ABSTRACT

KOOMMOO-WELCH, PENNY. Measurement Invariance in Performance Appraisal Ratings of US Army Special Forces Soldiers. (Under the direction of Mark A. Wilson.)

The purpose of the present study was to examine the equivalence of mental models of job performance between rater groups among US Army Special Forces Soldiers. Performance appraisals are often completed in organizations by individual raters, whose ratings are then compared to one another in order to make inferences of the ratee's performance on the job. Disagreements in ratings can lead to erroneous conclusions unless it is first established that comparisons between the rater groups are appropriate. Ratings of soldiers by two supervisory rater groups (N = 1052 and N = 910) on an appraisal instrument designed specifically for the Special Forces were examined. An exploratory factor analysis of the ratings indicated a four-factor model, which was then subsequently used to test for measurement invariance between the rater groups using multiple-group confirmatory factor analysis and item response theory. Fit indices indicated reasonable fit of the model, and ratings were concluded to be invariant at the rater group level of analysis, indicating that the rater groups refer to similar mental models of performance when rating individual soldiers.

Measurement Invariance in Performance Appraisal Ratings of

US Army Special Forces Soldiers

by

PENNY KOOMMOO-WELCH

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

PSYCHOLOGY

Raleigh

2005

APPROVED BY:

_____

M. A. Wilson, Chair

_____ _____

D. W. Drewes S. B. Craig

_____

M. G. Sanders

BIOGRAPHY

Penny Koommoo-Welch was born on December 27, 1977 in Albany, Georgia to parents Anan and Penpit Koommoo. She graduated from Lovejoy High School in 1996, and after attending the Georgia Institute of Technology her freshman year of college, she transferred to Emory University in Atlanta, where she graduated with a Bachelor of Arts in Psychology in 2000. Not wanting to lose momentum, Penny applied to and was accepted into the Industrial/Organizational Psychology program that same year at North Carolina State University in Raleigh, NC, where she resides today working towards her doctorate.

Penny has completed internships with Corporate Insights and Development, a human resources consulting firm in Atlanta, Georgia and global pharmaceuticals company GlaxoSmithKline in Research Triangle Park, North Carolina. She has served as a laboratory instructor for undergraduate psychology courses in Research Methods and Statistics for 4 years, and is active in campus politics as Vice-President of the Graduate Association of Students in Psychology as well as acting President of the University Graduate Student Association of North Carolina State University. She was recently married to Ryan S. Welch, her high school sweetheart of 9 years, who is currently serving in Iraq with the 1-41 Field Artillery Division of the US Army.

ACKNOWLEDGMENTS

First, I would like to thank my committee members, Dr. Mark Wilson, Dr. Bart Craig, Dr. Don Drewes, and Dr. Mike Sanders, for their incredible support and patience in this marathon endeavor of a thesis – I know I probably made each of them crazy at some point in the long process, and sincerely thank each of them for not giving up on me or this project.

Many thanks to Dr. Jat Thompson for his patience in answering my questions in the early stages of this project and beyond, and Dr. Adam Meade for helping me through occasional moments of panic while dealing with baffling software and confusing output. Thanks also to Dr. Tommy Powell, for his support, encouragement, and attempts to make sure I wasn't neglecting my thesis in favor of other pursuits. I would also like to thank Dr. Becky Rufty, Assistant Dean of the Graduate School, for her words of encouragement and support as I juggled various activities, and Dr. Don Mershon for his help and actions on my behalf when the Graduate School Handbook was less than crystal clear.

I would also like to thank my parents, Anan and Penpit Koommoo, and my parents-in-law, Rick and Vicky Welch, for their love, support and frequent calls to make sure that I was getting enough to eat, dressing warmly, and generally not making myself sick. I'd like to extend special thanks to my 'partners-in-crime' in the Psychology Department: Christina Costanzo Mendat, Kari Yoshimura, Becca Baker, and Tara Shetye for their encouragement and 'cheerleading,' and for humoring me when I became just a bit too melodramatic. I also want to thank my 'girls from home': Val Weisser, Jennifer Osse, and Heather Dean, for always being supportive in spirit.

Last but not least, I'd like to give special loving thanks my husband and best friend Ryan, for his support, patience, and occasional 'kick in the pants' when I was getting too comfortable as a penniless graduate student – I miss you, and I couldn't have finished without you.

TABLE OF CONTENTS

LIST OF TABLES

Measurement Invariance in Performance Appraisal Ratings of US Army Special Forces Soldiers

The use of individual ratings as measures of job performance is by no means a new concept. The performance appraisal process for both administrative and developmental purposes is not only commonplace, but also often expected in modern organizations (Murphy, Cleveland, & Mohler, 2001). Historically, supervisory ratings have been of primary interest. Ratings from supervisors are less disputed and appear to be perceived as more valid than those from peers or subordinates, primarily due to the assumption that supervisors' in-role behaviors should include observing and evaluating individuals, whereas others have no such similar responsibility (e.g., subordinates' job descriptions do not typically include evaluation of their superiors; Murphy at al., 2001). Despite this, research has highlighted a number of problems with basing recommendations on supervisory ratings alone, including errors of halo, bias, and leniency (e.g., Harris & Schaubroeck, 1988; Conway & Huffcutt, 1997). There is also some evidence that supervisors rely heavily on outcomes rather than actual performance behaviors when evaluating subordinates, thus calling into question the accuracy of performance ratings captured after the behavior has occurred (Carson, Cardy, & Dobbins, 1991).

In response to these issues, the traditional supervisory rating system has been supplemented by one in which individuals are evaluated by multiple sources in addition to the supervisor. The most common variant of this multiple source feedback (MSF) system is known as 360-degree feedback, in which individuals (the ratees) are evaluated by several different sources (rater groups), which generally include any number of supervisors, peers, direct subordinates, as well as the ratees themselves. The implication of such a system is that an all-inclusive, total (360-degree) view of performance will be more comprehensive than an evaluation by any single individual (e.g., Farr & Newman, 2001; Harris & Schaubroeck, 1988). By one definition, the intent of MSF systems is to obtain assessments from multiple raters who are each witness to different facets of an individual's

performance, and thus are able to make evaluations based on different work behavior information (Farr & Newman, 2001). The increasing use of multiple source feedback in industry has naturally raised the issue of whether ratings obtained from these different sources are in fact comparable and valid (Bracken, Timmreck, & Church, 2001). While some studies have shown supervisory ratings to have consistently higher reliability than peer or self ratings (Viswesvaran, Ones, & Schmidt, 1997; Rothstein, 1990), other analyses of MSF have provided some evidence that ratings will most likely disagree, regardless of the organizational level of the raters, and even within the same organizational level (Murphy at al., 2001; Greguras & Robie, 1998). This general disagreement in ratings between raters has resulted in controversy over the usefulness of multiple source ratings, and several potential reasons have been suggested. These include, among others, differences between raters in their opportunities to observe target behaviors (Murphy & Cleveland, 1995), individual rater bias and inaccurate recall of behaviors (Wherry & Bartlett, 1982), and differences in rater frame-of-reference (FOR; Woehr & Huffcutt, 1994). Thus, ratings of performance may in the end be based less on actual ratee behavior and more on the individual rater herself (Scullen, Mount, & Goff, 2000; Wherry & Bartlett, 1982).

It has further been suggested that perhaps this problem of disagreement between rater sources actually lies in the rating instrument itself (Facteau & Craig, 2001; Scullen et al., 2000). That is, performance dimensions represented by a rating form may not translate equivalently from one rater to the next, in effect changing the nature of the instrument. These different conceptualizations, or mental models, of the performance dimensions being evaluated may then manifest themselves as disagreement in the ratings of any single individual. It has therefore been argued that, in order to compare the ratings of any two raters from different rater groups on a single ratee, it must be assumed not only that the ratings from each rater exist on the same measurement scale, but that the relationships between the indicators (i.e., items) and the constructs they are intended to measure are

equivalent across rater groups (Reise, Widaman, & Pugh, 1993). This procedure to describe the equivalency of an instrument from one individual or group to the next is typically referred to as measurement invariance (MI; or measurement equivalence, ME) and until recently was widely discussed, but rarely tested directly (Facteau & Craig, 2001; Maurer, Raju, & Collins, 1998). In structural equation modeling terms, measurement invariance of an instrument shows that the relationships between the indicators (items) and latent constructs they represent are stable across individual raters. An instrument that demonstrates such measurement invariance across raters provides support for the hypothesis that those raters completing the form do in fact share the same mental model of the dimensions being rated, and only after this step may we then compare ratings between different rater groups and speculate on why differences occur (Vandenberg & Lance, 2000).

Two methods of testing for measurement invariance have been most frequently suggested: confirmatory factor analyses (CFA) and item response theory (IRT). The use of CFA to determine equivalence of performance models appears most common; in their recent meta-analysis, Vandenberg and Lance (2000) reviewed 81 studies that utilized CFA as the primary means to establish measurement invariance. Exclusive use of IRT in the examination of ratings is far less common, though it is becoming more prevalent (Craig & Kaiser, 2003a; Barr & Raju, 2003). Over the last few years, researchers have taken to using both CFA and IRT in conjunction to establish measurement invariance, as each method supplies somewhat different information regarding rating scale characteristics (Maurer et al., 1998). Specifically, the strength of CFA lies in its evaluation of invariance between performance dimensions, while IRT (and its requirement of unidimensionality) is more specific in its evaluation of MI of actual items within a specific performance dimension. Maurer et al. (1998) and Facteau and Craig (2001) have recently used both methods to specifically compare ratings between rater groups, with the former finding support for measurement invariance in the ratings of managers by peers and subordinates, and the latter establishing MI in manager ratings

across self, peer, supervisor, and subordinate rater groups. Thus CFA and IRT can be used in conjunction to establish the measurement equivalence of a ratings form across two groups. The next sections will discuss, briefly, the rationale for testing MI and the proposed analyses (by supervisory level, rater tenure, and location). The study will then be put into context with a description of the unique organization of interest, and research questions will be addressed.

Rater agreement and measurement invariance. One critical assumption of a multiple source feedback system is that each rater is exposed to a unique (and oftentimes different) perspective of the ratee's performance, and that in order for their ratings to be considered legitimate by ratees, raters must have sufficient opportunity to observe the behaviors being evaluated (Farr & Newman, 2001; Murphy et al., 2001). This suggests that ratings from different rater groups collected as part of a MSF system are, in fact, not expected to agree, and researchers have used this to dispel the common conclusion that ratings which disagree are invalid (e.g., Lance & Bennett, 1997; Conway & Huffcutt, 1997). However, it must be clarified that there is a distinct difference between *agreement* in ratings between rater groups and *measurement invariance* between rater groups – disagreement between rater sources may prove valuable for developmental feedback purposes, but such comparisons cannot logically be made until it can be shown that the rater groups in question exhibit measurement invariance; that is, that they reference the same mental model (and scale structure) of performance when making ratings. Unless this step to show measurement invariance is undertaken, it cannot be determined whether disagreement in ratings is due to those factors discussed previously, such as opportunity to observe or bias, or due to the fact that raters possess different conceptualizations of performance (Viswesvaran, Schmidt, & Ones, 2002). Only after it is established that such measurement invariance exists between rater groups can we reasonably compare ratings between groups, and assess possible sources of disagreement. The present study was an attempt to determine

not whether the rater groups showed agreement in their ratings of subordinates, but whether their mental models of performance could be considered equivalent.

Supervisory levels. While some research maintains that ratings between rater sources will most likely disagree regardless of organizational level (e.g., Murphy et al., 2001; Greguras & Robie, 1998), other research examining multiple supervisory ratings of subordinates have found these ratings to show moderate agreement (Harris & Schaubroeck, 1988; Conway & Huffcutt, 1997). In their study of ratings convergence between organizational levels, Viswesvaran, Schmidt, and Ones (2002) suggested that disagreement in ratings between a ratee's peers and supervisors might be a result of his peers and supervisors rating somewhat different dimensions of job performance, due in part to perceived differences in the nature of the dimensions (what they termed 'construct-level divergence'). This notion, that different groups of raters may perceive and categorize job performance into different dimensions, can be extended in the present study to different types of raters within a single organizational level, insofar as their relationships to the ratee differ enough to allow them possibly unique perspectives of performance. It was through this rationale that the present study examined measurement invariance between rater groups at two different supervisory levels.

Only one study was found that examined the equivalence of a rating form between different supervisory levels. Using CFA and IRT methods to examine the archival performance ratings of peers at various organizational levels, the authors found MI to exist in the feedback instrument across all 3 rater groups, regardless of organizational level (Craig & Kaiser, 2003b). A similar study by the same authors examining measurement invariance in a large telecommunications company found similar results: equivalence was found in the instrument across several levels of top executives (S.B. Craig, personal communication, 10 November 2004). Although the authors concede to certain limitations within their studies (e.g., empirically, rather than theoretically, derived factors), the findings demonstrate that examination measurement invariance of the rating instrument has potential

in helping to examine agreement between different rater groups. In both cases, the rating form was found to be invariant across peer ratings. The present study examined measurement invariance between two rater groups whose differential relationships and interaction with ratees may result in different mental models of performance by which ratees are evaluated.

Rater tenure and experience. Although most research tends to focus on performance ratings at a global rater source level (such as by supervisor; Murphy, Cleveland, & Mohler, 2001), there is some evidence that other factors may play a role. It could be argued that raters with more experience in a particular position gradually learn how to appropriately evaluate individuals, suggesting that their ratings should be more informed than ratings by novices in the same position (an implication being that their mental models of performance become more refined). Evidence of the contribution of rater tenure/experience has been mixed, with some studies reporting a positive relationship between rater experience and ratings (Tesluk & Jacobs, 1998; Landy & Farr, 1980), and others finding rater tenure to be unrelated to ratings of others (Brutus, Fleenor, & McCauley, 1999; Judge & Ferris, 1993). It is unclear, however, whether experience/tenure of the rater can influence the measurement invariance of a rating form above and beyond that of rater group membership alone. Therefore the present study examined the possible influence of tenure/experience of raters only in cases where MI was not found at the supervisory level alone.

Location. There does not appear to be much research focusing on the performance of teams in different geographical locations, and what does exist seems to be rooted in examinations of "geographically dispersed teams" in which a single workgroup is scattered across two or more locations while working toward a single, common goal (Cramton, 2001). By contrast, "location" is used in the present context in the examination of different teams assigned to different locations, who work toward possibly different goals. This is an important distinction in that the present case suggests a possible difference in mental models of performance dependent on the specific mission

and performance requirements as determined by geographical assignment. That is, it could be argued that although all ratees are evaluated on a specific performance dimension (e.g., language proficiency), 'acceptable performance' on this dimension may differ depending on whether the ratee is expected to be proficient in a single language versus multiple languages. Thus geographical differences may translate into different mental models of performance, which would make comparisons between these different locations problematic unless measurement invariance at this level could be established. A recent study did find MI in a rating form across raters in different countries (that is, MI was tested and found across raters operating in the Unites States and raters in various other countries; S.B. Craig, personal communication, 18 April 2004), but aside from this study, there do not appear to be any other published examinations of geographical location as a variable influencing performance ratings in the context used here. The present study examined the possible influence of rater location only in cases where MI was not found at the supervisory level alone.

