater than .09 are significant at $p < .05$. This sample was created for a Post Hoc analysis. Correlations have been rounded to the nearest hundredth for display purposes. All CFA models were run using the correlation matrix prior to rounding.

Table C3.

Descriptive Statistics and Zero-Order Correlations for Question One, 1998 Sample

Variables		v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11
M		3.99	3.97	.65	.67	38.48	79.32	86.51	3.58	3.56	4.25	4.32
SD		.55	.43	.41	.43	7.04	13.60	8.18	.23	.27	1.01	1.01
N		346	346	346	346	346	346	346	346	346	346	346
Team (Ph1)	v1	--										
Effort (ISO)	v2	.49	--									
ISO Ranking	v3	.50	.62	--								
RS Ranking	v4	.42	.46	.65	--							
ISO Exam	v5	.15	.24	.17	.12	--						
Land Nav Exam	v6	.13	.04	.12	.08	.09	--					
SUT Exam	v7	.32	.25	.37	.23	.15	.21	--				
Negative Spots (Ph1)	v8	.10	.13	.12	.05	.16	.10	.05	--			
Negative Spots (Ph3)	v9	.02	.09	.08	.20	.13	.00	.06	.05	--		
SUT FTX	v10	.20	.10	.20	.13	.05	.10	.10	.28	.05	--	
STAR FTX	v11	.14	.11	.15	.11	.11	.04	.21	.17	.05	.24	--

Note. Correlations greater than .10 are significant at $p < .05$. This sample was created for a Post Hoc analysis. Correlations have been rounded to the nearest hundredth for display purposes. All CFA models were run using the correlation matrix prior to rounding.

Table C4.

Descriptive Statistics and Zero-Order Correlations for Question One, 1999 Sample

Variables		v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11
M		4.02	3.95	.66	.68	38.92	75.50	87.04	3.37	3.60	4.39	4.51
SD		.53	.41	.41	.40	6.01	14.48	8.73	.42	.18	.91	.87
N		367	367	367	367	367	367	367	367	367	367	367
Team (Ph1)	v1	--										
Effort (ISO)	v2	.44	--									
ISO Ranking	v3	.44	.64	--								
RS Ranking	v4	.33	.48	.65	--							
ISO Exam	v5	.15	.12	.16	.06	--						
Land Nav Exam	v6	.14	.05	.07	-.01	.06	--					
SUT Exam	v7	.24	.25	.30	.12	.21	.17	--				
Negative Spots (Ph1)	v8	.12	.10	.08	.07	.02	.14	.09	--			
Negative Spots (Ph3)	v9	.08	.09	.06	.07	.06	-.01	-.02	-.01	--		
SUT FTX	v10	.27	.14	.18	.11	.12	.11	.18	.25	-.05	--	
STAR FTX	v11	.27	.21	.17	.00	.13	.13	.24	.11	.08	.15	--

Note. Correlations greater than .10 are significant at $p < .05$. This sample was created for a Post Hoc analysis. Correlations have been rounded to the nearest hundredth for display purposes. All CFA models were run using the correlation matrix prior to rounding.

Table C5.

Descriptive Statistics and Zero-Order Correlations for Question One, 18A Sample

Variables		v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11
M		4.26	4.14	.92	.69	42.29	80.48	93.85	3.50	3.55	4.33	4.72
SD		.47	.36	.40	.42	5.50	14.07	5.89	.29	.30	.92	.60
N		163	163	163	163	163	163	163	163	163	163	163
Team (Ph1)	v1	--										
Effort (ISO)	v2	.43	--									
ISO Ranking	v3	.37	.64	--								
RS Ranking	v4	.37	.42	.59	--							
ISO Exam	v5	.13	.04	.03	.09	--						
Land Nav Exam	v6	.08	.07	.03	-.03	-.01	--					
SUT Exam	v7	.08	.06	.02	-.02	-.02	.03	--				
Negative Spots (Ph1)	v8	.07	.17	.15	.04	.07	.08	.08	--			
Negative Spots (Ph3)	v9	-.01	.00	-.03	.15	.00	.02	.06	-.12	--		
SUT FTX	v10	.17	.09	.06	.01	.14	.05	.07	.17	.02	--	
STAR FTX	v11	.00	.04	-.07	-.04	-.02	-.06	.06	-.13	.01	-.06	--

Note. Correlations greater than .15 are significant at $p < .05$. This sample was created for a Post Hoc analysis. Correlations have been rounded to the nearest hundredth for display purposes. All CFA models were run using the correlation matrix prior to rounding. 18A refers to the Special Forces team leader (i.e., officer).

