Technical Report 1030

Evidence for an Interpersonal Knowledge Factor: The Reliability and Factor Structure of Tests of Interpersonal Knowledge and General Cognitive Ability

Peter J. Legree and Frances C. Grafton
U.S. Army Research Institute

September 1995

United States Army Research Institute for the Behavioral and Social Sciences

Approved for public release; distribution is unlimited.

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency Under the Jurisdiction of the Deputy Chief of Staff for Personnel

EDGAR M. JOHNSON
Director

Technical review by

Robert N. Kilcullen
Douglas MacPherson

NOTICES

DISTRIBUTION: Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to: U.S. Army Research Institute for the Behavioral and Social Sciences, ATTN: PERI-POX, 5001 Eisenhower Ave., Alexandria, Virginia 22333-5600.

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U.S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave Blank)	2. REPORT DATE 1995, September	3. REPORT TYPE AND DATES COVERED FINAL 8/93 - 6/95

4. TITLE AND SUBTITLE
Evidence for an Interpersonal Knowledge Factor: The Reliability and Factor Structure of Tests of Interpersonal Knowledge and General Cognitive Ability

5. FUNDING NUMBERS
262785A
A791
1211
H1

6. AUTHOR(S)
Peter J. Legree and Frances C. Grafton

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
U.S. Army Research Institute for the Behavioral and Social Sciences
ATTN: PERI-RS
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Research Institute for the Behavioral and Social Sciences
5001 Eisenhower Ave.
Alexandria, VA 22333-5600

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
ARI Technical Report 1030

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words):

Many aptitude scales measure general or academic knowledge and utilize a forced choice response format in which answers are scored as either correct or incorrect. In contrast to this traditional scoring procedure, quantifying performance on scales developed to measure interpersonal skills requires the opinions of multiple experts, and individual responses cannot be easily or unambiguously evaluated. Given this type of uncertain knowledge domain, a Likert procedure was modified to measure expertise based on the distance between expert and subject ratings of the relative strengths of a set of probabilistic relationships. In Phase 1, data were collected and indicate that an improvement in the reliability of an existing measure of leadership could be traditional forced choice format. In Phase 2, data were collected with the leadership scale and two additional interpersonal knowledge scales using Air Force recruits for whom Armed Services Vocational Aptitude Battery (ASVAB) data were available. Confirmatory factor analyses indicate that the factor structure of the 13-test battery (ASVAB plus the experimental scales) could be best explained by hypothesizing the existence of a separate interpersonal knowledge factor in addition to the four factors that are typically extracted from the ASVAB. These results demonstrate (1) the applicability of the Likert response format to efficiently measure individual differences in nontraditional knowledge domains such as interpersonal skills, and (2) the existence of a separate first-order factor that is labeled Interpersonal Knowledge.

14. SUBJECT TERMS
Low fidelity simulation
Likert response format
Tacit knowledge scales
Personality
Social intelligence
Temperament

15. NUMBER OF PAGES 51

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

NSN 7540-01-280-5500

Standard Form 298 (Rev 2-89)
Prescribed by ANSI Std 239-18 298-102

Technical Report 1030

Evidence for an Interpersonal Knowledge Factor: The Reliability and Factor Structure of Tests of Interpersonal Knowledge and General Cognitive Ability

Peter J. Legree and Frances C. Grafton
U.S. Army Research Institute

Selection and Assignment Research Unit
Michael G. Rumsey, Chief

Personnel and Training Systems Research Division
Zita M. Simutis, Director

U.S. Army Research Institute for the Behavioral and Social Sciences
5001 Eisenhower Avenue, Alexandria, Virginia 22333-5600

Office, Deputy Chief of Staff for Personnel
Department of the Army

September 1995

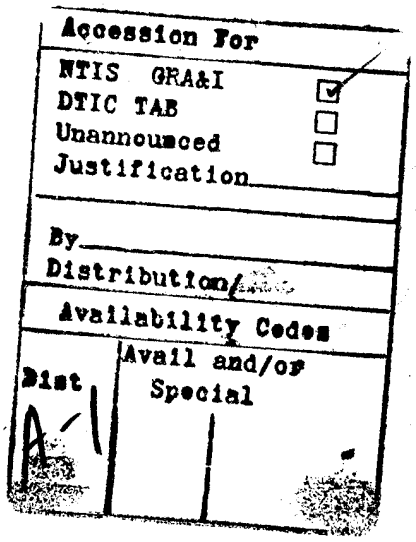
Army Project Number
2O262785A791

Education and Training Technology

FOREWORD

The U.S. Army has embarked upon a line of research to evaluate and improve its existing personnel selection and classification system. Toward this goal, the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) is committed to the exploration of alternative personnel testing and evaluation procedures. As part of this effort, this report describes and evaluates a methodological approach to develop tests for nontraditional content domains.

EDGAR M. JOHNSON
Director



ACKNOWLEDGMENTS

The author would like to recognize the contributions of colleagues who contributed to this report, including Dr. Michael Drillings and Dr. Douglas K. Detterman, who helped identify and formulate research questions, and Frances Grafton, who contributed substantial time to evaluate numerous theoretical issues and review drafts of this report. The research was supported by the Research and Advanced Concepts Office of the U.S. Army Research Institute for the Behavioral and Social Sciences. Data collection was possible through the cooperation of the U.S. Air Force Armstrong Laboratory under the Training and Personnel Systems Science and Technical Evaluation Management (TAPSTEM) initiative.