Models of performance. Several models of job performance have been proposed throughout the years, and the once pervasive notion of a single dimension of performance has recently given way to a general consensus among researchers that performance is more likely a multidimensional factor (Thompson, 2002). The rating instrument used in the present study was itself created based on a model of performance consisting of 3 factors: ***know***, ***do***, and ***extrarole*** (or ***be***; Grant, 1996), which share similarities with other proposed models in describing general knowledge, job-specific performance, and contextual performance, respectively (e.g., Borman & Motowidlo, 1993; Campbell, 1990). Items in the rating instrument were derived using a critical incidents technique, and described examples of each factor: Soldiering Skills (***know***), SF Specific Skills (***do***), and Team Member Skills (***be***; Wilson, Drewes, Cunningham, Sanders, Thompson, & Surface, 2001). Although this model was used to create the instrument, a decision was made against using the 3-factor model in the present

analysis for two primary reasons. First, as an exploratory study of the mental models of performance in a specific sample, it was initially believed that using the a priori model might be too restrictive in that the model may describe more how the instrument was created than how the raters appear to actually be using it. That is, it was believed that the practical utility of the analysis would be increased if analyses were performed on the models that were actually being used by raters. Second, it was speculated that there existed a better probability of establishing measurement invariance between rater groups if the groups themselves were allowed to determine the model. For these reasons, it was decided to forego the use of the a priori 3-factor measurement model and instead, a new measurement model was established by performing an exploratory factor analysis (EFA) on the combined sample (all ratings across both rater groups).

Study Context: United States Army Special Forces (SF)

Prior to presenting the formal research questions it will be necessary to provide more detail on the subject population studied. Data for this study were collected from United States Army Special Forces (USASF, or SF) soldiers who were currently serving on teams. Typically, an SF team, or Occupational Detachment Alpha (ODA), is comprised of a team leader (TL), team sergeant (TS), and up to ten additional team members. As commissioned officers, team leaders are the highest-ranking individuals on the team and are in command of the ODA. Team sergeants and team members are all noncommissioned officers (NCOs), and the team sergeant is the highest in rank among them. Before they are assigned to an ODA, potential team members must successfully complete two major hurdles: Special Forces Assessment and Selection (SFAS) and the Special Forces Qualification Course (SFQC). All potential team members (including TLs and TSs) receive the same training, except for the last phase of SFQC, in which Officers (potential TLs) attend a specialized training module different from the others, where they are given advanced leadership instruction. ODAs are assigned to and operate within five different geographical theaters throughout

the world (Southeast Asia, Africa, the Middle East, South/Central America, and Europe), with several ODA teams operating out of each theater at once. ODA teams function in several different military and humanitarian capacities within their geographical area, requiring each ODA to undergo specific language and cultural training for their particular region of the world.

In general, team leaders are in command of the ODA, and team sergeants act as support for team leaders. However, the nature of the ODA is such that, for any team, the TL and TS share responsibilities that would typically be assigned to a single supervisory position in a traditional organization: TLs deal extensively with external bureaucratic and administrative aspects of ODA operations in addition to leading and planning missions, while TSs generally work only within actual mission planning and operations. Also, the processes by which officers and noncommissioned officers gain entry into the ODA differ. Team leaders are commissioned officers who have completed the Army's Officer Candidate School (OCS), the SFAS and SFQC, and have been assigned to an ODA, where they typically stay for a relatively short period of time before advancing to other positions. Team sergeants are noncommissioned officers who began their military careers as enlisted men and have advanced to their current rank. They have completed the Army's Basic Noncommissioned Officer Course (BNOC), the SFAS and SFQC, and have been assigned to an ODA, where they tend to remain for an extended period of time until moved to other ODAs.

The SF organization as a whole is structured in such a way that although there exists a clear hierarchy of leadership, individual soldiers may receive orders from various different sources. This organization is also unique in a sense, in that each team has a leader as well as an unofficial co-leader, and while the co-leader does not have the same authority in the hierarchy as the leader, he does have responsibilities that make him quite influential. Thus in terms of nonmilitary organizations, we might think of each soldier on a team as having two supervisors, one of which is higher in terms of organizational status, but the other probably having somewhat more intimate and

varied interactions with the ratees. Additionally, differences between the entry processes of TLs and TSs into SF suggest differential leadership training, and the disparate lengths of time in an ODA suggest differences in experience, both as a soldier in general and as a leader in particular.

Ratings used in the present analysis were obtained through large-scale data collection in which all SF soldiers were evaluated by their respective team leaders and team sergeants using a behaviorally based ratings form developed specifically for the USASF (Wilson et al., 2001). Team leaders and team sergeants completed evaluations of each soldier under their command, as well as on each other, but did not complete self-evaluations. For the purposes of this study, only those ratings by the TLs and TSs of their subordinate soldiers were examined.

Study Research Questions

The usefulness of performance ratings from multiple sources has found mixed support since it is unclear whether such ratings can be considered comparable. Logically, comparisons of ratings between rater groups should not be attempted until it can be determined that the rater groups refer to the same conceptualizations of performance when making such ratings, thereby demonstrating that measurement invariance between groups exists. Previous assessments of ratings by different rater groups seem to suggest low to moderate agreement, however in only a few of these has measurement invariance been established. Even in those cases, the emphasis seems to be focused more on peer ratings than the traditional supervisory ratings used in many organizations. Additionally, very little is known about the possible influence, if any, of rater tenure/experience and geographical location on the measurement invariance between rater groups. Thus of primary interest in this study was whether measurement invariance existed in the rating form between the two rater groups (first and second level supervisors) when examined by supervisory level alone, and whether this invariance could be established using the complimentary statistical methods of CFA and IRT. If MI was not found at this first level of analysis, emphasis was shifted to examination of other possible factors that may have

influence, specifically supervisory experience level (as measured by tenure on the team) and geographical field location assignment, respectively. If MI was not found at the supervisory level alone, these two variables were examined in conjunction with supervisory level. However, if MI was found in the rating instrument across the rater groups at the supervisory level alone, analysis was halted and examination of experience and location did not proceed.

Methods

Participants

The individuals of interest were all members of the US Army Special Forces, assigned to a SF team or Occupational Detachment Alpha (ODA). All members were male, and two primary samples were examined in this study: the first sample ('TL') included the ratings of 1052 team members by their respective team leaders (such that most, if not all, raters evaluated multiple individuals within the sample). The second sample ('TS') included ratings of 910 team members by their respective team sergeants (such that most, if not all, raters evaluated multiple individuals within the sample, including many, if not all, of the same individuals evaluated by their TL in the previous sample). Although each TL and TS was required to complete an evaluation of each subordinate soldier, incomplete observations were dropped from analysis, resulting in different sample sizes for the TL and TS rater groups. Collectively, these samples were designated the 'combined' sample, as they consisted of the combined ratings of subordinates by both supervisor levels. No other samples were used, but each of these samples was additionally partitioned into subsamples stratified by tenure (this variable was dichotomized into 'high tenure' and 'low tenure,' due to the observation that nearly 80 percent of the sample had been with their current team for less than 18 months at the time of the ratings) and geographical location (Groups 1, 3, 5, 7, and 10, as assigned by the US Army's numerical scheme for SF teams) when necessary to continue analysis if measurement invariance was not found at the initial supervisor level.

Rating Instrument

The instrument used in this study is the SF Team Member Performance Rating Form (PRF; see Appendix A), a field performance rating form developed specifically for the US Army Special Forces, (Wilson et al., 2001) primarily to improve selection, assessment, and training processes. The rating form is comprised of four sections. Section I collects administrative and individual data, including rater and ratee identification numbers, ODA, rater position (TL or TS), and rater and ratee length of time on the team (tenure). Section II consists of a Mixed Standard Rating Scale (MSRS) of 33 behavioral items, which condense into 9 sub-dimensions, and further into 3 performance dimensions. Appendix B lists these behaviors by sub-dimension, ordered as high behavior (I), average behavior (II), and low behavior (III). Raters are asked to rate the individual as always (+), sometimes (0), or never (-) performing the behavior indicated. The nature of the MSRS format is such that only a set number of response combinations for each sub-dimension are logically consistent, thus reducing halo and leniency errors (for a more detailed description of the MSRS format for this particular instrument, see Thompson, 2002). Section III asks the rater to evaluate the individual on a 7-point scale (low, effective, high) on each of three performance dimensions (Soldiering Skills, SF Specific Skills, Team Member Skills). Section IV consists of a single 11-point scale on which raters are required to force-rank each team member. Based on conversations with a member of the team who helped create the PRF, only the first 27 items from Section II (of the total 33) were included in analyses, as the last 6 were added to the form by end-user request rather than through traditional reliability and validity testing (J.A.Thompson, personal communication, 20 July 2003).

Procedure

Ratings used in the present analysis were obtained through large-scale data collection as part of the initial validation procedures for the PRF (Wilson et al., 2001). Field performance rating forms

were distributed to teams within each of the five active-duty SF Groups. Team leaders and team sergeants completed evaluations of each of the soldiers under their command, resulting in two ratings per soldier. Additionally, each team leader rated his team sergeant, and the team sergeant rated his team leader (although these ratings were not analyzed in the present study due to insufficient sample sizes). Ratings were given only by TLs and TSs, and individual team members did not complete self evaluations, nor did they provide ratings for any other individual. Usable ratings (following removal of incomplete data) comprise the samples used here.

Results

The present study used two complimentary statistical methods to examine measurement invariance across two rater groups: confirmatory factor analysis (CFA) and item-response theory (IRT). Prior to conducting these two procedures an exploratory factor analysis (EFA) was performed in order to determine the factors underlying the PRF, as CFA requires the determination of a factor structure for analysis of appropriate fit of the data, and IRT, as a unidimensional analysis, requires a priori establishment of the manifest variables comprising each factor. The rating form itself was initially developed using a 3-factor a priori model of performance (Wilson et al., 2000), but the decision was made to forgo examination of the a priori model in favor of development of a new model using exploratory factor analysis. This decision was made based on a personal belief that the examination of MI should be based on the performance model that raters appear to be using, rather than the model raters ***should be*** using.

Procedures for determining measurement invariance across team leaders and team sergeants were based primarily on procedures described by Vandenberg and Lance (2000), Facteau and Craig (2001), and Maurer et al. (1998), as these sources discuss the specific recommended methodology for examining measurement invariance of ratings, and Reise at al. (1993). The logic underlying each of these methods will be examined briefly, but the reader is invited to refer to Vandenberg and Lance

(2000) and Maurer et al. (1998) for more detailed discussions of measurement invariance using CFA and IRT, respectively. Both techniques have been assessed as sufficient for testing measurement invariance between rater groups (Facteau & Craig, 2001), but neither has been hailed as more informative than the other, as both provide unique information regarding the comparability of ratings. Due to the complexity of the analyses, the analyses and results are presented together in the next section to facilitate reading. The CFA analyses and results are described first, followed by the IRT analyses and results. Relationships to the proposed research questions are addressed prior to each methodology.

EFA Results. As previously described, a decision was made against testing the a priori 3-factor model in the present analysis in favor of performing an EFA to determine the baseline model to the tested using CFA and IRT procedures. All responses to the first 27-items of the rating form were subjected to an exploratory factor analysis (combined N = 1962), and all EFA procedures were performed using the SAS statistical program, version 8 (SAS Institute, 1999). The maximum likelihood method was used to extract the factors, and a scree test indicated four meaningful factors, which were retained for oblique rotation. In interpreting the rotated factor pattern, an item was associated with a given factor if the factor loading was .35 or greater for that factor, and less than .35 for all others factors (Hatcher, 1994). Using these criteria, 9 items were found to load on the first factor (labeled ‘Leadership’), 8 items loaded on the second (‘Initiative and Effort’), 7 items loaded on the third factor (‘Environment Adaptation’), and 3 items loaded on the fourth factor (‘Interpersonal Skills’). Items and corresponding factor loadings are presented in Table 1. The model generated by the EFA did not appear to parallel the three factors proposed by the a priori model used to create the instrument: the EFA-generated Leadership factor appears to include the majority of positive behaviors from across the a priori factors, while the Initiative and Effort factor includes primarily negative items, also from across the a priori factors. The Environment Adaptation factor is

comprised of those items, both positive and negative, which are concerned with dealings with indigenous populations, and most closely resembles the SF Specific Skills factor from the a priori model. Lastly, the EFA-generated Interpersonal Skills factor includes all 3 items associated with one subfactor of the a priori Team Member Skills factor (specifically, the Handling Interpersonal Situations subfactor). These findings would imply that attempts to fit the a priori model to the samples would have resulted in poor fit, as both the number of factors and items associated with each factor differ between the a priori model used to create the instrument and the model extracted in the exploratory factor analysis.

CFA Analysis. Following the EFA on the total combined sample, simultaneous multiple-group confirmatory factor analyses (CFAs) were then performed, and model fit was examined to determine MI. All CFA analyses were performed using the Mplus statistical program, version 2 (Muthen & Muthen, 2000). Examination of measurement invariance across rater groups was performed first at the rater group level. Measurement invariance testing using confirmatory factor analysis was performed using techniques prescribed by Vandenberg and Lance (2000) condensed into 4 levels (from least to most stringent in terms of showing MI) and described here as Levels 1, 2, 3, and 4 to facilitate reading. The levels of invariance are discussed, followed by analysis results.

The least stringent requirement of demonstrating measurement invariance (discussed here as ‘Level 1’) involves allowing both item factor loadings (measurement model) and covariance/correlation matrices to vary between rater groups, effectively testing whether the number of factors alone is equivalent within each group. This is done by testing each sample independently against the baseline model to establish that the number of factors produced by the baseline model fit in each sample. Acceptance of invariance at this level (but no higher) would indicate that although the rater groups appear to partition performance into the same number of dimensions, the dimensions themselves are not equivalent across rater groups.

The next requirement of invariance ('Level 2') is somewhat more stringent and involves fixing the covariance/correlation matrices to be equivalent between the rater groups but allowing the measurement model to vary between them. In this test, the rater groups are analyzed simultaneously (rather than independently as in the Level 1 analysis), and acceptance of invariance at this level (but no higher) would indicate not only that the rater groups partition performance into the same number of dimensions, but that these dimensions also seem to be related to each other by a similar manner across rater groups.

The Level 3 test of invariance examines the converse of Level 2, and now involves fixing the measurement model to be equivalent between the rater groups but allowing the covariance/correlation matrices to vary (thus allowing different factor relationships within each rater group). Again, analysis is performed across both rater groups simultaneously, and acceptance at this level (but no higher) would indicate that the individual factors (and associated items) themselves are invariant between rater groups, but the way in which those factors are related to one another differs between groups.

The final and most stringent test of invariance (Level 4) tests the equality of both the measurement model and covariance/correlation matrices across both rater groups (i.e., equal factor loadings of each item across rater groups). Acceptance at this final level would indicate that the model does indeed fit in both rater groups and that both rater groups appear to use the same mental model of performance when completing performance evaluations.

This four-step test of invariance was used to test the fit of the model obtained from the EFA to the rater group samples. For each step, a chi-square (χ^2) value and a relative chi-square (χ^2/df) value were computed, in addition to other goodness-of-fit indices. Since the χ^2 statistic tests the null hypothesis that the model fits the data, χ^2 values produced in each CFA were expected to be relatively small (with respect to the appropriate degrees of freedom) and nonsignificant ($p > .05$) if

the model was indeed a good fit to the data. Kline (1998) also states that χ^2/df values should be 3.0 or less to be considered as evidence of acceptable fit. However, χ^2 is extremely sensitive to sample size (often resulting in the rejection of well-fitting models, as trivial differences between the observed model and the perfect-fit model may be found significant; Hu & Bentler, 1999; Hatcher, 1998). Therefore, other fit indices were also examined to evaluate model fit; specifically, the comparative fit index (CFI; which should be close to or above .90 to be accepted as evidence of acceptable fit), the root mean square error of approximation (RMSEA; where a value between 0 and .05 indicates a close fit to the model in the population, .08 or less indicates reasonable fit, and anything over .08 indicates poor fit), and the standardized root mean squared residual (SRMR), which should be less than .08 for acceptable fit (Hu & Bentler, 1999). Although ideal, it has been argued that perfect model fit is an unreasonable expectation and models that display 'close fit' should be considered adequate for measurement invariance purposes (Facteau & Craig, 2001). Therefore, for all analyses, goodness-of-fit indices were evaluated based on compliance with the aforementioned fit statistic acceptability guidelines, and model fit was accepted based on 'close fit.'