Table C6.

Descriptive Statistics and Zero-Order Correlations for Question One, 18B Sample

Variables		v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11
M		4.03	3.95	.70	.73	37.59	75.70	85.62	3.44	3.60	4.38	4.46
SD		.53	.43	.44	.44	6.54	14.41	7.94	.37	.20	.89	.84
N		205	205	205	205	205	205	205	205	205	205	205
Team (Ph1)	v1	--										
Effort (ISO)	v2	.40	--									
ISO Ranking	v3	.44	.66	--								
RS Ranking	v4	.38	.51	.71	--							
ISO Exam	v5	.09	.14	.12	.13	--						
Land Nav Exam	v6	.13	.05	.17	.10	.02	--					
SUT Exam	v7	.25	.21	.33	.24	.00	.23	--				
Negative Spots (Ph1)	v8	.07	.06	.02	.02	-.04	.11	.00	--			
Negative Spots (Ph3)	v9	-.05	.04	.08	.04	.26	-.07	-.03	.01	--		
SUT FTX	v10	.22	.10	.25	.15	.07	.12	.14	.10	-.04	--	
STAR FTX	v11	.24	.05	.16	.05	-.02	.02	.28	.03	.05	.09	--

Note. Correlations greater than .13 are significant at $p < .05$. This sample was created for a Post Hoc analysis. Correlations have been rounded to the nearest hundredth for display purposes. All CFA models were run using the correlation matrix prior to rounding. 18B refers to the Special Forces weapons sergeant.

Table C7.

Descriptive Statistics and Zero-Order Correlations for Question One, 18C Sample

Variables		v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11
M		3.90	3.90	.57	.71	38.15	76.40	84.31	3.43	3.60	4.24	4.25
SD		.52	.46	.37	.43	6.55	14.63	8.39	.43	.21	1.04	1.11
N		181	181	181	181	181	181	181	181	181	181	181
Team (Ph1)	v1	--										
Effort (ISO)	v2	.53	--									
ISO Ranking	v3	.54	.66	--								
RS Ranking	v4	.41	.57	.73	--							
ISO Exam	v5	.16	.22	.11	.04	--						
Land Nav Exam	v6	.21	.02	.12	.09	.04	--					
SUT Exam	v7	.23	.23	.29	.30	.14	.17	--				
Negative Spots (Ph1)	v8	.19	.10	.14	.16	.06	.19	.12	--			
Negative Spots (Ph3)	v9	.18	.24	.19	.17	.25	.01	.09	.00	--		
SUT FTX	v10	.41	.11	.27	.15	.09	.15	.23	.39	.08	--	
STAR FTX	v11	.31	.21	.23	.16	.23	.07	.08	.17	.13	.37	--

Note. Correlations greater than .14 are significant at $p < .05$. This sample was created for a Post Hoc analysis. Correlations have been rounded to the nearest hundredth for display purposes. All CFA models were run using the correlation matrix prior to rounding. 18C refers to the Special Forces engineering sergeant.

Table C8.

Descriptive Statistics and Zero-Order Correlations for Question One, 18D Sample

Variables		v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11
M		3.93	3.98	.60	.69	38.27	79.12	87.06	3.48	3.65	4.15	4.49
SD		.52	.37	.34	.40	5.98	13.35	7.59	.32	.14	1.12	.84
N		111	111	111	111	111	111	111	111	111	111	111
Team (Ph1)	v1	--										
Effort (ISO)	v2	.44	--									
ISO Ranking	v3	.43	.48	--								
RS Ranking	v4	.30	.44	.76	--							
ISO Exam	v5	.01	.27	.15	.17	--						
Land Nav Exam	v6	.10	-.11	-.07	-.07	-.16	--					
SUT Exam	v7	.17	.05	.04	.13	.06	.13	--				
Negative Spots (Ph1)	v8	-.02	-.11	-.14	-.09	-.20	.24	-.01	--			
Negative Spots (Ph3)	v9	-.03	.10	.19	.17	.23	.06	-.03	-.10	--		
SUT FTX	v10	.29	.17	.13	.07	-.06	.06	.15	.15	.04	--	
STAR FTX	v11	-.03	.02	.02	-.01	.11	.16	.09	.09	.02	.24	--