EVIDENCE FOR AN INTERPERSONAL KNOWLEDGE FACTOR: THE RELIABILITY AND FACTOR STRUCTURE OF TESTS OF INTERPERSONAL KNOWLEDGE AND GENERAL COGNITIVE ABILITY

EXECUTIVE SUMMARY

Requirement:

The Army strives toward efficient personnel selection and classification methods. Although considerable progress has been made over the years, the more the Army can learn about testing procedures, the more effective its personnel management decisions can be. The goal of this effort was to explore and develop methodological approaches that could be used to improve personnel testing technology. This report describes the application of a Likert-based procedure to measure interpersonal knowledge.

Procedure:

The impact of a Likert-based testing format was evaluated for a supervisory ability scale that had been developed with a traditional response format. The Likert procedure was also used to develop tests for two content domains corresponding to commonly occurring interpersonal situations. The psychometric properties and factor structure of these scales was evaluated with a population of Air Force recruits.

Findings:

The data indicate that the Likert-based testing method could be used to develop tests in a cost-effective manner that efficiently measures knowledge for nontraditional content domains. Analyses of these knowledge scales indicate a separate factor that has a substantial loading on psychometric g. This factor may correspond to the interpersonal content of the scales or to the ability to quantify uncertain or probabilistic relationships.

Utilization of Findings:

The approach explored in this research can be applied to the development of maximal performance measures for other nontraditional content domains such as personality and temperament. Additional research is suggested to better identify the nature of the demonstrated factor, social versus probabilistic ability, and to explore cultural and group differences.

EVIDENCE FOR AN INTERPERSONAL KNOWLEDGE FACTOR: THE RELIABILITY AND FACTOR STRUCTURE OF TESTS OF INTERPERSONAL KNOWLEDGE AND GENERAL COGNITIVE ABILITY

CONTENTS

LIST OF TABLES

EVIDENCE FOR AN INTERPERSONAL KNOWLEDGE FACTOR: THE RELIABILITY AND FACTOR STRUCTURE OF TESTS OF INTERPERSONAL KNOWLEDGE AND GENERAL COGNITIVE ABILITY

The research described in this report explores the hypothesis that individual differences in interpersonal skills reflect a social intelligence factor that is moderately to highly correlated with psychometric g. Past attempts to demonstrate a social intelligence factor have been only partially successful in that, while social intelligence factors have been extracted, scales loading on these factors have consisted of behavior rating scales completed by peers and teachers (Tisak & Ford, 1983; Marlowe, 1986). In general, factor analyses of aptitude-based scales, i.e., measures of maximum performance, have not provided convincing evidence for a separate social intelligence factor (Walker & Foley, 1973; Keating, 1978).

Consistent with the factor analyses of the behavior rating scales, which may be described as measures of typical performance, is the demonstration that individuals perceive social and practical intelligence as distinct from academic intelligence (Sternberg, Conway, Ketron & Bernstein, 1981). That society views social intelligence as separate from general intelligence is also apparent in long-standing legal definitions of mental retardation that require special services to be provided for children scoring below a specified level on both a social functioning scale and on a general intelligence scale (Hallahan & Kauffman, 1991).

One important consideration relating to the development of social intelligence scales is that it is difficult to justify imposing a value structure to score a Social Intelligence scale, but some sort of structure is required to score any scale. This limitation makes measuring social intelligence difficult and could lead to considerable controversy. In contrast, items on traditional intelligence scales can often be linked to an academic knowledge base and there is less potential controversy. It seems relevant to suggest that the concept of a single correct action for a complicated social situation may often be unrealistic; rather several fairly appropriate actions are available for most social situations and identifying a single correct answer may not be logical from a testing perspective.

Low Fidelity Simulations And Situational Judgment Scales

Although not readily available, situational judgment scales have been developed to measure individual differences in interpersonal skills in a number of areas including telephone sales representative skills (Phillips, 1992), collection agency

negotiation skills (Phillips, 1993), administrative and interpersonal skills of educators (Ostroff, 1991), leadership skills of Non-Commissioned Officers (Hanson & Borman, 1992), and managerial skills (Motowidlo, Dunnette & Carter, 1990). These scales can be described as either situational judgment scales or low fidelity simulations of work sample tasks.

All of these scales are similar in that a forced choice format was adopted and each test item is composed of a relatively long problem scenario and a number of actions that might be followed to try to resolve the dilemma. The subject is required to identify the most appropriate action, and sometimes the least appropriate action, based on the problem description, professional knowledge and past experience. A response is scored as either correct or incorrect on the basis of agreement with subject matter expert (SME) opinions.

Measuring performance with a low fidelity simulation may appear analogous to testing academic knowledge in that the correct response is externally verified and a response is scored as either correct or not correct. In addition, the low fidelity simulation methodology allows item level statistics to be computed to identify problematic items and refine the instrument. However, there are several important differences between testing academic knowledge and assessing individual differences in interpersonal skills via a low fidelity simulation.

Item Length

One important difference relates to test item length. Academic test items can often be written with relatively short item stems and response choices. In fact, brevity may be recommended to avoid ambiguity in the test, limit the effect of test-wiseness and maximize the number of items for which data can be collected within some fixed testing period. One implication of this terseness is that the reading requirement of the average item tends to be relatively minimal for many academic knowledge tests.

In contrast, the description of a problem scenario for a low fidelity simulation is often lengthy because of the ill-defined and complex nature of these items. Short problem scenarios often cannot support the complexity inherent in interpersonal conflicts and problem scenarios. The response choices also tend to be necessarily longer than those distractors for an academic knowledge test. Thus simulating complex problem scenarios can result in extremely long tests given the number of items for which data are collected. This tendency towards lengthy elaboration is evident in the Army Situational Judgment Test, which contains 49 test items and averages one-half page of text (169 words) per item.