CFA Results. At the rater group level of analysis, the baseline model determined by the EFA was independently tested for fit in both the TL sample (N = 1052) and TS sample (N = 910) as prescribed by the first test of measurement invariance. As Table 2 presents, the chi-square values were statistically significant for both rater groups ($p < .001$), and relative chi-square values were all greater than 3. However, as these indicators were sensitive to sample size, they were used in conjunction with other goodness-of-fit indices to determine model fit (Hu & Bentler, 1999). Although not exceeding the conventional .90 level, CFI was relatively close to .90, and RMSEA and SRMR were within the acceptable boundaries (less than .08) for both rater groups. It appears not inappropriate then to conclude that the rater groups satisfied the requirements for the first test of measurement invariance, indicating that the rater groups do appear to share the same number of

performance dimensions. Since the rater groups passed this first test of invariance, analysis proceeded to the remaining tests of invariance, and results are presented in Table 2. The rater groups were tested simultaneously at the second level of invariance (allowing the measurement model to vary while fixing the covariance/correlation matrices to be equal) and results demonstrated acceptable fit according to the criteria specified previously ($\chi^2(642) = 3358.53$, $p < .001$, CFI = .880, RMSEA = .066, SRMR = .053). The rater groups were then tested simultaneously at the third level of invariance (allowing the covariance/correlation matrices to vary while fixing the measurement model to be equal) and results again demonstrated acceptable fit according to recommended guidelines ($\chi^2(663) = 3420.51$, $p < .001$, CFI = .878, RMSEA = .065, SRMR = .061). The final test of measurement invariance (fixing the measurement model and covariance/correlation matrices to be equivalent) also showed acceptable simultaneous fit of the rater groups ($\chi^2(669) = 3424.53$, $p < .001$, CFI = .878, RMSEA = .065, SRMR = .061). Acceptance at this most stringent level of measurement invariance indicated that the baseline model specified by the EFA showed reasonable fit to both rater groups. Measurement invariance was therefore established for both groups, and it was concluded that the two rater groups do indeed appear to be employing similar mental models of performance in their ratings of subordinates. Since MI was found to exist at the general rater group level of analysis, further analyses by tenure and location were not required.

IRT Analysis. Because the CFA analysis generally tests overall measurement invariance of performance dimensions between rater groups (Facteau & Craig, 2001), the complementary method of testing measurement invariance analysis using item response theory was performed to examine the invariance of specific items within each dimension.

Within-factor MI can be tested using IRT by performing a series of analyses using the "differential functioning of items and tests" (DFIT) framework, (Raju, 1999). In IRT, an individual's level or ability on some latent trait (θ) is estimated based on his or her response to some test item or

rating scale. Probability levels of response (the probability of an individual responding in a certain way) are estimated based on 2 item parameters: the discrimination level of the item (*a;* i.e., how well the item distinguishes between high and low ability levels) and difficulty of the item (*b*). For rating scales, the *b*-parameter is typically referred to as a 'threshold' rather than 'difficulty' since it reflects the point on the latent trait (θ) where there is a 50% chance of the individual choosing that particular option or the one directly above it. Thus for any item on a rating scale of 1 to 3, there are 2 *b*-values: one for the threshold between selecting '2' over '1', and another threshold for selecting '3' instead of '2.' In contrast, the *a*-parameter represents the degree to which an item discriminates between different levels of ability (how sensitive the item is to changes in ability). For the present study, this ability level is representative of the rater's mental model of the associated item. An item with a high *a*-value is better able to distinguish between two individuals with adjacent ability levels (mental models) than an item with a low *a*-value (Maurer et al., 1998).

DFIT analysis requires determining both the item parameters associated with the sample, and the person parameters (IRT-based scale scores; θ) for each rater in the sample. Because IRT and DFIT require unidimensionality of items, the baseline model established in the initial EFA was again used to determine the appropriate associations of manifest items to performance dimensions. Following acceptable model fit under the CFA analysis, each performance dimension specified by the model was examined independently across both rater groups simultaneously to determine whether specific items could be isolated as sources of variance between rater groups. To obtain the parameters for each behavioral item, each performance dimension was run as a graded response model (GRM) using the MULTILOG program (Thissen, 1995), and then linked (matched for scale) via the EQUATE program (Baker, 1995; see Table 3). Once the samples were linked, differential functioning was examined with the DFIT program, which calculated differential functioning indices and chi-square values for each item within the performance dimension. Differential functioning

illuminates any differences in the expected responses of raters that could be attributed to their membership in different groups (Facteau & Craig, 2001). Two types of differential functioning of interest here were: differential item functioning (DIF), which could exist between individual (single) items across rater groups, and differential test functioning (DTF), which examines the entire dimension as a whole across rater groups. There are two indices of DIF at the individual item level: noncompensatory differential item functioning (NCDIF), and compensatory differential item functioning (CDIF). DIF is a property of any item that is represented differently for raters of the same theta level (i.e., mental model) across groups. Thus if an item was identified as being a DIF item, it was concluded that the item in question did not represent the same performance dimension for both rater groups. NCDIF is a purely item-level statistic that considers each item separately, regardless of all other items, and assumes that all items except the one being examined lack DIF (i.e., all other items are equivalent between the rater groups). NCDIF, then, reflects true response differences between the two samples (Facteau & Craig, 2001), and therefore was of primary concern for this study. By contrast, CDIF is an additive statistic that, when summed, will produce DTF, which represents the proportion of the dimension as a whole that differs between the rater groups. Significant values are determined in part by the number of possible responses for any given item, and as such significant values were identified as those items that exceeded NCDIF = .024 or DTFs exceeding .216, .192, .168, and .072, for dimensions 1 through 4 respectively, occurring alongside significant chi-square values at $p < .01$ (Raju, van der Linden, & Fleer, 1995). NCDIF and DTF above these critical values would indicate that differential functioning exists within the dimension, and thus the rater groups would not be considered invariant with regard to the item or dimension in question.

IRT Results. Because reasonable fit of the model was found at the rater group level using CFA (significant χ^2 values, but acceptable goodness-of-fit indices; establishing between-factor MI),

further IRT analyses were performed within each of the factors to also establish within-factor MI between the rater groups. These indices are presented in Table 4. As described previously, two criteria are required to show differential functioning (and hence noninvariance between rater groups): significant χ^2 values and either NCDIF (for individual items) or DTF (for whole dimensions) above the set critical values. As the table shows, although 3 of the 4 dimensions demonstrated significant χ^2 values (which, as in the CFA analysis, was to be expected due to the large sample sizes; Hatcher, 1998), there were no instances where NCDIF or DTF exceeded the critical values in the comparisons of the rater groups. This indicated that the items within each dimension showed MI across both rater groups, and this was true for all four dimensions. Taken together, both the CFA and IRT analyses suggest both rater groups possess similar mental models of the performance dimensions delineated in the performance rating form, and that ratings between the two rater groups are comparable.

Thus, to answer the primary research question of whether the PRF was invariant across two rater groups at different supervisory levels, it has been shown using the techniques of CFA and IRT that the Field Performance Rating Form shows at least reasonable MI across the rater groups, at both the between- and within- factor levels of analysis. Because of this finding, additional examinations by experience (tenure) and geographical location were not performed.

Discussion

Determining whether measurement invariance exists in a rating form intended to be used by different rater groups has important implications for the usage of the resultant ratings. Much of the existing research on multiple source feedback systems has attempted to determine the accuracy of agreement between different rater sources without first determining whether the sources' underlying mental models of performance are equivalent enough to make such comparisons valid. Researchers who have undertaken this critical step have been mixed in their evaluation of the invariance of performance ratings, and in these cases the primary focus appears to be examination of sources at

different formal organizational levels (e.g., top executive versus first-line manager) rather than at levels determined by relationship to the ratee as in this study. The purpose of the present research was to determine the congruence of mental models of performance across two rater groups through the examination of measurement invariance of the Special Forces Field Performance Rating Form using CFA and IRT. Results of the CFA analyses found reasonably close fit of the model to the data with respect to goodness-of-fit statistics using the most stringent test of measurement invariance, suggesting that the PRF measures the same underlying performance dimensions in each rater group. IRT analysis also showed the invariance of all items within all dimensions, and although nearly all chi-square analyses were significant, no differential functioning was found for any of the 27 items. These findings led to the conclusion that the PRF was invariant across the two supervisor rater groups, suggesting that team leaders and team sergeants do appear to share the same mental models of the latent performance dimensions represented in the PRF.

This finding of MI of the PRF is an ideal outcome, in that it provides several benefits to the organization in using these ratings. As previously discussed, one critical assumption of multiple source feedback systems is that each rater is privy to a unique perspective of the ratee's performance that must be sufficiently observed in order for the rating to be considered legitimate by ratees (Farr & Newman, 2001; Murphy et al., 2001). This suggests that ratings from different groups collected as part of a MSF system are, in fact, not expected to be equivalent (e.g., Lance & Bennett, 1997; Conway & Huffcutt, 1997). However, this 'expectation' of disagreement between raters is fundamentally rooted in the assumption that the conceptualizations of performance used by each rater are equivalent between raters and that the criteria required by a rater to assign a certain rating is equivalent to the criteria required by any other rater (e.g., such that a rating of "4" represents the same level of performance for each rater, and that a "4" for one rater does not translate to a "2" for another). With regard to the organization in the present study, a rating given to a soldier by his team

sergeant has now been determined to exist on the same scale (i.e., have the same meaning) as a rating given by the soldier's team leader – thus, if a soldier's team leader rates him as possessing "excellent" navigational skills (e.g., a "5"), and his team sergeant rates him as having only "good" navigational skills (e.g., a "4"), we know that the ratings of "excellent" and "good" are on the same scale for both the team leader and team sergeant, and that "excellent" for one rater does not in fact mean "good" for the other. As it were, we can now say that this disagreement between the team leader's and team sergeant's ratings for the same individual are likely based on some other variable, such as opportunity to observe or inaccurate recall, and not differential mental models of performance between the two raters as measured by the rating form. By the same token, if the team leader and team sergeant both agree that the soldier has "excellent" navigational skills, we can say with some certainty that both the team leader and team sergeant do conceive of "excellence" on that task using the same model, and that they are not "speaking different languages." It has been cautioned, however, that finding measurement invariance does not necessarily assume that ratings from different rater groups will completely agree, or that they are indeed accurate (Facteau & Craig, 2001). Still, this allows the organization to assess an individual's performance from different perspectives while establishing that ratings between raters are conceptualized the same within raters.

There are implications for the present study's findings with regard to non-military organizations as well. Establishment of measurement invariance for two levels of supervisors could allow for the comparison of a single employee's performance ratings by different supervisors, which might be particularly salient in cases where there is high career mobility, and employees in new positions are evaluated by new, different levels of supervisors. Even if ratings between the former and new supervisors do not agree quantitatively, if measurement invariance can be determined to exist between the rater groups, it could at the very least be argued that the qualitative performance

scale on which they each rated the employee was consistent, allowing for examination of that employee's performance over time regardless of the rater.

An additional finding with regard to team performance: it has also been suggested that team or group performance is more effective when mental models are similar across team-members, with the implication being that team members who share mental models of performance are better able to anticipate and interpret needs of their teammates, thereby allowing them to work more fluidly and efficiently than teams whose members must constantly explain needs and expectations to one another (Smith-Jentsch, Campbell, Milanovich, & Reynolds, 2001). This is particularly important for the organization of interest in this study, as there is a possibility that chains-of-command may be interrupted due to the casualties of war; in which case it would be important to completion of missions that replacement members be able to immediately understand the needs of the others. Finding invariance, then, has implications for the effectiveness of highly trained teams in the field.

Limitations. The present study had several limitations that may have contributed to the 'moderate' invariance of the PRF across the two rater groups. To start, the PRF was developed as a mixed standard rating scale (MSRS), in which a critical incident behavior was essentially partitioned into three items: one item describing a low/less desirable variation of the behavior, one item describing medium/normal behavior, and one item representing the high/most desirable version of the behavior. This low-medium-high design was implemented in order to alleviate potential bias and/or leniency in ratings through the implementation of a logical scale. As such, a grouping of items – each designated as a low, medium, or high variation of a single behavior – can only be considered 'logically consistent' if, on the 3-point scale from '*Never*' to '*Always*' each is assigned a rating, but no two items in the set receive the same rating (Thompson, 2002). It might be possible that model fit (and thus invariance) could have been improved had analysis been restricted to only logically consistent raters; however, this tactic may be flawed in that selecting the sample based on 'correct'

responses to previous questions would violate the IRT assumption of local independence (S.B. Craig, personal communication, 1 December 2004). Future research in measurement invariance should consider the nature of the scale in question before deciding on an appropriate way to examine measurement invariance for the instrument. Cursory examination of all raters in the present study suggested that only a very small proportion of raters could actually be considered logically consistent on all 27 items, thereby reducing the sample size available for analysis. As such, logically inconsistent raters were retained in this analysis. Additionally, it has been suggested that perhaps the medium level items should be dropped entirely, since the behavioral items are based on critical incidents, with the middle item suggesting 'normal behavior'—a behavior rater often find difficult to accurately assess (M.A. Wilson, personal communication, 23 January 2004). Further examination of this instrument may be warranted to determine whether such an action would improve invariance.

Another limitation involves the decision to perform an exploratory factor analysis on the combined sample rather than using either the 3-factor a priori model used to create the instrument, or finding the best-fitting model for each rater group independently (e.g., Facteau & Craig, 2001). Specifically, there were concerns that performing an EFA on the combined sample compromises the integrity of the subsequent CFA analysis on the same sample. Although admittedly unconventional, the rationale behind the decision to obtain the baseline model from the combined sample was intended to determine the best general model that represented both rater groups as they were currently (and practically) using the instrument. There were also concerns that no clear indication was made of whether individual raters were using the same mental model of performance to rate each individual (since a single rater evaluated multiple ratees within the sample). While an intriguing question, current thought suggests that testing for measurement invariance in this capacity would be inappropriate, since it would require testing the invariance of each rater in a sample against every

other rater in the same sample, which cannot be done (S.B. Craig, personal communication, December 2004).