Note. Correlations greater than .18 are significant at $p < .05$. This sample was created for a Post Hoc analysis. Correlations have been rounded to the nearest hundredth for display purposes. All CFA models were run using the correlation matrix prior to rounding. 18D refers to the Special Forces medical sergeant.

Table C9.

Descriptive Statistics and Zero-Order Correlations for Question One, 18E Sample

Variables		v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11
M		3.81	3.86	.47	.51	38.01	75.99	83.97	3.45	3.56	4.11	4.24
SD		.58	.37	.30	.34	6.80	14.43	8.08	.31	.21	1.10	1.07
N		162	162	162	162	162	162	162	162	162	162	162
Team (Ph1)	v1	--										
Effort (ISO)	v2	.30	--									
ISO Ranking	v3	.30	.44	--								
RS Ranking	v4	.38	.38	.68	--							
ISO Exam	v5	.05	.10	.09	.13	--						
Land Nav Exam	v6	.05	.13	.07	.07	.11	--					
SUT Exam	v7	.08	.09	.11	.00	.18	.10	--				
Negative Spots (Ph1)	v8	.15	.26	.02	.07	.16	.19	.02	--			
Negative Spots (Ph3)	v9	.17	.13	.13	.13	-.01	-.08	.12	.07	--		
SUT FTX	v10	.25	.16	.11	.18	.02	.06	.05	.32	-.04	--	
STAR FTX	v11	.10	.14	-.01	.01	.09	.19	.19	.08	.07	.27	--

Note. Correlations greater than .15 are significant at $p < .05$. This sample was created for a Post Hoc analysis. Correlations have been rounded to the nearest hundredth for display purposes. All CFA models were run using the correlation matrix prior to rounding. 18E refers to the Special Forces communications sergeant.

Table C10.

Descriptive Statistics and Zero-Order Correlations for Question One, NCO Sample

Variables		v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11
M		3.93	3.92	.59	.66	37.96	76.54	85.10	3.45	3.60	4.24	4.35
SD		.54	.42	.38	.42	6.51	14.32	8.10	.37	.20	1.03	.98
N		659	659	659	659	659	659	659	659	659	659	659
Team (Ph1)	v1	--										
Effort (ISO)	v2	.42	--									
ISO Ranking	v3	.45	.59	--								
RS Ranking	v4	.39	.50	.72	--							
ISO Exam	v5	.08	.17	.10	.10	--						
Land Nav Exam	v6	.12	.04	.10	.06	.02	--					
SUT Exam	v7	.20	.18	.24	.20	.09	.17	--				
Negative Spots (Ph1)	v8	.11	.09	.03	.06	.02	.17	.04	--			
Negative Spots (Ph3)	v9	.09	.15	.14	.13	.18	-.02	.06	.01	--		
SUT FTX	v10	.30	.13	.21	.15	.04	.10	.15	.24	.01	--	
STAR FTX	v11	.19	.13	.14	.08	.10	.10	.17	.10	.09	.26	--

Note. Correlations greater than .07 are significant at $p < .05$. This sample was created for a Post Hoc analysis. Correlations have been rounded to the nearest hundredth for display purposes. All CFA models were run using the correlation matrix prior to rounding. NCO refers to non-commissioned officers (i.e., all members of the 18B, 18C, 18D, and 18E MOS).

Table C11.