Scoring

Another important difference between a standard academic knowledge test and a situational judgment scale involves the procedures used to identify the correct response for a specific item. The scoring key of an academic knowledge test can usually be verified by referencing explicit facts derived from academic theories or listed in reference books. These facts are used to develop a scale that can then be pilot tested. Item level statistics may be computed to identify and either modify or delete problematic test items.

In contrast to the fact-based scoring procedure used for an academic test, scoring a situational judgment scale must often be based on Subject Matter Expert (SME) opinions. Because SME do not always agree, a relatively large number of expert opinions may be required to produce a credible scoring key. For example, Phillips (1992, 1993) required that 75 percent of approximately 20 experts agree that a specific response was "most appropriate", i.e., "correct", in order for that scenario (and response) to be included on a situational judgment scale.

Although 75 percent may appear to be a reasonable agreement criterion, the implication is that for any specific item, up to 25 percent of the experts may disagree as to the most appropriate, i.e., "correct", response alternative. It is relevant to note that with a 75 percent agreement criterion, which is equivalent to an "up to" 25 percent disagreement criterion, the performance of many experts would be far from perfect on a situational judgment scale. In contrast, near perfect performance would often be expected of experts on an academic knowledge test.

It is tempting to conclude that while verification of the correctness of a response alternative is relatively straightforward for academic knowledge, the procedure is quantitatively more complex when applied to the practical knowledge underlying situational judgment scales. However, this interpretation belies the possibility that these two types of knowledge may be qualitatively different. It can be argued that the correctness of an (exam) assertion, given some academic knowledge base (facts and theory), is usually unambiguously dichotomous, i.e., either correct or not correct.[1] In contrast, situational judgment scales attempt to simulate everyday problem situations but

---

[1] It would be possible to create an academic test that has intentionally ambiguous answers. For example, a test could require students to rate the relative clarity of 20 sentences. However, I know of no academic test or scale that utilizes this format. Instead, ambiguous test items are generally dropped after pretest analyses are completed.

usually cannot present enough information to allow the formulation of unambiguously "correct" solutions. This is not to suggest that if low fidelity simulations could present additional information, then the ambiguity would disappear. This ambiguity partially reflects real-world interpersonal interactions; these are often ambiguous because behavior can be multidetermined and because individuals are dynamic, complex and sometimes disingenuous. It follows that one qualitative difference between academic and everyday interpersonal knowledge is the presence or absence of the certainty that can be attached to the correctness of specific assertions or to the likely result of specific actions given a particular situation or problem.

It is important to recognize that as a general rule, the "correct" response alternative for a low fidelity simulation scenario cannot be guaranteed to lead to a satisfactory resolution of the simulated problem. Nonetheless, experts will generally agree that some alternatives are much more likely to result in a reasonable solution. It seems plausible to describe this type of interpersonal knowledge as probabilistic, fuzzy, or uncertain. A more veridical simulation of expertise for an ill-defined problem situation might require subjects to estimate the relative quality of proposed solutions and compare these estimates to expert ratings. This type of task recognizes and models this qualitative difference between general-academic and interpersonal knowledge.

Likert Scales

An alternative to the use of the forced choice format is evident in the tacit knowledge scales developed by Wagner and Sternberg (1986). These scales match a single scenario with approximately ten actions and the subjects must rate the appropriateness of all the actions on a Likert scale. The subject ratings can then be transformed to eliminate response bias, and a distance is calculated for each item between the transformed subject and expert ratings.

One practical advantage to the Likert format is that one datum is collected for each response alternative as opposed to one or two data per scenario. Therefore this format can be used to collect much more data per unit of text than is possible with the conventional forced choice format. For example, a typical scale developed by Wagner and Sternberg (1986) yields ten data points per scenario and requires approximately one page of text. In comparison, the Army Situational Judgment Test produces between two and four data points per page of text.

Another practical advantage to the Likert format is that interval data are computed for each item, i.e., the distances between the subject and expert ratings. This distance quantifies the correctness of the subject's response for a particular item

(response alternative) and allows the response to be characterized as varying along the dimension of correctness. In comparison, only dichotomous data are collected with the forced choice format with a concomitant loss of precision in the estimate.

A conceptual question relating to response format addresses the nature of the task. Most well-designed academic tests, which utilize a forced choice format, present a single correct answer per item. The primary purpose of the distractors is to limit the effect of guessing. On this type of scale, the subjects' task can be argued to be primarily an identification task in that a very knowledgeable subject can respond as soon as a correct response is read.

Unlike a conventional academic multiple choice test, the alternatives for most situational judgment scales are selected to range in correctness (i.e., appropriateness) with several "good" and several "bad" alternatives. When the forced choice format is adopted for a situational judgment scale, the task can no longer be considered an identification task because all the alternatives may be somewhat correct without any being optimal. Due to the ambiguity in the problem scenario, none of the actions may be necessarily "best" or even "good." The subject is presented with a comparison task that requires the understanding of nuances of vocabulary and meaning, rather than the direct application of the subject's expertise. In addition, this type of task has a substantial memory and reading requirement in order for the lengthy alternatives to be compared. This suggests that the psychometric properties of an existing situational judgment scales could be improved by utilizing the Likert response format to collect data in a more efficient manner. This type of modification could be implemented with only minor changes to many existing scales.