Other limitations include the possibility of rater bias, and the inability to examine ratings at a different supervisor level of analysis. One assumption that was made but not tested was that all raters were equivalent within their respective rater groups; however, since it was not the intent of this study to determine the equivalence of the rater groups, but measurement invariance across these rater groups, the within-groups variance was not examined. Additionally, it has been suggested that this particular analysis could not have been performed within the context of the current study as it would require testing for measurement invariance of each individual rater against all the others (essentially testing an N = 1; S.B. Craig, personal communication, 1 December 2004). Another limitation stems from missing data: although ideally each ratee would have been represented exactly twice, resulting in definite and equal sets of known raters, unclear and missing responses forced some evaluations to be discarded. It is therefore unclear whether the ratees in one sample (those soldiers rated by team sergeants) were adequately duplicated in the second (those soldiers rated by team leaders). Also, it is unknown exactly how ratings were collected—whether raters evaluated ratees all in one sitting, distributed across a period of time, or a mixture of both. Such information could have implications for the assessment of the actual quality of the ratings, as rater fatigue or bias may have contributed to variance. The hierarchy of the organization provided an interesting component in that although both rater groups were treated as supervisors, one group (TL) was also the supervisor of the other group (TS). Performance evaluations were also exchanged between these two groups, such that each TL rated his TS, and each TS rated his TL. An analysis at this level would have been interesting to examine whether factors such as seniority or expectations of performance could have an impact on ratings. However, in accordance with CFA suggested guidelines (Baggaley, 1983; Hatcher, 1994), a minimally adequate sample size for analysis should be the larger of 100 subjects or 5 times the

number of variables being analyzed. Twenty-seven items were analyzed and, after accounting for missing data, there were not enough raters in each rater group to meet minimum sample size criterion (N = 135); thus tests of measurement invariance at this second supervisory-level were not performed.

Conclusions. The present research found a performance rating form to be invariant across two groups of raters in complementary (but not identical) supervisory positions, using the statistical methods of confirmatory factor analysis and item response theory. This finding indicates that raters from these groups do appear to share similar mental models of the performance dimensions captured by the rating form, and that ratings between the groups are comparable in the sense that they are made using the same model of performance. Future research in this area should consider examining the ratings across more diverse levels of analysis (not performed here due to inadequate sample sizes), removing ambiguous items (the middle/neutral item in this study), and potentially using modification indices to revise the model obtained from the initial EFA in order to obtain a better fit to the data (not performed here due to the exploratory nature of the study). The performance rating form can be interpreted as a measure of the same performance dimensions in each rater group, and items are equally effective indicators of the performance dimensions represented by the PRF across the two rater groups. The meanings of the ratings can be considered reasonably congruent, regardless of who is rating, and ratings may be considered comparable to the extent that disagreement between raters is not likely due to ‘misinterpretation’ of the rating form by different raters. These findings are supportive and congruent with those found by Facteau and Craig (2001) and Maurer et al. (1998), who also found measurement invariance between different rater groups using both CFA and IRT methods.

Table 1

Item Factor Loadings

Item	Item Name	Factor 1	Factor 2	Factor 3	Factor 4
4	PLAN-1	.56*	.24	.06	-.06
5	TS-2	.36*	.09	-.08	-.05
7	TCH-1	.65*	-.12	.24	-.02
13	PLAN-2	.55*	.17	-.04	-.02
15	TEAM-1	.48*	.15	.11	.12
16	TCH-2	.66*	-.17	.28	-.03
20	NAV-1	.45*	.26	-.08	-.01
23	TS-1	.57*	.11	.09	.02
24	TEAM-2	.48*	.14	.09	.16
1	EFF-1	.29	.58*	-.03	.02
2	NAV-2	.28	.45*	-.10	.03
6	TEAM-3	-.11	.39*	.09	.13
10	EFF-2	.30	.60*	-.05	-.03
11	NAV-3	.05	.46*	.01	.02
14	TS-3	.07	.57*	.15	-.02
19	EFF-3	.01	.66*	.04	.06
22	PLAN-3	.10	.45*	.13	.01
8	LANG-2	.20	-.01	.51*	-.04
9	INDP-3	-.19	.20	.57*	.09
17	LANG-3	-.16	.24	.63*	-.04
18	INDP-1	.26	-.01	.51*	.08
25	TCH-3	.14	.31	.35*	-.10
26	LANG-1	.13	-.09	.54*	-.02
27	INDP-2	.19	.04	.46*	.09
3	INTP-3	-.07	.13	-.04	.67*
12	INTP-1	.18	.03	.04	.61*
21	INTP-2	-.04	-.03	.03	.81*

Note. Values greater than 0.35 are indicated with an '*'.

Table 2

Fit Indices for Iterative CFA Tests of Measurement Invariance

Model	χ^2	df	χ^2/df	CFI	RMSEA	SRMR
Level 1a: Test of equivalent number of factors – team leaders	1856.94*	318	5.839	0.875	0.068 (0.065 – 0.071)	0.053
Level 1b: Test of equivalent number of factors – team sergeants	1494.99*	318	4.701	0.886	0.064 (0.061 – 0.067)	0.051
Level 2: Free measurement model, fixed covariance/ correlation matrices	3358.53*	642	5.231	0.880	0.066 (0.063 – 0.068)	0.053
Level 3: Free covariance/correlation matrices, fixed measurement model	3420.51*	663	5.159	0.878	0.065 (0.063 – 0.067)	0.061
Level 4: Fixed measurement model, fixed covariance/ correlation matrices	3424.53*	669	5.119	0.878	0.065 (0.063 – 0.067)	0.061

Note. χ^2/df = relative chi-square index; CFI = comparative fit index; RMSEA = root mean square error of approximation (and 90% confidence interval); SRMR = standardized root mean square residual.

*$p < .001$.

Table 3

Individual Item Parameters by Dimension

Dimension and Item	TL Group Item Parameters			TS Group Item Parameters[a]		
	a (SE)	b_1 (SE)	b_2 (SE)	*a* (SE)	b_1 (SE)	b_2 (SE)
Dimension 1: Leadership						
Item 4	2.969 (0.25)	-2.576 (0.18)	-.477 (0.05)	2.640 (0.22)	-2.727 (0.18)	-.500 (0.05)
Item 5	.689 (0.11)	-3.984 (0.74)	-.322 (0.15)	.896 (0.12)	-3.803 (0.51)	-.178 (0.11)
Item 7	2.786 (0.21)	-2.121 (0.12)	-.184 (0.04)	2.677 (0.20)	-2.163 (0.12)	-.172 (0.05)
Item 13	2.581 (0.23)	-3.342 (0.37)	-.770 (0.06)	1.620 (0.16)	-3.515 (0.38)	-.717 (0.08)
Item 15	2.472 (0.19)	-2.164 (0.13)	-.314 (0.05)	2.847 (0.23)	-2.047 (0.11)	-.347 (0.05)
Item 16	2.912 (0.23)	-2.488 (0.15)	-.316 (0.04)	2.563 (0.21)	-2.455 (0.14)	-.402 (0.05)
Item 20	1.620 (0.17)	-3.520 (0.42)	-1.211 (0.10)	1.875 (0.18)	-3.283 (0.31)	-1.079 (0.08)
Item 23	2.881 (0.22)	-2.437 (0.16)	-.482 (0.05)	2.716 (0.23)	-2.483 (0.14)	-.491 (0.05)
Item 24	2.732 (0.22)	-2.113 (0.12)	-.429 (0.05)	2.761 (0.23)	-2.238 (0.12)	-.471 (0.05)
Dimension 2: Initiative and Effort						
Item 1	3.969 (0.35)	-2.239 (0.14)	-.814 (0.04)	4.826 (0.52)	-2.228 (0.11)	-.661 (0.04)
Item 2	2.591 (0.25)	-2.732 (0.24)	-.983 (0.06)	2.880 (0.28)	-2.604 (0.19)	-.946 (0.05)
Item 6	1.425 (0.15)	-2.094 (0.22)	-.848 (0.10)	1.286 (0.15)	-2.172 (0.23)	-.924 (0.11)
Item 10	3.600 (0.35)	-2.657 (0.20)	-.903 (0.05)	3.787 (0.36)	-2.278 (0.14)	-.813 (0.05)
Item 11	2.338 (0.29)	-2.663 (0.26)	-1.483 (0.10)	1.846 (0.25)	-2.855 (0.32)	-1.697 (0.14)
Item 14	3.078 (0.28)	-2.345 (0.18)	-1.021 (0.06)	3.007 (0.31)	-2.225 (0.14)	-1.204 (0.07)
Item 19	3.333 (0.29)	-2.046 (0.12)	-.857 (0.05)	2.904 (0.27)	-2.095 (0.13)	-.935 (0.06)
Item 22	2.262 (0.22)	-2.737 (0.24)	-1.137 (0.08)	1.988 (0.24)	-2.784 (0.30)	-1.269 (0.10)
Dimension 3: Environment Adaptation						
Item 8	2.143 (0.17)	-2.302 (0.16)	-.623 (0.06)	1.918 (0.16)	-2.588 (0.26)	-.588 (0.07)
Item 9	2.066 (0.20)	-2.596 (0.22)	-1.252 (0.09)	2.600 (0.22)	-2.396 (0.22)	-1.126 (0.07)
Item 17	2.449 (0.21)	-2.112 (0.14)	-1.032 (0.06)	1.954 (0.17)	-2.477 (0.24)	-1.085 (0.09)
Item 18	3.081 (0.25)	-2.357 (0.13)	-.717 (0.05)	3.734 (0.28)	-2.172 (0.14)	-.615 (0.05)
Item 25	1.948 (0.18)	-2.986 (0.27)	-1.080 (0.08)	1.790 (0.17)	-3.086 (0.37)	-1.264 (0.11)
Item 26	1.742 (0.13)	-1.788 (0.12)	.293 (0.06)	1.859 (0.14)	-1.970 (0.16)	.169 (0.07)
Item 27	2.725 (0.24)	-2.707 (0.20)	-.904 (0.06)	3.094 (0.23)	-2.342 (0.18)	-.756 (0.05)
Dimension 4: Interpersonal Skills						
Item 3	2.292 (0.18)	-2.142 (0.14)	-.807 (0.06)	2.331 (0.19)	-2.109 (0.12)	-.835 (0.06)
Item 12	2.693 (0.20)	-2.269 (0.13)	-.434 (0.05)	2.663 (0.20)	-2.120 (0.11)	-.473 (0.05)
Item 21	6.004 (0.60)	-2.169 (0.09)	-.672 (0.03)	3.103 (0.24)	-2.229 (0.11)	-.727 (0.05)

Note. TS = team sergeant rater group; TL = team leader rater group; *a* = item discrimination parameter; b_1 = item threshold parameter for rating '2' over '1'; b_2 = item threshold parameter for rating '3' over '2'; (SE) = standard errors.

[a]Equated item parameters. Transformation coefficients: Dimension 1: slope (A) = 1.0123, intercept (K) = -0.1650; Dimension 2: A = 1.0137, K = -0.1324; Dimension 3: A = 0.8732, K = -0.1672; Dimension 4: A = 1.0171, K = -0.1732.

Table 4

Differential Fit Indices for IRT/DFIT Tests of Measurement Invariance; by Dimension and Item

Dimension and Item	DTF	χ^2	df	NCDIF	Significance indicated by:
Factor 1: Leadership	.00135	923.72	909		
Item 4				.000	DTF > .216 Significant χ^2 (p < .05) Item NCDIF > .024
Item 5				.002	
Item 7				.000	
Item 13				.004	
Item 15				.001	
Item 16				.000	
Item 20				.000	
Item 23				.000	
Item 24				.000	
Factor 2: Initiative and Effort	.00035	3807.47*	909		
Item 1				.000	DTF > .192 Significant χ^2 (p < .05) Item NCDIF > .024
Item 2				.000	
Item 6				.000	
Item 10				.000	
Item 11				.000	
Item 14				.000	
Item 19				.000	
Item 22				.000	
Factor 3: Environment Adaptation	.00100	4574.85*	909		
Item 8				.000	DTF > .168 Significant χ^2 (p < .05) Item NCDIF > .024
Item 9				.000	
Item 17				.000	
Item 18				.000	
Item 25				.000	
Item 26				.002	
Item 27				.000	
Factor 4: Interpersonal Skills	.00101	1075.64*	909		DTF > .072 Significant χ^2 (p < .05) Item NCDIF > .024
Item 3				.000	
Item 12				.000	
Item 21				.002	

Note. Differential item functioning (DIF) is indicated by NCDIF item values greater than .024.

Differential test functioning (DTF) critical values are calculated by multiplying the critical NCDIF

value by the number of items in the dimension of interest ([.024] x [the number of items in the dimension]).

*$p < .05$.

References

Baker, F. (1995). *EQUATE 2.1: Computer program for equating two metrics in item response theory.* Madison: University of Wisconsin, Laboratory of Experimental Design.

Baggaley, A.R. (1983). Deciding on the ratio of number of subjects to number of variables in factor analysis. Multivariate Experimental Clinical Research, 6(2), 81-85.

Barr, M.A. & Raju, N.S. (2003). IRT-based assessments of rater effects in multiple-source feedback instruments. *Organizational Research Methods, 6(1),* 15-43.

Bentler, P.M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107(2),* 238-246.

Borman, W. C., & Motowidlo, S. J. (1993). Expanding the criterion domain to include elements of contextual performance. In N. Schmitt & W. C. Borman (Eds.), Personnel selection in organizations. San Francisco, CA: Jossey Bass.

Borman, W.C., White, L.A., Pulakos, E.D., & Oppler, S.H. (1991). Models of supervisory job performance ratings. *Journal of Applied Psychology, 76(6),* 863-872.

Bracken, D.W., Timmreck, C.W., & Church, A.H. (2001). *The Handbook of Multisource Feedback*. Jossey-Bass Inc: San Francisco, CA.

Brutus, S., Fleenor, J.W., & McCauley, C.D. (1999). Demographics and personality predictors of congruence in multi-source ratings. *Journal of Management Development, 18(5),* 417-435.

Byrne, B.M., Shavelson, R.J., & Muthen, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105(3),* 456-466.

Campbell, J. P. (1990). Modeling the performance prediction problem in Industrial and Organizational Psychology. In M. D. D. L. M. Hough (Ed.), Handbook of Industrial and

Organizational Psychology (2 ed., Vol. 1, pp. 687-732). Palo Alto, CA: Consulting Psychologist Press.

Carson, K.P., Cardy, R.L., & Dobbins, G.H. (1991). Performance appraisal as effective management or deadly management disease: Two initial empirical investigations. *Group and Organizational Studies, 16*, 143-159.

Conway, J.M. & Huffcutt, A.I. (1997). Psychometric properties of multisource performance ratings: A meta-analysis of subordinate, supervisor, peer, and self-ratings. *Human Performance, 10(4),* 331-360.

Craig, S.B. & Kaiser, R.B. (2003a). Applying item response theory to multisource performance ratings: What are the consequences of violating the independent observations assumption? *Organizational Research Methods, 6(1),* 44-60.

Craig, S.B. & Kaiser, R.B. (2003b). Using item response theory to assess measurement equivalence of 360° performance ratings across organizational levels. In A. Meade (Chair) Applications of Item Response Theory to Measurement in Organizations. Symposium presented at the annual conference of the Society for Industrial-Organizational Psychology in Orlando, Florida.

Cramton, C.D. (2001). The mutual knowledge problem and its consequences for dispersed collaboration. *Organization Science, 12(3),* 346-371.

Facteau, J.D. & Craig, S.B. (2001). Are performance appraisal ratings from different rating sources comparable? *Journal of Applied Psychology, 86(2),* 215-227.

Farr, J.L. & Newman, D.A. (2001). Rater Selection: Sources of Feedback. In Bracken, D.W., Timmreck, C.W., & Church, A.H. (Eds.) *The Handbook of Multisource Feedback.* Jossey-Bass Inc: San Francisco, CA. (pp. 96-113).

Ferris, T.A. & Judge, G.R. (1993). Social context of performance evaluation decisions. *Academy of Management Journal, 36,* 80-105.

Grant, L.D. (1996). A comprehensive examination of the latent structure of job performance. Doctoral Dissertation, North Carolina State University, Raleigh.

Greguras, G.J. & Robie, C. (1998). A new look at within-source interrater reliability of 360-degree feedback ratings. *Journal of Applied Psychology, 83(6),* 960-968.