Fit Indices for Question One Post Hoc Models

Model	*N*	X^2	*df*	X^2/df	GFI	CFI	NNFI	NFI	SRMSR	RMSEA
Sampling Issue										
All Question One Data	822	139.63 (<.0001)	32	4.36	.97	.93	.90	.91	.05	.06 (.05-.08)
Sample A	410	79.03 (<.0001)	32	2.47	.96	.94	.91	.90	.05	.06 (.04-.08)
Sample B	412	86.95 (<.0001)	32	2.70	.93	.93	.90	.89	.05	.06 (.05-.08)
1997 Sample	62	Insufficient sample size for modeling.								
1998 Sample	346	59.35 (.002)	32	1.85	.97	.96	.94	.92	.04	.05 (.03-.07)
1999 Sample	367	81.39 (<.0001)	32	2.54	.96	.93	.90	.89	.06	.07 (.05-.08)
2000 Sample	47	Insufficient sample size for modeling.								
MOS										
18A	The BE KNOW DO model would not converge for 18A. Investigation of the manifest correlation matrix revealed the measures (i.e., tests) that comprise the KNOW factor were uncorrelated. A two-factor (BE DO) model was confirmed.									
	163	19.20 (.12)	13	1.48	.97	.97	.95	.91	.04	.05 (.00-.10)
	Latent Variable Correlations			BE and DO	.28					
18B	205	45.55 (.06)	32	1.42	.96	.96	.95	.89	.06	.05 (.00-.07)
	Latent Variable Correlations			BE and KNOW BE and DO KNOW and DO	.52 .60 .91					
	Deviations	Phase One Negative Spot Reports and Phase Three Isolation exam have non-significant loadings in the 18B model.								

Table C11 (continued).

Model	*N*	X^2	*df*	X^2/df	GFI	CFI	NNFI	NFI	SRMSR	RMSEA
18C	181	73.90 (<.0001)	32	2.31	.93	.91	.87	.85	.07	.09 (.06-.11)
	Latent Variable Correlations			BE and KNOW		.58				
				BE and DO		.39				
				KNOW and DO		.54				
18D	111	57.73 (.0035)	32	1.80	.91	.86	.80	.74	.09	.09 (.05-.12)
	Latent Variable Correlations			BE and KNOW		-.29				
				BE and DO		-.15				
				KNOW and DO		.79				
	Deviations		Phase One SUT Exam, SUT Times to Criterion, and STAR Times to Criterion have non-significant loadings in the 18D model. Isolation exam has an inverse relationship with other exams in the 18D model.							
18E	162	53.10 (.01)	32	1.66	.94	.90	.86	.80	.07	.07 (.03-.09)
	Latent Variable Correlations			BE and KNOW		.29				
				BE and DO		.27				
				KNOW and DO		.37				
NCO Model	659	112.07 (<.0001)	32	3.50	.98	.94	.91	.91	.05	.06 (.05-.07)
	Latent Variable Correlations			BE and KNOW		.50				
				BE and DO		.36				
				KNOW and DO		.53				

Note. This is the same model utilized in Figure 13. 18A = Special Forces (SF) team leader; 18B = SF weapons sergeant; 18C = SF engineering sergeant; 18D = SF medical sergeant; and 18E = SF communications sergeant. NCO refers to non-commissioned officer. GFI = goodness-of-fit index; CFI = comparative fit index; NNFI = non-normed fit index; NFI = normed fit index; SRMSR = standardized root mean square residual; RMSEA = root mean square error of approximation estimate 90% confidence interval shown in parentheses. The *p* values for the X^2 are reported in parentheses below the statistic's value.

Appendix D

Question Two Post Hoc Analyses

The main goal of research question two was to understand training performance across time. The models in Figures 19 through 24 provided insights into how all three of the BE KNOW DO model constructs are related between Phase One and Phase Three. To glean additional understanding, a series of post hoc models were tested. These models deconstructed the BE KNOW DO model into a series of one-construct and two-construct models. Although testing the entire BE KNOW DO model resulted in valuable information, it was thought that exploring the constructs independently and in pairs would provide a more complete picture of the relationship over time.