General Research Objectives

At present, there is no literature that empirically estimates the effect of utilizing the Likert response format on either the reliability of a situational judgment scale or on the empirical relationship of this type of scale to related constructs such as general intelligence. On the basis of classical test theory, i.e., the Spearman-Brown Prophecy formula, it can be hypothesized that an improvement in the reliability of a situational judgment scale would be realized by substituting the Likert format for the forced choice format because the amount of information collected by the scale is greater. One goal of this research was to determine whether or not the Likert response format represents a viable method to improve the reliability of an existing situational judgment scale.

A related question is to determine whether the Tacit Knowledge format (Wagner & Sternberg, 1986), could be used to develop interpersonal knowledge scales in a cost effective manner. Most low fidelity simulations are based on the collection of critical incident data that are analyzed and represented in the form of problem scenarios and possible solutions. Additional subject matter experts are then required to rate the problems and verify the scoring key. In contrast, descriptions of the development of Tacit Knowledge Scales (Wagner & Sternberg, 1986) suggest that scales can be developed in a more cost-effective manner with this format.

Another general goal of this project was to produce a preliminary factor structure defined by general aptitude scales and by low fidelity scales that appear to measure interpersonal skills and knowledge. This was accomplished in Phase 2 by collecting data with an existing situational judgment scale and two additional interpersonal knowledge scales that were developed to capitalize on the efficiencies inherent to Wagner and Sternberg's methodology (1986), using subjects for whom recent ASVAB data were available.

Phase 1

As stated, one objective of this research was to estimate the extent to which the reliability of an existing situational judgment scale could be improved by utilizing a Likert response format. This was accomplished by modifying an existing instrument and testing two groups of subjects using either the Likert or the forced choice response format.

The Army Situational Judgment Test (SJT) was developed to support Project A research[2] as a test of NCO supervisory ability (Campbell & Zook, 1991). The SJT was selected for modification because it is typical of situational judgment scales in length and response format; in addition, the reported reliability estimates for the scale are in the moderate range. Refer to Hanson and Borman (1992) for information concerning test construction.

Data were collected at the U.S. Air Force Armstrong Data Collection Facility at Lackland AFB. The Lackland subject pool consists of Air Force recruits in their 21st day of basic

---

[2] Project A was a seven year effort designed to validate and improve the procedures used to select and classify Army soldiers. One aspect of Project A was the development and validation of new and existing predictors against new and existing job related criteria including the SJT.

training. The use of this population necessitated that the SJT content be slightly modified by substituting specific Air Force terms for the Army equivalents. For example, the Air Force rank, Airman, was substituted for the equivalent Army rank, Private.

One concern with using the SJT is that it might be too difficult to use as an individual difference measure when administered to Air Force recruits because it was developed as a Project A criterion to measure the supervisory ability of U.S. Army Non-commissioned Officers (NCO). In other words, a floor effect might occur and result in the attenuation of the reliability of the scale. However, many of the SJT items appear to tap a general interpersonal domain that might transcend the military-specific experiences of U.S. Army NCO and data had never been collected with the SJT using a group of subjects with little military experience.

One goal of Phase 1 was to determine the feasibility of using the SJT to collect individual difference data given an Air Force recruit population. This issue is important because a cost-effective procedure to develop measures of social intelligence would be to refine existing instruments. The feasibility issue was addressed by altering the terminology used in the SJT to reflect Air Force vernacular and collecting data on two small groups of subjects utilizing either a forced choice or a Likert response format for each group.

Materials

The SJT consists of 49 problem scenarios with between three and five solutions proposed for each scenario. For Project A, the correct response for each scenario on the SJT had been identified on the basis of SME ratings. The same answer key was used to score the SJT for this research.

The Project A procedure required the subject to read each problem scenario and then to identify the alternative that the subject felt was most appropriate and the alternative that the subject felt was least appropriate. In the current research, the forced choice version of the SJT was administered in accordance with instructions and scoring procedures that were essentially identical to those used for Project A.

The forced choice version was adapted to the Likert format by appending the following stem to the end of each scenario, "Please rate the appropriateness of the following actions". The instructions for both conditions, which include an example scenario, are contained in Appendix A. Note that the Likert response format requires the subject to rate each response alternative, as opposed to selecting the most or least appropriate response.

The subjects were required to rate the appropriateness of each action on an 11 point bipolar scale. The ends of the scale were anchored with the terms "Extremely Inappropriate" and "Extremely Appropriate", the midpoint was labeled "Neither Appropriate Nor Inappropriate". An 11-point scale was used in recognition of the possibility that some scenarios might contain only appropriate or inappropriate alternatives; in such a case, it was felt that a larger interval scale would allow subjects to make finer gradations in their ratings. In addition, Wagner and Sternberg (1986) successfully utilize an 11-point scale; recognizing the logic of adopting and extending a proven technique, the larger scale was utilized throughout this research.

Scoring the Forced Choice SJT

Hanson and Borman (1992) describe a number of ways to score the forced choice version of the SJT including: proportion of "most" appropriate hits, proportion of "least" appropriate hits, mean effectiveness SME rating of actions selected as "most" appropriate responses, mean effectiveness SME rating of actions selected as "least" appropriate responses, and the difference between the SME ratings of the "most" and "least" appropriate responses for each scenario. Other situational judgment scales have tended to adopt the simplest of these scoring procedures, i.e., the proportion correct "most" measure.

The two proportion measures were calculated by defining the response alternatives that were rated highest and lowest by the SME as the correct "most" appropriate and "least" appropriate response for each scenario. Individual difference scores were calculated as the proportion of correct responses for the two dimensions.