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

Harris, M.M. & Schaubroeck, J. (1988). A meta-analysis of self-supervisor, self-peer, and peer-supervisor ratings. *Personnel Psychology, 41,* 43-62.

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

Judge, T.A. & Ferris, G.R. (1993). Socail context of performance evaluation decisions. *Academy of Management Journal, 36(1),* 80-105.

Kline, R.B. (1998). *Principles and practice of structural equation modeling*. New York: Guilford Press.

Lance, C.E. & Bennett, W.J. (2000). Replication and extension of models of supervisory job performance ratings. *Human Performance, 13(2),* 139-158.

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

Maurer, T.J., Raju, N.S., & Collins, W.C. (1998). Peer and subordinate performance appraisal measurement equivalence. *Journal of Applied Psychology, 83(5),* 693-702.

Mount, M.K., Judge, T.A., Scullen, S.E., Systsma, M.R., & Hezlett, S.A. (1998). Trait, rater and level effects in 360-degree performance ratings. *Personnel Psychology, 51,* 557-576.

Murphy, K.A. & Cleveland, J.N. (1995). *Understanding performance appraisal: Social, organizational, and goal based perspectives.* Thousand Oaks, CA: Sage.

Murphy, K.A., Cleveland, J.N., & Mohler, C.J. (2001). Reliability, validity, and meaningfulness of multisource ratings. In Bracken, D.W., Timmreck, C.W., & Church, A.H. (Eds.) *The Handbook of Multisource Feedback.* Jossey-Bass Inc: San Francisco, CA (pp. 130-148).

Muthen, L.K., & Muthen, B.O. (2000). *Mplus: Statistical analyses with latent variables. User's guide (Version 2)*. Los Angeles: Author.

Raju, N.S. (1999). *DFIT4P: A computer program for analyzing differential item and test functioning.* Chicago: Illinois Institute of Technology.

Raju, N.S, van der Linden, W.J. & Fleer, P.F. (1995). IRT-based internal measures of differential functioning of items and tests. Applied Psychological Measurement, 19(4), 353-368.

Reise, S.P., Widaman, K.F., & Pugh, R.H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin, 114(3),* 552-556.

Rothstein, H.R. (1990). Interrater reliability of job performance ratings: Growth to asymptote level with increasing opportunity to observe. *Journal of Applied Psychology, 75(3),* 322-327.

SAS Institute (1999). SAS onlinedoc, Version 8. Cary, NC: SAS Institute.

Smith-Jentsch, K.A., Campbell, G.E., Milanovich, D.M., & Reynolds, A. M. (2001). Measuring teamwork mental models to support training needs assessment, development, and evaluation: Two empirical studies. *Journal of Organizational Behavior, 22*, 179-194.

Scullen, S.E., Mount, M.K., & Goff, M. (2000). Understanding the latent structure of job performance ratings. *Journal of Applied Psychology, 85(6),* 956-970.

Thissen, D. (1995). *MULTILOG 6.3: A computer program for multiple, categorical item analysis and test scoring using item response theory.* Chicago: Scientific Software, Inc.

Tesluk, P.E. & Jacobs, R.R. (1998). Toward an integrated model of work experience. *Personnel Psychology, 51,* 321-355.

Thompson, J.A. (2002). The relationship between scaled behavioral ratings, performance dimension ratings and rankings using United States Army Special Forces Soldiers. Unpublished master's thesis, North Carolina State University, Raleigh.

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(1),* 4-70.

Viswesvaran, C., Schmidt, F.L., & Ones, D.S. (2002). The moderating influence of job performance dimensions on convergence of supervisory and peer ratings of job performance: Unconfounding construct-level convergence and rating difficulty. *Journal of Applied Psychology, 87(2),* 345-354.

Viswesvaran, C., Ones, D.S., & Schmidt, F.L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81(5),* 557-574.

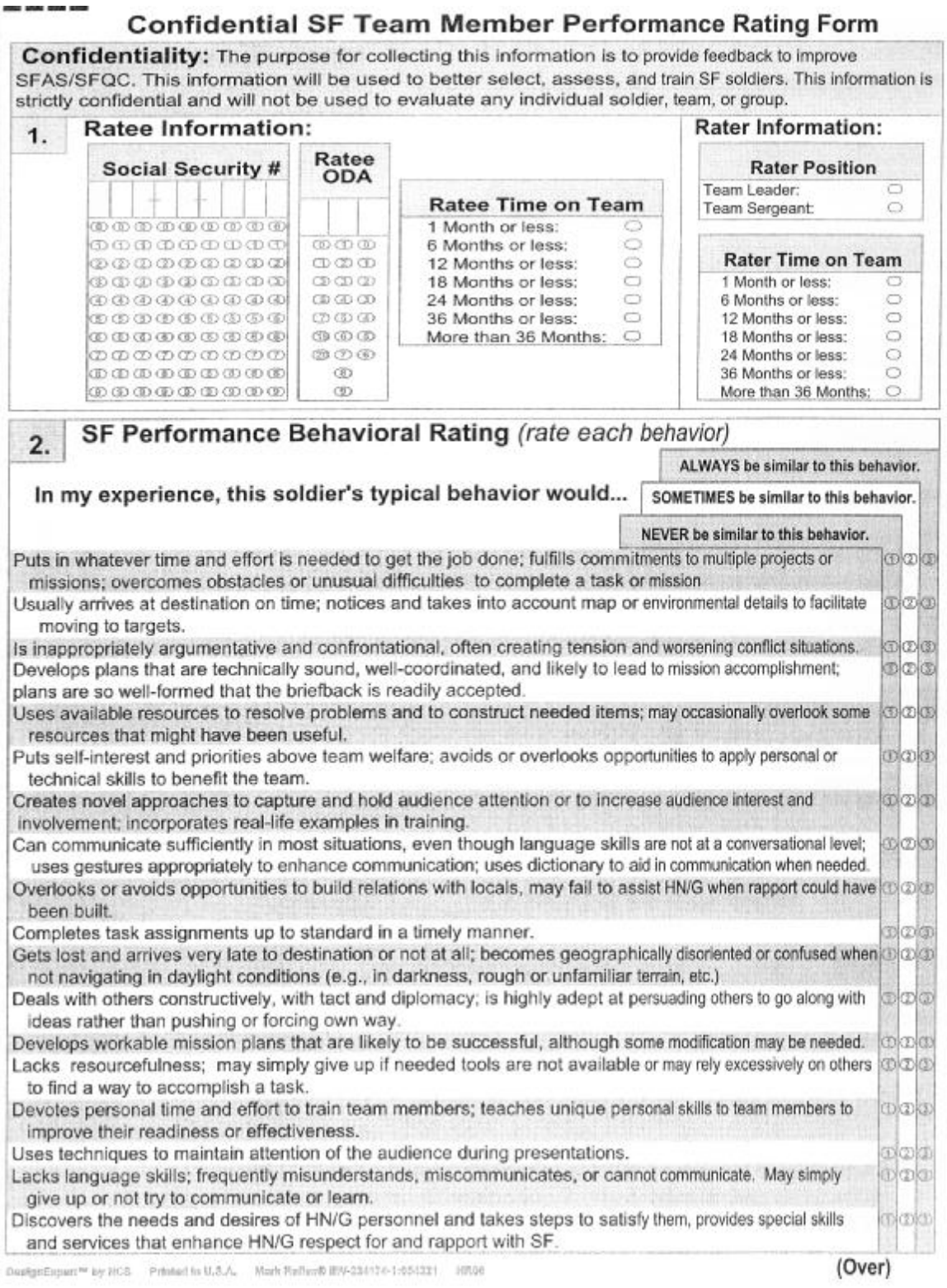
Wherry, R.J. Sr., & Bartlett, C.J. (1982). The control theory of bias in ratings: A theory of rating. *Personnel Psychology, 35,* 521-551.

Wilson, M.A., Drewes, D.W., Cunningham, J.W., Sanders, M.G., Thompson, J.A., & Surface, E.A. (2001). *Modeling Special Forces soldier performance.* Fort Bragg, NC: Army Research Institute Fort Bragg Field Office.

Woehr, D.J. & Huffcutt, A.I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology, 67*, 189-205.

Appendices

Appendix A



Confidential SF Team Member Performance Rating Form

Confidentiality: The purpose for collecting this information is to provide feedback to improve SFAS/SFQC. This information will be used to better select, assess, and train SF soldiers. This information is strictly confidential and will not be used to evaluate any individual soldier, team, or group.

1. Ratee Information:

Social Security #

Ratee ODA

Ratee Time on Team	
1 Month or less:	○
6 Months or less:	○
12 Months or less:	○
18 Months or less:	○
24 Months or less:	○
36 Months or less:	○
More than 36 Months:	○

Rater Information:

Rater Position	
Team Leader:	○
Team Sergeant:	○

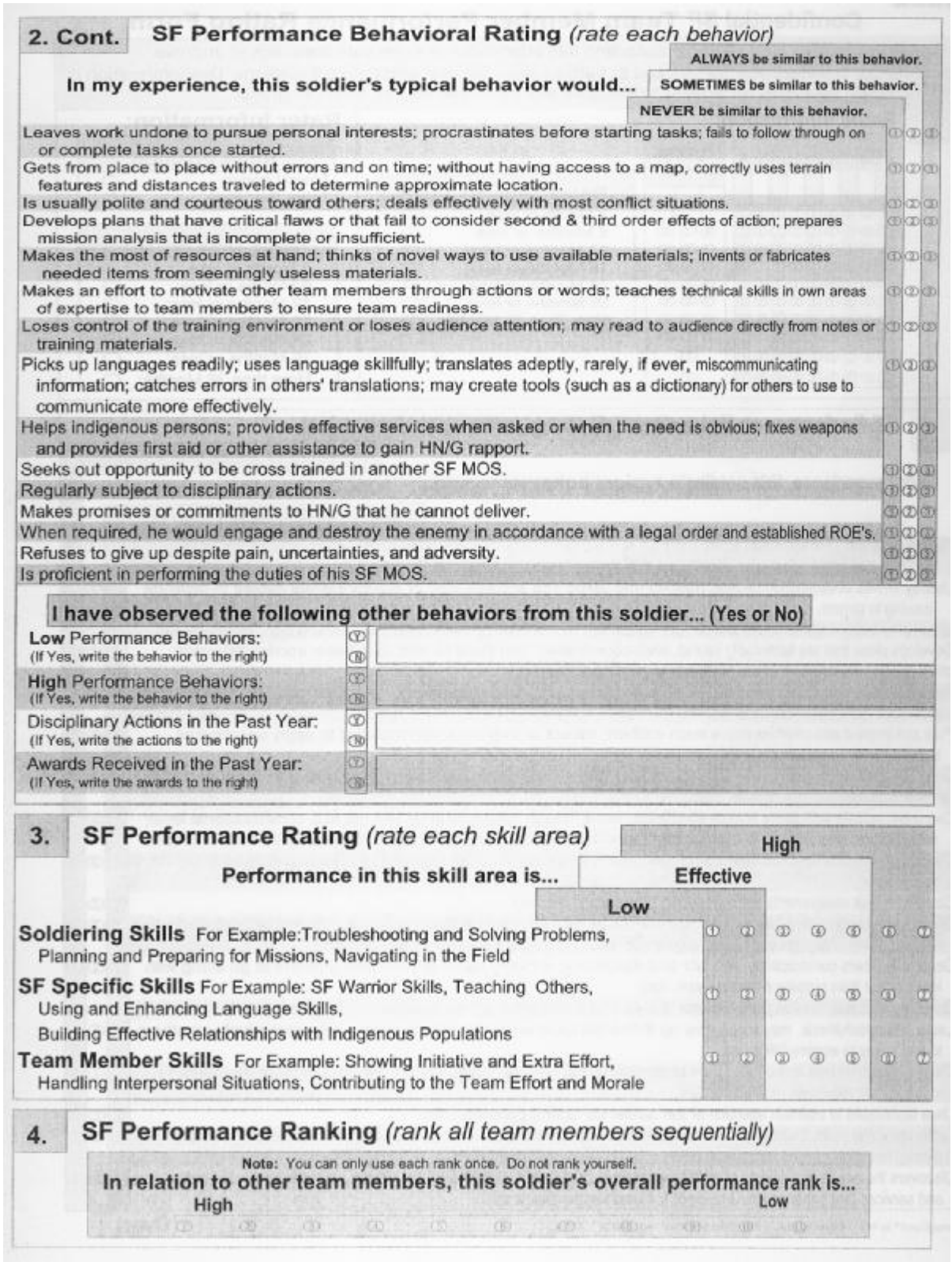
Rater Time on Team	
1 Month or less:	○
6 Months or less:	○
12 Months or less:	○
18 Months or less:	○
24 Months or less:	○
36 Months or less:	○
More than 36 Months:	○

2. SF Performance Behavioral Rating *(rate each behavior)*

In my experience, this soldier's typical behavior would...	ALWAYS be similar to this behavior.	SOMETIMES be similar to this behavior.	NEVER be similar to this behavior.
Puts in whatever time and effort is needed to get the job done; fulfills commitments to multiple projects or missions; overcomes obstacles or unusual difficulties to complete a task or mission	①	②	③
Usually arrives at destination on time; notices and takes into account map or environmental details to facilitate moving to targets.	①	②	③
Is inappropriately argumentative and confrontational, often creating tension and worsening conflict situations.	①	②	③
Develops plans that are technically sound, well-coordinated, and likely to lead to mission accomplishment; plans are so well-formed that the briefback is readily accepted.	①	②	③
Uses available resources to resolve problems and to construct needed items; may occasionally overlook some resources that might have been useful.	①	②	③
Puts self-interest and priorities above team welfare; avoids or overlooks opportunities to apply personal or technical skills to benefit the team.	①	②	③
Creates novel approaches to capture and hold audience attention or to increase audience interest and involvement; incorporates real-life examples in training.	①	②	③
Can communicate sufficiently in most situations, even though language skills are not at a conversational level; uses gestures appropriately to enhance communication; uses dictionary to aid in communication when needed.	①	②	③
Overlooks or avoids opportunities to build relations with locals, may fail to assist HN/G when rapport could have been built.	①	②	③
Completes task assignments up to standard in a timely manner.	①	②	③
Gets lost and arrives very late to destination or not at all; becomes geographically disoriented or confused when not navigating in daylight conditions (e.g., in darkness, rough or unfamiliar terrain, etc.)	①	②	③
Deals with others constructively, with tact and diplomacy; is highly adept at persuading others to go along with ideas rather than pushing or forcing own way.	①	②	③
Develops workable mission plans that are likely to be successful, although some modification may be needed.	①	②	③
Lacks resourcefulness; may simply give up if needed tools are not available or may rely excessively on others to find a way to accomplish a task.	①	②	③
Devotes personal time and effort to train team members; teaches unique personal skills to team members to improve their readiness or effectiveness.	①	②	③
Uses techniques to maintain attention of the audience during presentations.	①	②	③
Lacks language skills; frequently misunderstands, miscommunicates, or cannot communicate. May simply give up or not try to communicate or learn.	①	②	③
Discovers the needs and desires of HN/G personnel and takes steps to satisfy them, provides special skills and services that enhance HN/G respect for and rapport with SF.	①	②	③

DesignExpert™ by HCS. Printed in U.S.A. Mark Reflex® (BW-234174-1-051331) HR06

(Over)