Table D1 presents the fit indices for the post hoc models, and Table D2 displays the standardized parameter estimates for the one-construct models. No diagrams are provided; however, Figures 23 and 24 can be used to visualize the models by covering the two non-represented constructs. For the most part, the one-construct models achieved good to excellent fit. The KNOW construct model provided the best fit with a non-significant chi-square value for both samples, .93 (2), p = .63, and .25 (2), p = .88, for samples 1 and 2 respectively. The other fit indices suggested excellent fit for the KNOW model over time as well. For DO across time, fitting the model to both samples resulted in a non-significant chi-square value, 1.50 (20), p = .47, for sample 1 and a significant chi-square value, 10.15 (2), p = .0063, for sample 2. The other fit indices confirmed that the model was an excellent fit for sample 1. The other fit indices were mixed for sample 2. The chi-square/degrees of freedom ratio (5.08), the SRMSR (.07), and the RMSEA (.12) were all beyond their thresholds for good fit for the model on sample 2. The DO consisted of two manifest indicators for each phase. This might have made the RMSEA more sensitive to misspecification of indicators. Fitting the BE model to both samples resulted in very significant chi-square statistics, 33.42

(8), $p < .0001$, and 31.34 (8), $p < .0001$, for samples 1 and 2 respectively. The chi-square/degrees of freedom ratios were 4.18 and 3.98 respectively. However, the other fit indices suggested an excellent fit with the exception of the RMSEA, which was beyond the acceptable fit threshold for both samples (.10 [.07-.15], .10 [.07-.14]). The RMSEA, which is sensitive to indicator misspecification, might have been reacting to an unspecified factor because the LaGrange multipliers suggested correlating the error terms of several indicators. Taken collectively, the one-construct models provided good fit for the data. Interestingly, the results for the DO construct were the most mixed with the model demonstrating excellent fit on one sample and acceptable fit on the other.

Table D3 displays the standardized parameter estimates for the two-construct models. No diagrams are provided; however, Figures 23 and 24 can be used to visualize the models by covering the one non-represented construct. The BE-KNOW model was fit to both samples and provided good fit with the exception of significant chi-squares. The BE-DO model was fit to both samples and provided a poor for the data. Interestingly, correlating the disturbance terms as done with the final predictive model (see Figures 23 and 24) dramatically improved the model fit. The KNOW-DO model was fit to both samples and provided mixed results. The chi-square, the chi-square/degrees of freedom ratio, the SRMSR, and the RMSEA were all at or beyond their thresholds for acceptable fit. The CFI and NNFI indicated acceptable to good fit. Interestingly, the predictive relationship across the phases changed for KNOW and DO when they were paired together versus paired with BE. However, the predictive relationship was very consisted between Phase One BE and Phase Three BE regardless of whether it was paired with KNOW (.33, .33 for samples 1 and 2) or DO (.34, .35 for samples 1 and 2). The results of the two-construct models suggest that the

DO factor influences model fit when combined with the other two. The fact that the model fit improved greatly when the phase three disturbance terms for BE and DO were correlated suggests the problem was with the indicators of DO and BE sharing variance for an unspecified construct.

Table D1.

Fit Indices for Question Two Post Hoc Models

Model	N	X^2	*df*	X^2/df	GFI	CFI	NNFI	NFI	SRMSR	RMSEA
Be1—Be3										
Sample 1	279	33.42	8	4.18	.96	.99	.97	.98	.03	.10 (.07-.15)
Sample 2	279	31.84	8	3.98	.96	.99	.98	.98	.03	.10 (.07-.14)
Know1—Know3										
Sample 1	279	.93 (.63)	2	.47	1.00	1.00	1.06	.98	.02	.00 (.00-.09)
Sample 2	279	.25 (.88)	2	.13	1.00	1.00	1.06	1.00	.01	.00 (.00-.06)
Do1—Do3										
Sample 1	279	1.50 (.47)	2	.75	1.00	1.00	1.00	1.00	.03	.00 (.00-.11)
Sample 2	279	10.15 (.0063)	2	5.08	.98	.99	.96	.98	.07	.12 (.05-.20)
Be/Know1—Be/Know3										
Sample 1	279	54.40 (.008)	32	1.70	.96	.99	.98	.97	.04	.05 (.03-.07)
Sample 2	279	79.91	32	2.50	.95	.98	.97	.96	.06	.07 (.05-.09)
Know/Do1—Know/Do3										
Sample 1	279	53.76	19	2.83	.95	.94	.92	.92	.09	.08 (.06-.10)
Sample 2	279	56.81	19	2.99	.95	.95	.92	.92	.10	.08 (.06-.11)

Table D1 (continued).