The mean effectiveness rating scoring procedures weighted the "most" and "least" appropriate response for each scenario by the mean SME ratings for those responses. Thus if a subject selected a response alternative with a mean SME rating of 5.27 as the "most" appropriate response, then a value of 5.27 would be assigned for that item. Accordingly, better performance is indicated by higher scores for the "most" appropriate responses and by lower scores for the "least" appropriate responses. The difference measure was calculated by averaging (across scenarios) the difference in the weightings associated with the "most" appropriate and "least" appropriate responses. In this study, all five procedures were used.

Scoring the Likert SJT

The procedure used to score the Likert version of the SJT is dissimilar from any typically used to calculate individual

differences on ability scales[3]. The procedure produces interval data for each item as a function of the distance between the subject's rating and the mean expert rating for that response alternative. The average distance across items is then computed to estimate individual differences in performance on the task. However, several transformations of the data are required to eliminate response bias and to score tests for which answer keys are not available in Phase 2. (This second point will be discussed under Phase 2.)

Response bias is an important issue because the scoring procedure is intended to quantify individual differences in the ability to estimate the relative appropriateness of alternate solutions given a specific problem scenario. If ignored, response bias could have a dramatic effect for subjects who use only part of the rating scale. For example, if the ratings of a particular subject were biased towards the "Inappropriate" segment of the scale, then the distances calculated for all but the most inappropriate alternatives would be overestimated.

To resolve the response bias problem, the ratings produced by each subject were transformed to yield standard scores with a mean of 0.0 and a standard deviation of 1.0. A similar transformation was conducted on the expert ratings of the effectiveness of the alternatives described for the scenarios. These SME ratings had been collected as part of Project A. A distance was then calculated for each item as the square of the difference between the transformed expert and subject ratings. Individual difference scores were computed as the mean item distance for each subject. Using this procedure, better performance is indicated by lower values.

Subjects

Forty-eight male Air Force recruits in their 21st day of basic training at Lackland AFB participated in this study. Twenty-four subjects were assigned to each group.

Procedure

Data were collected after breakfast over a two week period between 7:00 and 9:00 AM. Subjects were alternately assigned to a condition, i.e., Likert versus forced choice. The subjects were seated in a classroom and instructed to follow the

---

[3] Some important terminology changes must be noted. A Likert alternative corresponds in content to a Forced Choice response alternative, but data are collected for all Likert alternatives while most response alternatives are distractors within the Forced Choice format. The terms "distractor" and "p-value" are meaningless from a Likert perspective.

instructions in the SJT test book. The recruits were tested in groups of up to 20 subjects and were told to wait at their desks until the session was completed.

Results

Reliability estimates were calculated for the two versions of the SJT and suggest an increase in the reliability estimate of the Likert format relative to the forced choice format. Table 1 contains estimates of the reliability, the mean performance of the subjects and the standard deviation of performance by scoring procedure.

Table 1. Test Means, Standard Deviations and Internal Consistencies.

Scoring Procedure	Mean	SD	Reliability
Forced Choice Format (49 scenarios)			
Most Proportion Correct	.46	.07	.27
Least Proportion Correct	.45	.08	.31
Most Weighting	4.79	.21	.37
Least Weighting	3.38	.17	.26
Difference Weighting	1.41	.36	.51
Refined (42 scenarios)			.65
Likert Format (202 items)			
Alternative Level	1.19	.18	.62
Refined (145 items)			.84

It may be of some interest that one advantage to the Likert format is that items may be eliminated at the alternative level. Of course, an analogous procedure can be followed at the scenario level for the forced choice format, but this requires eliminating an entire scenario and all the associated alternatives. To demonstrate the effect of this procedure, total and item scores were correlated across the 202 Likert items. Fifty-seven items with negative full scale correlations were eliminated from the Likert scale and the reliability of the new scale was estimated

to be .84 (Refer to Table 1).

For the purpose of comparison, the analogous procedure was followed for the most reliable scoring method that is associated with the forced choice format, i.e., the Difference weighting procedure. Seven scenarios with negative total score correlations were deleted from the scale, with the result that the reliability of the revised scale increased to .65 (Refer to Table 1).

As noted in the introduction to Phase 1, one major difference between this research and Project A is that the Project A subjects had substantial military experience when the SJT was administered, while the Air Force recruits were in their 21st day of basic training. This difference could be important because the SJT scenarios require the subjects to assume the role of a military supervisor confronted with a variety of personnel and supervisory problems.

If performance on the SJT reflects either explicit military doctrine or general military experience, then there should be a substantial difference in mean performance between the Army and Air Force samples. Descriptive statistics obtained from Hanson and Borman (1992) and computed for the Force Choice condition in this research are reported in Table 2. Effect sizes were calculated in accordance with the approach described by Bloom (1984) and are reported in Table 2. The Army variance estimates were used because these are based on a much larger sample size. The effect size estimates indicate that performance on the SJT is not highly influenced by military experience. Three of the five comparisons, including the most reliable scale, actually favor the Air Force subjects. These comparisons, however, must be interpreted cautiously because: (1) the small sample size associated with the Air Force population, and (2) the summary statistics reported by Hanson and Borman were based on a shorter version of the SJT that contained 35 scenarios.

One question revolving around the use of SME ratings to reference subject performance relates to the relationship between the Army SME and the Air Force subject mean ratings across the 202 Likert alternatives. Recall that the Army means are used as a reference pattern to score the SJT-Likert scale. Agreement in mean ratings was assessed by correlating the two sets of mean ratings ($r_{AF,Army}$=.72, p<.01).