2. Cont. SF Performance Behavioral Rating *(rate each behavior)*

In my experience, this soldier's typical behavior would...	ALWAYS be similar to this behavior.	SOMETIMES be similar to this behavior.	NEVER be similar to this behavior.
Leaves work undone to pursue personal interests; procrastinates before starting tasks; fails to follow through on or complete tasks once started.	1	2	3
Gets from place to place without errors and on time; without having access to a map, correctly uses terrain features and distances traveled to determine approximate location.	1	2	3
Is usually polite and courteous toward others; deals effectively with most conflict situations.	1	2	3
Develops plans that have critical flaws or that fail to consider second & third order effects of action; prepares mission analysis that is incomplete or insufficient.	1	2	3
Makes the most of resources at hand; thinks of novel ways to use available materials; invents or fabricates needed items from seemingly useless materials.	1	2	3
Makes an effort to motivate other team members through actions or words; teaches technical skills in own areas of expertise to team members to ensure team readiness.	1	2	3
Loses control of the training environment or loses audience attention; may read to audience directly from notes or training materials.	1	2	3
Picks up languages readily; uses language skillfully; translates adeptly, rarely, if ever, miscommunicating information; catches errors in others' translations; may create tools (such as a dictionary) for others to use to communicate more effectively.	1	2	3
Helps indigenous persons; provides effective services when asked or when the need is obvious; fixes weapons and provides first aid or other assistance to gain HN/G rapport.	1	2	3
Seeks out opportunity to be cross trained in another SF MOS.	1	2	3
Regularly subject to disciplinary actions.	1	2	3
Makes promises or commitments to HN/G that he cannot deliver.	1	2	3
When required, he would engage and destroy the enemy in accordance with a legal order and established ROE's.	1	2	3
Refuses to give up despite pain, uncertainties, and adversity.	1	2	3
Is proficient in performing the duties of his SF MOS.	1	2	3

I have observed the following other behaviors from this soldier... (Yes or No)

Low Performance Behaviors: (If Yes, write the behavior to the right)	1 / 2	
High Performance Behaviors: (If Yes, write the behavior to the right)	1 / 2	
Disciplinary Actions in the Past Year: (If Yes, write the actions to the right)	1 / 2	
Awards Received in the Past Year: (If Yes, write the awards to the right)	1 / 2	

3. SF Performance Rating *(rate each skill area)*

Performance in this skill area is...	Low		Effective			High	
Soldiering Skills For Example:Troubleshooting and Solving Problems, Planning and Preparing for Missions, Navigating in the Field	1	2	3	4	5	6	7
SF Specific Skills For Example: SF Warrior Skills, Teaching Others, Using and Enhancing Language Skills, Building Effective Relationships with Indigenous Populations	1	2	3	4	5	6	7
Team Member Skills For Example: Showing Initiative and Extra Effort, Handling Interpersonal Situations, Contributing to the Team Effort and Morale	1	2	3	4	5	6	7

4. SF Performance Ranking *(rank all team members sequentially)*

Note: You can only use each rank once. Do not rank yourself.

In relation to other team members, this soldier's overall performance rank is...

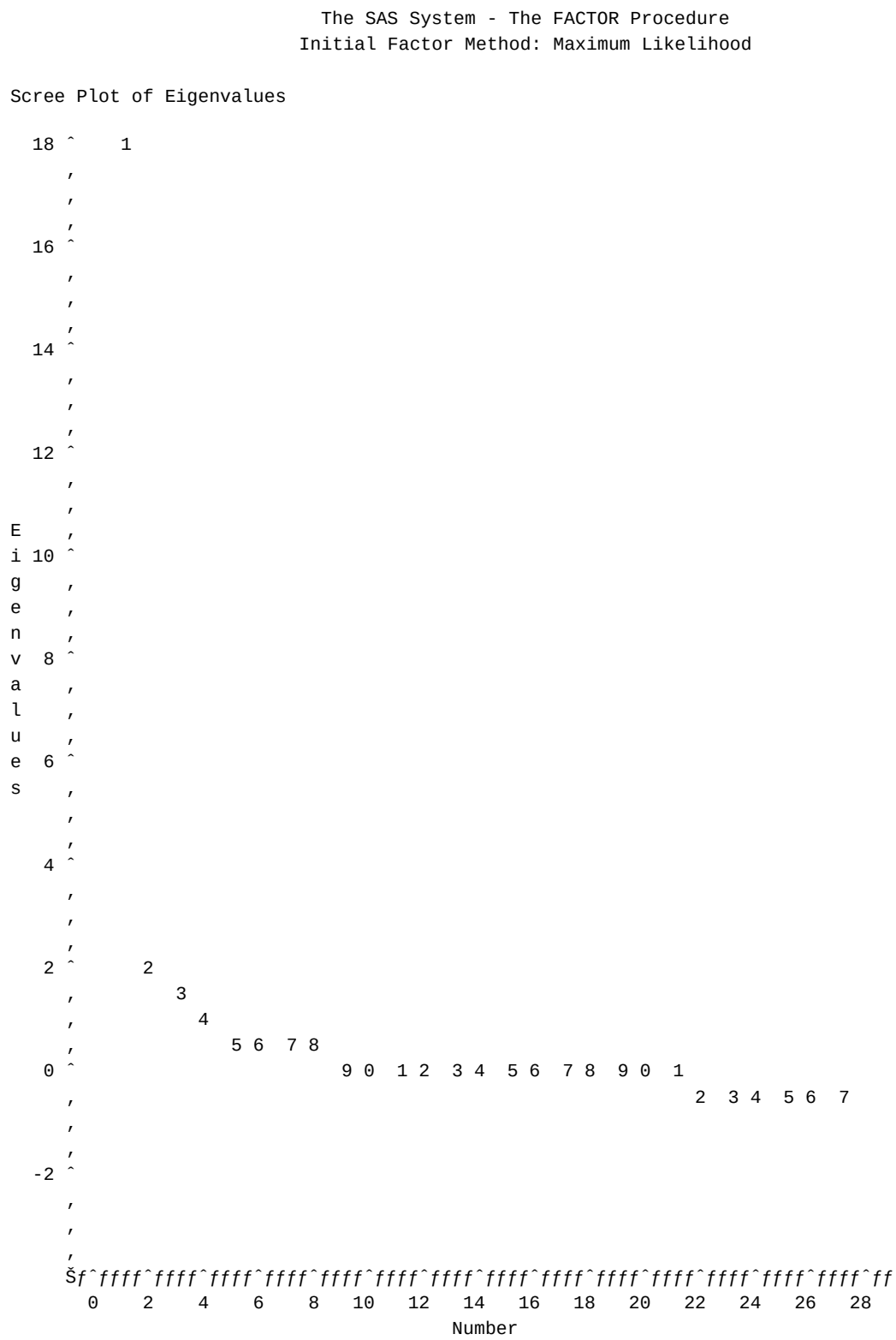
High 1 2 3 4 5 6 7 8 9 10 Low

Appendix B

Item Order on Ratings Form/ Variable name	I = + II = 0 III = -	Description
Troubleshooting and Solving Problems (Soldiering Skills)		
23 TS-1	I	Makes the most of resources at hand; thinks of novel ways to use available materials; invents or fabricates needed items from seemingly useless materials.
5 TS-2	II	Uses available resources to resolve problems and to construct needed items; may occasionally overlook some resources that might have been useful.
14 TS-3	III	Lacks resourcefulness; may simply give up if needed tools are not available or may rely excessively on others to find a way to accomplish a task.
Planning and Preparing for Missions (Soldiering Skills)		
4 PLAN-1	I	Develops plans that are technically sound, well-coordinated, and likely to lead to mission accomplishment; plans are so well-formed that the briefback is readily accepted.
13 PLAN-2	II	Develops workable mission plans that are likely to be successful, although some modification may be needed.
22 PLAN-3	III	Develops plans that have critical flaws or that fail to consider second & third order effects of action; prepares mission analysis that is incomplete of insufficient.
Navigating in the Field (Soldiering Skills)		
20 NAV-1	I	Gets from place to place without errors and on time; without having access to a map, correctly uses terrain features and distances traveled to determine approximate location.
2 NAV-2	II	Usually arrives at destination on time; notices and takes into account map or environmental details to facilitate moving to targets.
11 NAV-3	III	Gets lost and arrives very late to destination or not at all; becomes geographically disoriented or confused when not navigating in daylight conditions (e.g., in darkness, rough or unfamiliar terrain, etc.)
Teaching Others (SF Specific Skills)		
7 TCH-1	I	Creates novel approaches to capture and hold audience attention or to increase audience interest and involvement; incorporates real-life examples in training.
16 TCH-2	II	Uses techniques to maintain attention of the audience during presentations.
25 TCH-3	III	Loses control of the training environment or loses audience attention; may read to audience directly from notes or training materials.
Using and Enhancing Language Skills (SF Specific Skills)		
26 LANG-1	I	Picks up languages readily; uses language skillfully; translates adeptly, rarely, if ever, miscommunicating information; catches errors in others' translations; may create tools (such as a dictionary) for others to use to communicate more effectively.
8 LANG-2	II	Can communicate sufficiently in most situations, even though language skills are not at a conversational level; uses gestures appropriately to enhance communication; uses dictionary to aid in communication when needed.
17 LANG-3	III	Lacks language skills; frequently misunderstands, miscommunicates, or cannot communicate. May simply give up or not try to communicate or learn

Item Order on Ratings Form/ Variable name	I = + II = 0 III = -	Description
Building Effective Relationships with Indigenous Populations (SF Specific Skills)		
18 INDP-1	I	Discovers the needs and desires of HN/G personnel and takes steps to satisfy them, provides special skills and services that enhance HN/G respect for an rapport with SF.
27 INDP-2	II	Helps indigenous persons; provides effective services when asked or when the need is obvious; fixes weapons and provides first aid or other assistance to gain HN/G rapport.
9 INDP-3	III	Overlooks or avoids opportunities to build relations with local, may fail to assist HN/G when rapport could have been built.
Showing Initiative and Extra Effort (Team Member Skills)		
1 EFF-1	I	Puts in whatever time and effort is needed to get the job done; fulfills commitments to multiple projects or missions; overcomes obstacles or unusual difficulties to complete a task or mission.
10 EFF-2	II	Completes task assignments up to standard in a timely manner.
19 EFF-3	III	Leaves work undone to pursue personal interests; procrastinates before starting tasks; fails to follow through on or complete tasks once started.
Handling Interpersonal Situations (Team Member Skills)		
12 INTP-1	I	Deals with others constructively, with tact and diplomacy; is highly adept at persuading others to go along with ideas rather than pushing or forcing own way.
21 INTP-2	II	Is usually polite and courteous toward others; deals effectively with most conflict situations.
3 INTP-3	III	Is inappropriately argumentative ad confrontational, often creating tension and worsening conflict situations.
Contributing to the Team Effort and Morale (Team Member Skills)		
15 TEAM-1	I	Devotes personal time and effort to train team members; teaches unique personal skills to team members to improve their readiness or effectiveness.
24 TEAM-2	II	Makes an effort to motivate other team members through action or words; teaches technical skills in own areas of expertise to team members to ensure team readiness.
6 TEAM-3	III	Puts self-interest and priorities above team welfare; avoids or overlooks opportunities to apply personal or technical skills to benefit the team.
Miscellaneous (added at end-user request; not included in analyses)		
28 / M-1	I	Seeks out opportunity to be cross trained in another SF MOS.
29 / M-2	III	Regularly subject to disciplinary actions.
30 / M-3	III	Makes promises or commitments to HN/G that he cannot deliver.
31 / M-4	I	When required, he would engage and destroy the enemy in accordance with a legal order and established ROE's.
32 / M-5	I	Refuses to give up despite pain, uncertainties, and adversity.
33 / M-6	I	Is proficient in performing the duties of his SF MOS.

Appendix C


The SAS System - The FACTOR Procedure
Initial Factor Method: Maximum Likelihood
Scree Plot of Eigenvalues
Eigenvalues
Number

Appendix D

Eigenvalues of the Weighted Reduced Correlation Matrix: Total = 24.5094287 Average = 0.90775662

	Eigenvalue	Difference	Proportion	Cumulative
1	18.8733350	16.5694068	0.7700	0.7700
2	2.3039282	0.1318607	0.0940	0.8640
3	2.1720675	1.0119678	0.0886	0.9527
4	1.1600997	0.5194806	0.0473	1.0000
5	0.6406191	0.0650374	0.0261	1.0261
6	0.5755817	0.1692200	0.0235	1.0496
7	0.4063618	0.0419290	0.0166	1.0662
8	0.3644328	0.1720770	0.0149	1.0811
9	0.1923558	0.0375376	0.0078	1.0889
10	0.1548182	0.0733172	0.0063	1.0952
11	0.0815010	0.0198950	0.0033	1.0986
12	0.0616060	0.0225875	0.0025	1.1011
13	0.0390185	0.0111319	0.0016	1.1027
14	0.0278867	0.0635213	0.0011	1.1038
15	-0.0356346	0.0212681	-0.0015	1.1024
16	-0.0569027	0.0223888	-0.0023	1.1000
17	-0.0792915	0.0377954	-0.0032	1.0968
18	-0.1170869	0.0333547	-0.0048	1.0920
19	-0.1504415	0.0120950	-0.0061	1.0859
20	-0.1625366	0.0510358	-0.0066	1.0792
21	-0.2135724	0.0263416	-0.0087	1.0705
22	-0.2399139	0.0087625	-0.0098	1.0607
23	-0.2486764	0.0186063	-0.0101	1.0506
24	-0.2672827	0.0139960	-0.0109	1.0397
25	-0.2812787	0.0490573	-0.0115	1.0282
26	-0.3303360	0.0308937	-0.0135	1.0147
27	-0.3612296		-0.0147	1.0000

Appendix E

The FACTOR Procedure
Rotation Method: Promax (power = 3)

Rotated Factor Pattern (Standardized Regression Coefficients)

		Factor1	Factor2	Factor3	Factor4
x1	x1	29	58 *	-3	2
x2	x2	38	45 *	-10	3
x3	x3	-7	13	-4	67 *
x4	x4	56 *	24	6	-6
x5	x5	36	9	-8	-5
x6	x6	-11	39	9	13
x7	x7	65 *	-12	24	-2
x8	x8	20	-1	51 *	-4
x9	x9	-19	20	57 *	9
x10	x10	30	60 *	-5	-3
x11	x11	5	46 *	1	2
x12	x12	18	3	4	61 *
x13	x13	55 *	17	-4	-1
x14	x14	7	57 *	15	-1
x15	x15	48 *	15	11	12
x16	x16	66 *	-17	28	-3
x17	x17	-16	24	63 *	-4
x18	x18	26	-1	51 *	8
x19	x19	1	66 *	4	6
x20	x20	45 *	26	-8	-1
x21	x21	-4	-3	3	81 *
x22	x22	10	45 *	13	1
x23	x23	57 *	11	9	2
x24	x24	48 *	14	9	16
x25	x25	14	31	35	-10
x26	x26	13	-9	54 *	-2
x27	x27	19	4	46 *	9

Printed values are multiplied by 100 and rounded to the nearest integer. Values greater than 0.4 are flagged by an '*'.

Inter-Factor Correlations

	Factor1	Factor2	Factor3	Factor4
Factor1	100 *	55 *	60 *	47 *
Factor2	55 *	100 *	44 *	49 *
Factor3	60 *	44 *	100 *	50 *
Factor4	47 *	49 *	50 *	100 *

Printed values are multiplied by 100 and rounded to the nearest integer. Values greater than 0.4 are flagged by an '*'.