Model	N	X^2	df	X^2/df	GFI	CFI	NNFI	NFI	SRMSR	RMSEA
Be/Do1—Be/Do3										
Sample 1	279									
Model		340.44	32	10.6	.86	.88	.84	.87	.21	.17 (.17-.20)
Model with Correlated Disturbances		76.19	31	2.46	.95	.98	.98	.97	.03	.07 (.05-.09)
Sample 2	279									
Model		401.67	32	12.6	.85	.87	.82	.86	.23	.20 (.19-.22)
Model with Correlated Disturbances		73.40	31	2.37	.95	.99	.98	.97	.04	.07 (.05-.09)

Note. Fit indices are presented for post hoc models. Each construct in Phase One predicts its corresponding construct in Phase Three. No diagrams are provided but Figures 19 through 24 can be used to visualize the models by eliminating the non-represented constructs. Unless otherwise noted in parentheses, all X^2 values in the table are significant at $p < .0001$.

Table D2.

Standardized Parameter Estimates for Question Two Post Hoc One-Construct Models

Models	*N*	Standardized Parameter Estimates by Phase						
		Phase 1			Phase 3			
BE								
		Effort 1	Social 1	Team 1	Effort 3	Social 3	Team 3	BE1-BE3
Sample 1	279	.93	.85	.99	.95	.86	.98	.33
Sample 2	279	.95	.86	.98	.94	.87	.98	.33
KNOW								
		Land Navigation	Small Unit Tactics	Isolation	Comprehensive	KNOW1-KNOW3		
Sample 1	279	.15	1.00	.18	.65	.61		
Sample 2	279	.21	1.00	.37	.70	.59		
DO								
		STAR Exercise	SUT Exercise	Tactic 3	Leader 3	DO1-DO3		
Sample 1	279	.23	1.00	.99	.93	.26		
Sample 2	279	.28	.99	.98	.95	.16		

Note. Standardized parameter estimates are reported for post hoc one-construct predictive models. Each construct in Phase One predicts its corresponding construct in Phase Three. No diagrams are provided but Figures 19 through 24 can be used to visualize the models by eliminating the other two (non-represented) constructs. All parameters are significant at $p < .05$.

Table D3.

Standardized Parameter Estimates for Question Two Post Hoc Two-Construct Models

Manifest Indicators by Construct and Phase	Standardized Parameter Estimates					
	BE KNOW		KNOW DO		BE DO	
	Sample 1	Sample 2	Sample 1	Sample 2	Sample 1	Sample 2
BE						
Phase One						
Effort	.95	.93	-	-	.94	.95
Social	.86	.85	-	-	.85	.86
Team	.98	.99	-	-	.98	.97
Phase Three						
Effort	.94	.95	-	-	.95	.94
Social	.87	.86	-	-	.86	.87
Team	.98	.98	-	-	.98	.98
KNOW						
Phase One						
Land Navigation	.25	.23	.15	.21	-	-
Small Unit Tactics	.83	.68	.99	.99	-	-
Phase Three						
Isolation	.37	.19	.18*	.37	-	-
Comprehensive	.70	.60	.65	.70	-	-
DO						
Phase One						
STAR	-	-	.24	.29	.28	.46
SUT	-	-	.98	.97	.56	.39
Phase Three						
Tactic	-	-	.98	.98	.96	.96
Leader	-	-	.93	.95	.95	.96
Negative Spots	-	-	.08	.25	.10	.25
Correlations						
BE1—KNOW1	.37	.25	-	-	-	-
BE1—DO1	-	-	-	-	.70	.71
KNOW1—DO1	-	-	.15	.20	-	-
BE1—BE3	.33	.33	-	-	.34	.35
KNOW1—KNOW3	.76	.94	.61	.59	-	-
DO1—DO3	-	-	.27	.18	.51	.52

Note. Standardized parameter estimates are reported for post hoc one-construct predictive models. Each construct in Phase One predicts its corresponding construct in Phase Three. Figures 19 through 24 can be used to visualize the models.
*$p > .05$; All parameters are significant at $p < .05$.