This correlation, r=.72, could not be directly corrected for attenuation of reliability because the reliability of the mean Army ratings could not be located. However, a working paper was found that suggests that each mean Army rating was based on the ratings of 6 NCO. (In fairness to Project A, this would be adequate for the purpose for which the ratings were collected.) Assuming a similar level of agreement between the NCO raters

(n=6) and the Air Force recruits (n=24), the Spearman-Brown correction can be used to estimated the reliability of the mean NCO ratings at r_{xx}=.65, based on the value calculated for the Air Force sample, r_{xx}=.88, and the ratio between the number of observations for each group, .25. The reliability estimates can then be used to estimate the parameter correlation between the Army NCO mean ratings and Air Force recruit mean ratings based on the correction for attenuation of reliability, r_{tt}=.95.

Table 2. Comparison of Army and Air Force Levels of Performance.

Scoring Procedure	Army Mean (SD)	Army r_{xx}	Air Force Mean (SD)	Air Force r_{xx}	Effect Size[1]
Most Proportion Correct	.47 (.12)	.60	.46 (.07)	.27	.08
Least Proportion Correct	.42 (.11)	.57	.45 (.08)	.31	-.27[2]
Most Weighting	4.91 (.34)	.68	4.79 (.21)	.37	.35
Least Weighting[3]	3.54 (.31)	.68	3.38 (.17)	.26	.52[2]
Difference Weigthing	1.36 (.61)	.75	1.41 (.36)	.51	-.08[2]

[1] Calculated in accordance with Bloom (1984) with the Army SD as the reference value.
[2] The difference favors the Air Force sample.
[3] Low scores indicate better performance.

Discussion

The results of Phase 1 are consistent in support for the feasibility of obtaining analyzable data by administering the SJT to Air Force recruits. To summarize the major points: (1) the reliability data indicate adequate levels of reliability for the Likert and the more sophisticated forced choice scoring procedures for this population; (2) the mean performance level of the Air Force recruits was similar to that of the Army NCOs on the forced choice version; and (3) the untransformed mean ratings of the Air Force recruits and Army NCO's were highly correlated.

In interpreting the mean performance data, it is notable that the SJT was developed to measure general supervisory knowledge and the development of the SJT did not utilize explicit military doctrine. Instead, Army supervisors were contacted to identify problem situations requiring practical supervisory knowledge and abilities, as opposed to situations requiring

knowledge of explicit Army doctrine. The fact that the SJT scenarios reflect non-explicit supervisory knowledge necessitated that the correct "most" and "least" responses be based on NCO ratings.

One interpretation of the mean performance data is that the SJT measures general supervisory knowledge that transcends military or civilian settings. According to this interpretation, the knowledge and ability measured by the SJT can be gained through a variety of interpersonal experiences and only small differences in performance should be expected in the comparison of the Army and Air Force samples. This interpretation is further strengthened by the extremely high correlation between the mean untransformed ratings of the Army NCOs and Air Force recruits ($r_{AF,Army}$=.72) and the associated corrected correlation estimate (r_{tt}=.95). These values suggest that the SJT does not tap knowledge that can only be gained through military experience, rather the data are more consistent with the view that the SJT measures general supervisory and interpersonal knowledge.

A comparison of the Phase 1 reliability estimates suggests that the Likert format, when utilized in place of the forced choice format, results in more reliable individual difference estimates. This is to be expected because the number of data points is increased. Increasing the amount of collected data should, according to the Spearman-Brown prophecy formula, result in a more reliable scale. This demonstration would be trivial if the increase in reliability was simply due to increasing the length of the scale in a manner that increases the time required to administer the scale; but this was not the case. The reliability data are important specifically because the additional data were collected with only minimal differences in the length of the SJT text, i.e., the amount of information that was presented to the subjects.

Phase 2

In Phase 2, additional data were collected at Armstrong Labs in order to verify the results from Phase 1. Phase 2 was structured to extend the conclusions from Phase 1 and allow the hypothesized factor structure of three tests of interpersonal knowledge to be evaluated. This section is prefaced with a description of the development and rational used for Phase 2.

More Accurately Estimating Reliability Parameters.

One major implication of Phase 1 is that the SJT could be used to collect reliable individual difference data given a population with little military experience, e.g., Air Force

recruits. This demonstration is important to the development of a battery of interpersonal knowledge scales because the SJT was developed as an NCO criterion and the presence of a floor effect would limit the utility of the scale for this purpose. Furthermore it is logical to verify the Phase 1 results before investing too much effort in the refinement of the SJT.

From the Phase 1 reliability analyses, it seems reasonable to expect that an improvement in the reliability of the SJT would be realized by substituting the Likert format for the standard forced choice format. However, the reliability estimates were based on small sample sizes (n=24 per group), and it is possible that the observed differences reflect sampling error. In this regard, it is notable that inferential statistics are not typically used to test reliability differences because these are generally viewed as parameter estimates (McNemar, 1969).

A larger sample size was specified in the Phase 2 data collection effort in order to more accurately estimate the SJT forced choice and Likert reliability parameters. Based on the Phase 1 analyses, it was expected that the Likert version would be more reliable than the traditional forced choice version.

Efficiently Developing Interpersonal Knowledge Scales

From a methodological perspective, an important goal is the exploration of test construction options to score the SJT and simplify the development of other interpersonal knowledge scales. It has been speculated that this approach could be adapted to produce maximum performance measures for nontraditional test domains such as personality or temperament (Legree, 1994).