Appendix F

The SAS System
The CORR Procedure

27 Variables: x1 x2 x3 x4 x5 x6 x7
x8 x9 x10 x11 x12 x13 x14
x15 x16 x17 x18 x19 x20 x21
x22 x23 x24 x25 x26 x27

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
x1	1962	2.74567	0.46613	5387	1.00000	3.00000
x2	1962	2.78848	0.43274	5471	1.00000	3.00000
x3	1962	2.66157	0.57718	5222	1.00000	3.00000
x4	1962	2.64526	0.50550	5190	1.00000	3.00000
x5	1962	2.46891	0.61122	4844	1.00000	3.00000
x6	1962	2.60652	0.65691	5114	1.00000	3.00000
x7	1962	2.51835	0.56942	4941	1.00000	3.00000
x8	1962	2.62742	0.54881	5155	1.00000	3.00000
x9	1962	2.79409	0.45992	5482	1.00000	3.00000
x10	1962	2.77676	0.44149	5448	1.00000	3.00000
x11	1962	2.87309	0.38149	5637	1.00000	3.00000
x12	1962	2.57900	0.57050	5060	1.00000	3.00000
x13	1962	2.70591	0.47223	5309	1.00000	3.00000
x14	1962	2.81549	0.43742	5524	1.00000	3.00000
x15	1962	2.56473	0.57319	5032	1.00000	3.00000
x16	1962	2.59480	0.53097	5091	1.00000	3.00000
x17	1962	2.74975	0.51402	5395	1.00000	3.00000
x18	1962	2.68960	0.50291	5277	1.00000	3.00000
x19	1962	2.75484	0.49633	5405	1.00000	3.00000
x20	1962	2.77727	0.44117	5449	1.00000	3.00000
x21	1962	2.68400	0.52187	5266	1.00000	3.00000
x22	1962	2.81091	0.43134	5515	1.00000	3.00000
x23	1962	2.63761	0.52053	5175	1.00000	3.00000
x24	1962	2.61111	0.55419	5123	1.00000	3.00000
x25	1962	2.78287	0.44682	5460	1.00000	3.00000
x26	1962	2.32059	0.63122	4553	1.00000	3.00000
x27	1962	2.73598	0.47219	5368	1.00000	3.00000

Appendix G

The SAS System - The CORR Procedure
Pearson Correlation Coefficients, N = 1962
Prob > |r| under H0: Rho=0

	x1	x2	x3	x4	x5	x6	x7
x1	1.00000	0.56743	0.32435	0.51720	0.21297	0.31251	0.41048
x1		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x2	0.56743	1.00000	0.26246	0.45640	0.22094	0.24524	0.38102
x2	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001
x3	0.32435	0.26246	1.00000	0.25771	0.06845	0.24442	0.21130
x3	<.0001	<.0001		<.0001	0.0024	<.0001	<.0001
x4	0.51720	0.45640	0.25771	1.00000	0.21515	0.20906	0.50095
x4	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001
x5	0.21297	0.22094	0.06845	0.21515	1.00000	0.01651	0.21116
x5	<.0001	<.0001	0.0024	<.0001		0.4649	<.0001
x6	0.31251	0.24524	0.24442	0.20906	0.01651	1.00000	0.17063
x6	<.0001	<.0001	<.0001	<.0001	0.4649		<.0001
x7	0.41048	0.38102	0.21130	0.50095	0.21116	0.17063	1.00000
x7	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	
x8	0.30116	0.29713	0.17323	0.36521	0.13799	0.16318	0.41433
x8	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x9	0.31219	0.24993	0.26178	0.26690	0.05340	0.21441	0.30651
x9	<.0001	<.0001	<.0001	<.0001	0.0180	<.0001	<.0001
x10	0.61850	0.51343	0.27370	0.51097	0.22372	0.27721	0.37737
x10	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x11	0.32597	0.40568	0.26571	0.28736	0.10663	0.22592	0.22787
x11	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x12	0.41596	0.38892	0.48234	0.39252	0.13062	0.25171	0.37541
x12	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x13	0.41525	0.44407	0.21838	0.51977	0.31193	0.16104	0.40600
x13	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x14	0.53004	0.40526	0.31204	0.44645	0.17880	0.31511	0.35141
x14	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x15	0.53021	0.43454	0.30979	0.49112	0.22190	0.26676	0.50257
x15	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001

	x1	x2	x3	x4	x5	x6	x7
x16	0.38076	0.37916	0.20624	0.47874	0.15207	0.15524	0.65624
x16	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x17	0.32589	0.30524	0.22659	0.30973	0.09452	0.19603	0.31101
x17	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x18	0.39398	0.34020	0.26861	0.46128	0.18177	0.21359	0.48734
x18	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x19	0.56353	0.46360	0.30478	0.39506	0.14884	0.37809	0.28206
x19	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x20	0.35921	0.50903	0.17044	0.40926	0.20975	0.16373	0.35020
x20	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x21	0.31091	0.29097	0.54543	0.24368	0.10347	0.23658	0.27864
x21	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x22	0.41251	0.35114	0.25694	0.42657	0.21270	0.26459	0.33906
x22	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x23	0.47323	0.40888	0.25014	0.51895	0.26830	0.20914	0.50159
x23	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x24	0.52893	0.44784	0.34400	0.51029	0.20140	0.28124	0.48398
x24	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x25	0.40559	0.35840	0.22508	0.39257	0.13585	0.21956	0.38846
x25	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x26	0.25126	0.18304	0.14819	0.29906	0.15868	0.11006	0.36881
x26	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x27	0.36665	0.34298	0.23705	0.38507	0.15883	0.20744	0.38786
x27	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001

---

	x8	x9	x10	x11	x12	x13	x14
x1	0.30116	0.31219	0.61850	0.32597	0.41596	0.41525	0.53004
x1	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x2	0.29713	0.24993	0.51343	0.40568	0.38892	0.44407	0.40526
x2	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x3	0.17323	0.26178	0.27370	0.26571	0.48234	0.21838	0.31204
x3	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x4	0.36521	0.26690	0.51097	0.28736	0.39252	0.51977	0.44645
x4	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001

	x8	x9	x10	x11	x12	x13	x14
x5	0.13799	0.05340	0.22372	0.10663	0.13062	0.31193	0.17880
x5	<.0001	0.0180	<.0001	<.0001	<.0001	<.0001	<.0001
x6	0.16318	0.21441	0.27721	0.22592	0.25171	0.16104	0.31511
x6	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x7	0.41433	0.30651	0.37737	0.22787	0.37541	0.40600	0.35141
x7	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x8	1.00000	0.31613	0.29636	0.18323	0.31149	0.32667	0.27642
x8		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x9	0.31613	1.00000	0.25318	0.25497	0.33217	0.20001	0.33829
x9	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001
x10	0.29636	0.25318	1.00000	0.36761	0.36768	0.42851	0.49165
x10	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001
x11	0.18323	0.25497	0.36761	1.00000	0.24642	0.29091	0.38522
x11	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001
x12	0.31149	0.33217	0.36768	0.24642	1.00000	0.33329	0.34248
x12	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001
x13	0.32667	0.20001	0.42851	0.29091	0.33329	1.00000	0.32967
x13	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001
x14	0.27642	0.33829	0.49165	0.38522	0.34248	0.32967	1.00000
x14	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	
x15	0.35147	0.32720	0.46419	0.26729	0.43425	0.38781	0.39749
x15	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x16	0.43014	0.27000	0.36877	0.19915	0.36918	0.39088	0.33004
x16	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x17	0.42853	0.38589	0.30648	0.21243	0.30656	0.22395	0.34566
x17	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x18	0.40481	0.46651	0.36988	0.24111	0.40277	0.40347	0.34919
x18	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x19	0.23736	0.32382	0.55299	0.30960	0.37189	0.30578	0.49150
x19	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x20	0.24470	0.20110	0.39389	0.36220	0.31612	0.40261	0.31809
x20	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x21	0.24571	0.28965	0.29125	0.19547	0.57717	0.23313	0.28729
x21	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001

	x8	x9	x10	x11	x12	x13	x14
x22	0.23432	0.29460	0.42893	0.34373	0.29387	0.36525	0.45555
x22	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x23	0.34469	0.28456	0.45328	0.22538	0.39438	0.44790	0.45423
x23	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x24	0.37510	0.33189	0.45783	0.25607	0.47706	0.39674	0.41698
x24	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x25	0.32291	0.30840	0.40560	0.28700	0.31138	0.30867	0.43155
x25	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x26	0.41416	0.25736	0.22584	0.10128	0.27586	0.23776	0.20695
x26	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x27	0.38569	0.41875	0.35802	0.21305	0.38034	0.34912	0.34670
x27	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001

---

	x15	x16	x17	x18	x19	x20	x21
x1	0.53021	0.38076	0.32589	0.39398	0.56353	0.35921	0.31091
x1	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x2	0.43454	0.37916	0.30524	0.34020	0.46360	0.50903	0.29097
x2	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x3	0.30979	0.20624	0.22659	0.26861	0.30478	0.17044	0.54543
x3	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x4	0.49112	0.47874	0.30973	0.46128	0.39506	0.40926	0.24368
x4	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x5	0.22190	0.15207	0.09452	0.18177	0.14884	0.20975	0.10347
x5	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x6	0.26676	0.15524	0.19603	0.21359	0.37809	0.16373	0.23658
x6	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x7	0.50257	0.65624	0.31101	0.48734	0.28206	0.35020	0.27864
x7	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x8	0.35147	0.43014	0.42853	0.40481	0.23736	0.24470	0.24571
x8	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x9	0.32720	0.27000	0.38589	0.46651	0.32382	0.20110	0.28965
x9	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x10	0.46419	0.36877	0.30648	0.36988	0.55299	0.39389	0.29125
x10	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001

	x15	x16	x17	x18	x19	x20	x21
x11	0.26729	0.19915	0.21243	0.24111	0.30960	0.36220	0.19547
x11	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x12	0.43425	0.36918	0.30656	0.40277	0.37189	0.31612	0.57717
x12	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x13	0.38781	0.39088	0.22395	0.40347	0.30578	0.40261	0.23313
x13	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x14	0.39749	0.33004	0.34566	0.34919	0.49150	0.31809	0.28729
x14	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x15	1.00000	0.49422	0.30684	0.48634	0.43313	0.41096	0.35140
x15		<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x16	0.49422	1.00000	0.32519	0.48742	0.26335	0.33509	0.25907
x16	<.0001		<.0001	<.0001	<.0001	<.0001	<.0001
x17	0.30684	0.32519	1.00000	0.37204	0.29308	0.19483	0.25443
x17	<.0001	<.0001		<.0001	<.0001	<.0001	<.0001
x18	0.48634	0.48742	0.37204	1.00000	0.33035	0.30881	0.35664
x18	<.0001	<.0001	<.0001		<.0001	<.0001	<.0001
x19	0.43313	0.26335	0.29308	0.33035	1.00000	0.31176	0.30122
x19	<.0001	<.0001	<.0001	<.0001		<.0001	<.0001
x20	0.41096	0.33509	0.19483	0.30881	0.31176	1.00000	0.21685
x20	<.0001	<.0001	<.0001	<.0001	<.0001		<.0001
x21	0.35140	0.25907	0.25443	0.35664	0.30122	0.21685	1.00000
x21	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	
x22	0.38676	0.28427	0.32236	0.32874	0.40028	0.34132	0.29170
x22	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x23	0.52047	0.49982	0.26505	0.47006	0.38428	0.35893	0.28406
x23	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x24	0.61145	0.50227	0.31874	0.46321	0.43001	0.36929	0.34892
x24	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x25	0.37149	0.41997	0.40719	0.36712	0.35311	0.30815	0.22390
x25	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x26	0.34078	0.36039	0.39042	0.39556	0.20705	0.18147	0.23029
x26	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x27	0.45507	0.43547	0.34323	0.61032	0.33292	0.31976	0.34002
x27	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001

	x22	x23	x24	x25	x26	x27
x1	0.41251	0.47323	0.52893	0.40559	0.25126	0.36665
x1	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x2	0.35114	0.40888	0.44784	0.35840	0.18304	0.34298
x2	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x3	0.25694	0.25014	0.34400	0.22508	0.14819	0.23705
x3	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x4	0.42657	0.51895	0.51029	0.39257	0.29906	0.38507
x4	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x5	0.21270	0.26830	0.20140	0.13585	0.15868	0.15883
x5	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x6	0.26459	0.20914	0.28124	0.21956	0.11006	0.20744
x6	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x7	0.33906	0.50159	0.48398	0.38846	0.36881	0.38786
x7	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x8	0.23432	0.34469	0.37510	0.32291	0.41416	0.38569
x8	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x9	0.29460	0.28456	0.33189	0.30840	0.25736	0.41875
x9	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x10	0.42893	0.45328	0.45783	0.40560	0.22584	0.35802
x10	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x11	0.34373	0.22538	0.25607	0.28700	0.10128	0.21305
x11	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x12	0.29387	0.39438	0.47706	0.31138	0.27586	0.38034
x12	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x13	0.36525	0.44790	0.39674	0.30867	0.23776	0.34912
x13	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x14	0.45555	0.45423	0.41698	0.43155	0.20695	0.34670
x14	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x15	0.38676	0.52047	0.61145	0.37149	0.34078	0.45507
x15	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x16	0.28427	0.49982	0.50227	0.41997	0.36039	0.43547
x16	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x17	0.32236	0.26505	0.31874	0.40719	0.39042	0.34323
x17	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001

	x22	x23	x24	x25	x26	x27
x18	0.32874	0.47006	0.46321	0.36712	0.39556	0.61032
x18	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x19	0.40028	0.38428	0.43001	0.35311	0.20705	0.33292
x19	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x20	0.34132	0.35893	0.36929	0.30815	0.18147	0.31976
x20	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x21	0.29170	0.28406	0.34892	0.22390	0.23029	0.34002
x21	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001
x22	1.00000	0.33286	0.32366	0.36896	0.19654	0.31309
x22		<.0001	<.0001	<.0001	<.0001	<.0001
x23	0.33286	1.00000	0.55065	0.36971	0.31962	0.43420
x23	<.0001		<.0001	<.0001	<.0001	<.0001
x24	0.32366	0.55065	1.00000	0.39403	0.28807	0.47072
x24	<.0001	<.0001		<.0001	<.0001	<.0001
x25	0.36896	0.36971	0.39403	1.00000	0.32467	0.32757
x25	<.0001	<.0001	<.0001		<.0001	<.0001
x26	0.19654	0.31962	0.28807	0.32467	1.00000	0.36282
x26	<.0001	<.0001	<.0001	<.0001		<.0001
x27	0.31309	0.43420	0.47072	0.32757	0.36282	1.00000
x27	<.0001	<.0001	<.0001	<.0001	<.0001	