Appendix E

Question Three Post Hoc Analyses

To gain more insight into the poor fit of the BE models, a series of two-construct post hoc models were tested. Given the correlations between the three factors, it was suspected that the personal discipline factor might not be a part of BE, especially since two of the Personnel Discipline indicators (negative spot reports) were primarily skill indicators (i.e., the negative spots are given more for skill failures than personal discipline failures, although they are given for both). Plus, the negative spot reports indicators failed to performance well in other models. Table E1 presents the fit indices for these models along with the correlation between the two latent variables in each model. No diagrams are presented. The Effort-Personal Discipline model provided mixed results in terms of fit indices across the two model versions and samples (e.g., the second model fit to sample 2 resulted in excellent fit) but yielded very consistent correlations between the two factors ranging from .81 to .83 across the four analyses. The Team-Personal Discipline model provided poor fit across all four analyses and consistent correlations between the latent variables across the initial model (.81, .82) and the second model (.51, .50) for both samples. The Effort-Team model resulted in very poor fit indices. However, the correlation between the constructs was very consistent across the four analyses (1.04, 1.03, .96, .98). The team and effort correlations suggest they are measuring the same construct.

Table E1.

Fit Indices for Post Hoc Two-Factor Models Exploring the Dimensionality of BE

Model	N	X^2	*df*	X^2/df	ΔX^2	GFI	CFI	NNFI	NFI	SRMSR	RMSEA
Effort & Personal Discipline											
Sample 1	340										
Initial Model		264.56	26	10.18		.86	.79	.71	.77	.07	.17 (.15-.18)
					Relationship between effort and personal discipline = .81						
Second Model		54.58	13	4.20	209.98	.95	.92	.87	.90	.05	.10 (.07-.13)
					Relationship between effort and personal discipline = .81						
Sample 2	345										
Initial Model		231.11	26	8.89		.87	.79	.71	.78	.07	.17 (.13-.17)
					Relationship between effort and personal discipline = .82						
Second Model		24.65 (.0256)	13	4.20	206.46	.98	.98	.96	.95	.04	.05 (.02-.08)
					Relationship between effort and personal discipline = .83						
Team & Personal Discipline											
Sample 1	340										
Initial Model		833.80	34	24.52		.70	.58	.45	.58	.12	.26 (.24-.28)
					Relationship between team and personal discipline = .81						
Second Model		425.44	19	22.39	408.36	.78	.64	.46	.63	.14	.25 (.23-.27)
					Relationship between team and personal discipline = .51						
Sample 2	345										
Initial Model		816.88	34	24.03		.72	.59	.46	.59	.12	.26 (.24-.27)
					Relationship between team and personal discipline = .82						
Second Model		520.08	19	27.37	296.80	.75	.58	.38	.58	.16	.28 (.26-.30)
					Relationship between team and personal discipline = .50						

Table E1 (continued).

Model	N	X^2	*df*	X^2/df	ΔX^2	GFI	CFI	NNFI	NFI	SRMSR	RMSEA
Team & Effort											
Sample 1	340										
Initial Model		2081.00	43	48.40		.48	.49	.35	.49	.15	.37 (.36-.39)
					Relationship between team and effort = 1.04						
Second Model		570.21	13	43.86	1510.79	.69	.71	.53	.70	.19	.36 (.33-.38)
					Relationship between team and effort = .96						
Sample 2	345										
Initial Model		1947.06	43	42.33		.50	.50	.36	.50	.17	.36 (.35-.37)
					Relationship between team and effort = 1.03						
Second Model		684.47	13	50.34	1292.59	.66	.67	.47	.67	.21	.38 (.35-.40)
					Relationship between team and effort = .98						

Note. In order to understand why question three (the dimensionality of BE) was not successful, a series of post hoc two-factor BE models were tested. No diagrams are presented for these models. However, Figures 11, 12, 25 and 26 can be used to visualize the models by covering up the missing latent variable. GFI = general fit index; CFI = comparative fit index; NNFI = non-normed fit index; NFI = normed fit index; SRMSR = standardized root mean square residual; RMSEA = root mean square error of approximation estimate 90% confidence interval shown in parentheses. Unless otherwise noted, all X^2 values in the table are significant at $p < .0001$.