One problem with developing an interpersonal knowledge scale is the lack of a credible knowledge base that can be referenced to verify the correctness of actions given some social scenario. This is largely due to the fact that the outcome of social situations can be multi-determined, and objectively verifying the relative appropriateness of an action may not be possible. One solution to this problem is to collect data from a group of SME to estimate the appropriateness of responses given some problem. This procedure seems to have provided reasonable scoring keys for the SJT and other situational judgment scales.

Regardless of the success of the SME-based procedure, it is not always possible to identify a group of experts for specific interpersonal situations. For example, identifying a group of individuals who are experts at acting in an appropriate manner at a cocktail party may be arbitrary unless some sort of cocktail party ability scale has already been developed. But a cocktail scale cannot be developed without first identifying a group of cocktail party experts using a traditional test development procedure.

As a means of circumventing this problem, this research explored the possibility that the opinions of a large number of knowledgeable non-experts can collectively reflect more expertise than the opinions of a few more knowledgeable experts for some knowledge/problem domains. It is assumed that the opinions of most individuals will reflect at least a substantial, although not necessarily an expert, level of knowledge for the selected (interpersonal) knowledge domains.

One justification for this position is the very high correlation between the mean Army NCO ratings and the mean Air Force recruit ratings ($r_{Army,AF}$=.72; r_{tt}=.95). These parameters suggest that the recruit data could be used to develop a reference pattern with which to score an NCO supervisory scale. Incidently, the Phase 1 analyses were repeated using the mean recruit ratings as a reference pattern, instead of the NCO ratings, with essentially identical results.

The use of knowledgeable non-experts to develop a reference pattern seems reasonable if the individual rating patterns are conceptualized as reflecting common and unique variance; the common variance corresponding to the general societal consensus and the unique variance being specific to the individual. If we assume that the general societal consensus represents expertise and that the unique variance is random across subjects, then expertise can be conceptualized as proportional to the magnitude of the common component. A reference pattern can be developed by averaging over the ratings of the non-expert subjects[4]. This procedure was used to develop the reference pattern mentioned in the preceding paragraph.

It follows that a more reliable reference pattern can be produced by surveying a large number of fairly knowledgeable nonexperts rather than a smaller number of more knowledgeable individuals. For Phase 2, all the Likert reference patterns were based on the mean Air Force recruit ratings. This procedure is an extension of the method, which was developed by Wagner and Sternberg (1986) to measure tacit knowledge in domains for which experts may be more easily and unambiguously identified.

One important twist, however, is that the distance scores would be distorted if an individual did not utilize the entire Likert scale, e.g., the ratings might all be compressed towards one end of the scale. This problem was addressed by converting all the ratings to z-scores and calculating item distances based

---

[4] For some knowledge domains, subjects may cluster into meaningful groups, in which case the scoring procedure would have to be group referenced. For example, males and females may disagree as to the meaning, implications, or appropriateness of specific phrases and actions that are sexually loaded.

on the differences between the z-scores corresponding to each individual and the z-scores based on the mean reference pattern.

Implications Related to g

Associated with the development of additional interpersonal knowledge scales is the requirement to examine the factor structure of these scales in relation to measures of general cognitive ability. It cannot be denied that the lack of a convincing factor analysis supporting the existence of a Social Intelligence construct casts doubt on the logic of trying to develop a battery of social intelligence tests. A major advancement in the study of social intelligence could occur if it can be shown that a low fidelity simulation format may be used to develop interpersonal knowledge scales that will load on a factor that is separate from those that can be obtained by factoring the ASVAB. It is notable that the use of factor analysis techniques is at the heart of the criteria proposed by Jensen (1993) to demonstrate the existence of separate factors.

To address this question, two additional interpersonal knowledge scales were developed that measure knowledge relating to commonly occurring interpersonal experiences. Thus this design represents a weak test of the hypothesis that a separate interpersonal knowledge factor exists because at least three scales are required to define a factor.

Selection of Content Areas

Two dissimilar content areas were selected for the development of the two additional interpersonal knowledge scales to insure the generality of the factor analyses in accordance with the criteria set forth by Jensen (1993). The use of dissimilar content areas is important because spurious factors can be produced by factoring content areas that are too similar.

The first content area can be described as explicit and assesses knowledge of dinner behavior by requiring the recruits to rate the extent to which a variety of actions would be appropriate at a traditional family dinner. The description of this content area as explicit is based on the expectation that individuals are given very candid feedback from parents and family members concerning appropriate dinner actions and style. This scale was intended to probe for knowledge that would be gained through familial interactions.

The second content area can be described as tacit and assesses knowledge of subtle indicators of alcohol abuse. This scale required individuals to rate the extent to which a variety of observable behaviors suggest alcohol abuse. This knowledge is viewed as being primarily covert and is assumed to be typically implicitly learned. It is notable that lists of subtle alcohol

abuse indicators are not readily available, subtle indicators cannot be easily identified, and the indicators generally reflect weak relationships. It is difficult to imagine how this type of knowledge would be explicitly learned and who would teach it.

Description of Scale Packages

Two scale packages were developed and correspond to the Likert and the forced choice conditions. The administration of the two conditions was similar in that subjects were given a test book and were required to record their responses on a separate scannable sheet. All subjects had completed the Armed Services Vocational Aptitude Battery (ASVAB) prior to enlistment. The Defense Manpower Data Center was contacted in order to provide ASVAB test data and to verify the accuracy of the Air Force data base.