Appendix H

```
Mplus VERSION 3.11
MUTHEN & MUTHEN

Level 1 invariance: Team Leaders

SAMPLE STATISTICS
           Correlations
              Y1            Y2            Y3            Y4            Y5
              ____          ____          ____          ____          ____
 Y1             1.000
 Y2             0.552         1.000
 Y3             0.364         0.265         1.000
 Y4             0.484         0.436         0.271         1.000
 Y5             0.132         0.154         0.011         0.153         1.000
 Y6             0.290         0.248         0.219         0.210        -0.016
 Y7             0.374         0.347         0.204         0.512         0.168
 Y8             0.288         0.267         0.182         0.365         0.108
 Y9             0.307         0.222         0.261         0.259         0.041
 Y10            0.583         0.449         0.269         0.493         0.156
 Y11            0.384         0.429         0.258         0.330         0.062
 Y12            0.392         0.397         0.468         0.366         0.104
 Y13            0.436         0.454         0.225         0.566         0.259
 Y14            0.532         0.391         0.268         0.476         0.136
 Y15            0.479         0.406         0.309         0.463         0.197
 Y16            0.331         0.366         0.185         0.492         0.109
 Y17            0.327         0.317         0.206         0.298         0.112
 Y18            0.365         0.341         0.265         0.463         0.141
 Y19            0.565         0.465         0.286         0.399         0.130
 Y20            0.324         0.507         0.153         0.363         0.219
 Y21            0.310         0.274         0.576         0.242         0.025
 Y22            0.446         0.367         0.199         0.486         0.216
 Y23            0.433         0.362         0.241         0.538         0.234
 Y24            0.490         0.444         0.334         0.497         0.171
 Y25            0.401         0.364         0.194         0.400         0.108
 Y26            0.244         0.186         0.181         0.285         0.143
 Y27            0.343         0.364         0.217         0.399         0.158

              Y6            Y7            Y8            Y9            Y10
              ____          ____          ____          ____          ____
 Y6             1.000
 Y7             0.136         1.000
 Y8             0.143         0.475         1.000
 Y9             0.214         0.321         0.305         1.000
 Y10            0.294         0.345         0.306         0.256         1.000
 Y11            0.264         0.242         0.203         0.265         0.382
 Y12            0.233         0.363         0.306         0.340         0.356
 Y13            0.205         0.430         0.393         0.248         0.450
 Y14            0.325         0.378         0.293         0.342         0.499
 Y15            0.261         0.481         0.321         0.311         0.425
 Y16            0.135         0.668         0.452         0.258         0.321
 Y17            0.196         0.331         0.477         0.411         0.303
 Y18            0.167         0.480         0.427         0.461         0.337
 Y19            0.417         0.262         0.250         0.307         0.571
 Y20            0.228         0.322         0.235         0.188         0.336
 Y21            0.221         0.251         0.250         0.310         0.272
 Y22            0.210         0.384         0.283         0.326         0.448
 Y23            0.193         0.498         0.381         0.290         0.411
 Y24            0.266         0.462         0.381         0.338         0.417
 Y25            0.233         0.425         0.369         0.333         0.425
 Y26            0.128         0.389         0.446         0.289         0.232
 Y27            0.204         0.393         0.418         0.394         0.352
```

	Y11	Y12	Y13	Y14	Y15
Y11	1.000				
Y12	0.234	1.000			
Y13	0.376	0.352	1.000		
Y14	0.398	0.350	0.426	1.000	
Y15	0.270	0.411	0.403	0.405	1.000
Y16	0.224	0.361	0.431	0.338	0.470
Y17	0.193	0.344	0.300	0.320	0.325
Y18	0.284	0.424	0.466	0.379	0.438
Y19	0.294	0.349	0.370	0.471	0.426
Y20	0.391	0.310	0.383	0.330	0.384
Y21	0.199	0.606	0.232	0.266	0.330
Y22	0.347	0.325	0.453	0.487	0.428
Y23	0.256	0.377	0.508	0.478	0.481
Y24	0.270	0.475	0.449	0.404	0.608
Y25	0.323	0.308	0.383	0.452	0.376
Y26	0.142	0.288	0.273	0.231	0.355
Y27	0.270	0.391	0.430	0.377	0.434

	Y16	Y17	Y18	Y19	Y20
Y16	1.000				
Y17	0.328	1.000			
Y18	0.502	0.403	1.000		
Y19	0.252	0.275	0.307	1.000	
Y20	0.296	0.206	0.279	0.357	1.000
Y21	0.223	0.255	0.387	0.280	0.198
Y22	0.289	0.320	0.334	0.434	0.393
Y23	0.502	0.293	0.442	0.365	0.296
Y24	0.494	0.362	0.457	0.422	0.353
Y25	0.470	0.427	0.400	0.358	0.316
Y26	0.357	0.423	0.403	0.207	0.194
Y27	0.421	0.389	0.623	0.357	0.297

	Y21	Y22	Y23	Y24	Y25
Y21	1.000				
Y22	0.205	1.000			
Y23	0.246	0.370	1.000		
Y24	0.325	0.358	0.543	1.000	
Y25	0.216	0.386	0.385	0.436	1.000
Y26	0.241	0.243	0.317	0.329	0.332
Y27	0.339	0.367	0.431	0.469	0.369

	Y26	Y27
Y26	1.000	
Y27	0.360	1.000

THE MODEL ESTIMATION TERMINATED NORMALLY

TESTS OF MODEL FIT

Chi-Square Test of Model Fit

Value	1856.939
Degrees of Freedom	318
P-Value	0.0000

Chi-Square Test of Model Fit for the Baseline Model

Value	12695.629
Degrees of Freedom	351
P-Value	0.0000

CFI/TLI

CFI	0.875
TLI	0.862

Loglikelihood

H0 Value	-14614.299
H1 Value	-13685.830

```
Information Criteria
          Number of Free Parameters             60
          Akaike (AIC)                   29348.599
          Bayesian (BIC)                 29646.106
          Sample-Size Adjusted BIC       29455.536
            (n* = (n + 2) / 24)

RMSEA (Root Mean Square Error Of Approximation)
          Estimate                          0.068
          90 Percent C.I.                   0.065  0.071
          Probability RMSEA <= .05          0.000

SRMR (Standardized Root Mean Square Residual)
          Value                             0.053

R-SQUARE
    Observed
    Variable  R-Square
    Y1           0.597
    Y2           0.440
    Y3           0.434
    Y4           0.527
    Y5           0.066
    Y6           0.178
    Y7           0.490
    Y8           0.394
    Y9           0.311
    Y10          0.536
    Y11          0.277
    Y12          0.612
    Y13          0.461
    Y14          0.484
    Y15          0.492
    Y16          0.478
    Y17          0.372
    Y18          0.572
    Y19          0.497
    Y20          0.256
    Y21          0.614
    Y22          0.391
    Y23          0.507
    Y24          0.534
    Y25          0.366
    Y26          0.315
    Y27          0.504


Mplus VERSION 3.11
MUTHEN & MUTHEN

Level 1 invariance: Team Sergeants

SAMPLE STATISTICS
           Correlations
              Y1           Y2           Y3           Y4           Y5
              ____         ____         ____         ____         ____
 Y1            1.000
 Y2            0.581        1.000
 Y3            0.281        0.256        1.000
 Y4            0.546        0.475        0.240        1.000
 Y5            0.303        0.298        0.132        0.287        1.000
 Y6            0.332        0.240        0.267        0.205        0.053
 Y7            0.441        0.413        0.212        0.485        0.261
 Y8            0.311        0.328        0.159        0.362        0.174
 Y9            0.317        0.278        0.261        0.274        0.067
 Y10           0.646        0.571        0.272        0.525        0.296
 Y11           0.270        0.382        0.271        0.244        0.155
 Y12           0.433        0.378        0.493        0.415        0.159
```

	Y1	Y2	Y3	Y4	Y5
Y13	0.386	0.433	0.203	0.472	0.372
Y14	0.531	0.420	0.356	0.418	0.229
Y15	0.577	0.461	0.306	0.517	0.251
Y16	0.426	0.390	0.223	0.462	0.201
Y17	0.324	0.291	0.247	0.321	0.073
Y18	0.416	0.336	0.267	0.455	0.228
Y19	0.560	0.461	0.320	0.389	0.170
Y20	0.387	0.509	0.182	0.452	0.200
Y21	0.302	0.303	0.513	0.239	0.188
Y22	0.379	0.334	0.311	0.365	0.209
Y23	0.508	0.454	0.254	0.496	0.309
Y24	0.567	0.450	0.351	0.522	0.237
Y25	0.416	0.354	0.261	0.387	0.171
Y26	0.262	0.181	0.112	0.316	0.179
Y27	0.380	0.319	0.250	0.366	0.159

	Y6	Y7	Y8	Y9	Y10
Y6	1.000				
Y7	0.204	1.000			
Y8	0.183	0.344	1.000		
Y9	0.214	0.289	0.328	1.000	
Y10	0.259	0.403	0.284	0.250	1.000
Y11	0.188	0.212	0.161	0.244	0.353
Y12	0.268	0.383	0.314	0.323	0.373
Y13	0.115	0.377	0.256	0.151	0.401
Y14	0.305	0.326	0.260	0.334	0.488
Y15	0.271	0.522	0.382	0.343	0.497
Y16	0.174	0.642	0.404	0.281	0.410
Y17	0.195	0.287	0.371	0.356	0.311
Y18	0.259	0.490	0.376	0.472	0.395
Y19	0.338	0.300	0.222	0.341	0.536
Y20	0.096	0.374	0.251	0.213	0.441
Y21	0.248	0.299	0.236	0.268	0.299
Y22	0.318	0.291	0.181	0.262	0.411
Y23	0.223	0.501	0.302	0.278	0.488
Y24	0.296	0.504	0.366	0.324	0.496
Y25	0.206	0.351	0.271	0.281	0.394
Y26	0.089	0.347	0.376	0.220	0.223
Y27	0.208	0.377	0.349	0.445	0.356

	Y11	Y12	Y13	Y14	Y15
Y11	1.000				
Y12	0.256	1.000			
Y13	0.213	0.308	1.000		
Y14	0.374	0.336	0.244	1.000	
Y15	0.263	0.454	0.368	0.390	1.000
Y16	0.173	0.374	0.350	0.322	0.517
Y17	0.232	0.265	0.146	0.373	0.286
Y18	0.197	0.376	0.337	0.319	0.533
Y19	0.323	0.392	0.243	0.512	0.438
Y20	0.333	0.317	0.414	0.307	0.434
Y21	0.189	0.546	0.220	0.308	0.367
Y22	0.340	0.261	0.283	0.425	0.344
Y23	0.194	0.407	0.386	0.431	0.557
Y24	0.240	0.477	0.344	0.431	0.613
Y25	0.251	0.318	0.242	0.411	0.369
Y26	0.057	0.264	0.205	0.181	0.326
Y27	0.158	0.364	0.266	0.318	0.472

	Y16	Y17	Y18	Y19	Y20
Y16	1.000				
Y17	0.321	1.000			
Y18	0.469	0.337	1.000		
Y19	0.272	0.312	0.351	1.000	

	Y16	Y17	Y18	Y19	Y20
Y20	0.371	0.181	0.334	0.265	1.000
Y21	0.289	0.253	0.320	0.317	0.227
Y22	0.278	0.324	0.321	0.366	0.289
Y23	0.495	0.232	0.494	0.401	0.418
Y24	0.509	0.269	0.467	0.437	0.384
Y25	0.368	0.384	0.334	0.350	0.302
Y26	0.366	0.350	0.389	0.208	0.168
Y27	0.446	0.295	0.594	0.307	0.336

	Y21	Y22	Y23	Y24	Y25
Y21	1.000				
Y22	0.370	1.000			
Y23	0.315	0.294	1.000		
Y24	0.369	0.287	0.556	1.000	
Y25	0.236	0.353	0.355	0.349	1.000
Y26	0.221	0.145	0.325	0.240	0.316
Y27	0.333	0.259	0.433	0.470	0.289

	Y26	Y27
Y26	1.000	
Y27	0.370	1.000

TESTS OF MODEL FIT

Chi-Square Test of Model Fit

Value	1494.990
Degrees of Freedom	318
P-Value	0.0000

Chi-Square Test of Model Fit for the Baseline Model

Value	10635.103
Degrees of Freedom	351
P-Value	0.0000

CFI/TLI

CFI	0.886
TLI	0.874

Loglikelihood

H0 Value	-14249.993
H1 Value	-13502.498

Information Criteria

Number of Free Parameters	60
Akaike (AIC)	28619.986
Bayesian (BIC)	28908.793
Sample-Size Adjusted BIC (n* = (n + 2) / 24)	28718.241

RMSEA (Root Mean Square Error Of Approximation)

Estimate	0.064	
90 Percent C.I.	0.061	0.067
Probability RMSEA <= .05	0.000	

SRMR (Standardized Root Mean Square Residual)

Value	0.051

R-SQUARE

Observed Variable	R-Square
Y1	0.651
Y2	0.504
Y3	0.417
Y4	0.511
Y5	0.138
Y6	0.167
Y7	0.488
Y8	0.311

Observed Variable	R-Square
Y9	0.327
Y10	0.603
Y11	0.202
Y12	0.625
Y13	0.294
Y14	0.445
Y15	0.588
Y16	0.476
Y17	0.270
Y18	0.578
Y19	0.470
Y20	0.328
Y21	0.504
Y22	0.289
Y23	0.523
Y24	0.557
Y25	0.280
Y26	0.276
Y27	0.489

Mplus VERSION 3.11
Level 2 invariance - Free measurement model with fixed factor covariance/correlation matrices

TESTS OF MODEL FIT

Chi-Square Test of Model Fit

Value	3358.531	
Degrees of Freedom	642	
P-Value	0.0000	

Chi-Square Test of Model Fit for the Baseline Model

Value	23330.732	
Degrees of Freedom	702	
P-Value	0.0000	

CFI/TLI

CFI	0.880	
TLI	0.869	

Loglikelihood

H0 Value	-28867.593	
H1 Value	-27188.328	

Information Criteria

Number of Free Parameters	114	
Akaike (AIC)	57963.186	
Bayesian (BIC)	58599.503	
Sample-Size Adjusted BIC (n* = (n + 2) / 24)	58237.321	

RMSEA (Root Mean Square Error Of Approximation)

Estimate	0.066	
90 Percent C.I.	0.063	0.068

SRMR (Standardized Root Mean Square Residual)

Value	0.053	

```
Mplus VERSION 3.11
Level 3 invariance - fixed measurement model with free factor covariance/correlations matrices

TESTS OF MODEL FIT
Chi-Square Test of Model Fit
          Value                           3420.509
          Degrees of Freedom                   663
          P-Value                           0.0000

Chi-Square Test of Model Fit for the Baseline Model
          Value                          23330.732
          Degrees of Freedom                   702
          P-Value                           0.0000

CFI/TLI
          CFI                               0.878
          TLI                               0.871

Loglikelihood
          H0 Value                      -28898.582
          H1 Value                      -27188.328

Information Criteria
          Number of Free Parameters            93
          Akaike (AIC)                  57983.164
          Bayesian (BIC)                58502.264
          Sample-Size Adjusted BIC      58206.800
            (n* = (n + 2) / 24)

RMSEA (Root Mean Square Error Of Approximation)
          Estimate                          0.065
          90 Percent C.I.                   0.063  0.067

SRMR (Standardized Root Mean Square Residual)
          Value                             0.061


Mplus VERSION 3.11
Level 4 invariance - fixed measurement model with fixed factor covariance/correlation matrices

TESTS OF MODEL FIT
Chi-Square Test of Model Fit
          Value                           3424.534
          Degrees of Freedom                   669
          P-Value                           0.0000

Chi-Square Test of Model Fit for the Baseline Model
          Value                          23330.732
          Degrees of Freedom                   702
          P-Value                           0.0000

CFI/TLI
          CFI                               0.878
          TLI                               0.872

Loglikelihood
          H0 Value                      -28900.595
          H1 Value                      -27188.328
```

```
Information Criteria
          Number of Free Parameters             87
          Akaike (AIC)                  57975.190
          Bayesian (BIC)                58460.799
          Sample-Size Adjusted BIC      58184.397
            (n* = (n + 2) / 24)

RMSEA (Root Mean Square Error Of Approximation)
          Estimate                          0.065
          90 Percent C.I.                   0.063  0.067

SRMR (Standardized Root Mean Square Residual)
          Value                             0.061
```