The Likert package was produced by appending the two additional Likert scales, i.e., the Alcohol Abuse Indicators and Dinner Scenario scales, to the SJT Likert scale. Both of the additional scales are similar in that the Likert response format is utilized, however, the scales differ in that the Dinner scale incorporates a scenario, while the alcohol scale lists indicators of alcohol abuse without tying them to a specific scenario. Each scale contains 20 items; the length was determined on the basis of data (Wagner & Sternberg, 1986) indicating a reasonable level of reliability for Likert based knowledge scales of this length.

The forced choice condition corresponds to the package used in Phase 1 and is similar to the scale developed for Project A. Unfortunately the Dinner and Alcohol scales could not be meaningfully represented with a forced choice format and it was decided not to append the two Likert scales to the SJT forced choice scale because this would substantially complicate the administrative aspects of the task. Thus the package used for the forced choice condition only contained the SJT forced choice scale.

Subjects

Subjects consisted of 400 Air Force recruits in their 21st day of basic training. The subjects were divided into two groups of 200 subjects for the Likert and forced choice conditions.

Construction of Reference Patterns

The Likert scales were scored in accordance with the procedure described in Phase 1 with the exception that the Phase 2 reference patterns were based on the Phase 2 mean subject ratings. In contrast, the Phase 1 analyses had referenced the mean Army Noncommissioned Officer (NCO) ratings. As in Phase 1, subject ratings were transformed to insure equal means and

variances across subjects and within scales. The reference and subject ratings were standardized within each scale.

The Phase 2 Air Force rating means were used to score performance because these values were expected to better quantify the current level of appropriateness of the various response alternatives across the SJT problem scenarios. It seems reasonable to expect that the relative appropriateness of the response alternatives should slowly evolve to reflect current societal norms and from this perspective the Project A scoring key could be argued to be dated. In addition, the scale was originally based on NCO supervisory knowledge and it seems more reasonable to score Air Force recruits with a reference pattern reflecting the knowledge of Air Force recruits as opposed to knowledge that might be more closely tied to Army NCO experiences. Finally, the reliability of the mean scores would be extremely high because these means are based on the responses of 200 individuals. Adopting these values as a reference pattern is based on the assumptions that: many subjects will have some level of expertise that will be reflected in their ratings; and unique variance will be randomly distributed about the mean rating for each response alternative.

The forced choice condition was scored a second time using the Project A scoring pattern for the specific purpose of comparing the mean performance level of the Air Force recruits and the Project A soldiers.

Procedure

Data were collected after breakfast over a two month period between 7:00 and 9:00 AM. Subjects were alternately assigned to condition. Subjects were seated in a classroom and were instructed to follow the instructions described in the testbook. The subjects were tested in groups of 20 and were told to wait quietly at their desks until the session was completed.

Results

Mean Performance Level. Mean performance estimates for the Air Force samples are reported in Table 3 with estimates obtained from Project A that quantify mean performance on the 49 scenario SJT. All the values reported in Table 3 are based on the Project-A (Army NCO) scoring key. Table 3 indicates that the mean score of the Army sample exceeds that of the Air Force sample for four of the five scales. Given the large sample sizes used in this study (n=198) and in Project A (n=1,580), all four of these comparisons yield significant z-ratios.

Effect sizes are also reported in Table 3 and indicate that military experience had: (1) a moderate impact on the ability to identify the most appropriate solutions (.56 to .58 standard

deviation units), and (2) little impact on the ability to identify the least appropriate solutions (0 to .10 standard deviation units). Overall, military experience had only a minor impact on performance on the SJT as indicated by the effect size associated with the Difference Weighting score (.32 standard deviation units). With regards to interpreting the effect size estimates, it should be emphasized that Project A soldiers had approximately 3 years of military experience when tested, while the Air Force recruits were in their 21st day of basic training.

Table 3. Comparison of Army and Air Force Levels of Performance Based on the Project A Reference values.

Scoring Procedure	Army Mean (SD)	Air Force Mean (SD)	Effect Size[1]	Z-score Ratio	p<
Most Proportion	.53 (.12)	.44 (.10)	.58	11.67	.0001
Least Proportion	.45 (.10)	.45 (.10)	0	0	ns
Most Weighting	4.97 (.32)	4.76 (.28)	.56	9.78	.0001
Least Weighting[2]	3.35 (.29)	3.42 (.26)	.10	3.52	.001
Difference Weight	1.62 (.57)	1.35 (.50)	.32	7.05	.0001

[1] Calculated in accordance with Bloom (1984) with the Army SD as the reference value.
[2] Low scores indicate better performance. the reference value.

Reliability Data. Reliability estimates for the experimental scales are reported in Table 4 and generally replicate the Phase 1 estimates. Specifically, the SJT Likert scores were substantially more reliable than the Proportion Correct scores and were also more reliable than the weightings-based measures.

The reliability estimates for the two additional Likert scales can be considered moderate for the Dinner Scenario scale, although additional work could substantially improve its reliability, and high for the Alcohol Abuse scale considering that both scales: (1) address broad knowledge domains, (2) contain 20 items each, and (3) require approximately one-thirtieth and one-twentieth the administration time of the SJT scale based on the length of the SJT text. The number of words

in each scale is contained Table 4 and is presented as a proxy for the time required to administer the scales. It was not practical to collect administration time estimates.

Table 4. Internal Consistency Reliability Estimates.

Scale / Scoring Procedure	Words	Mean	Var	Reliability
SJT Forced Choice Format Measures				
Most Proportion Correct	8312	.47	.09	.44
Least Proportion Correct	8312	.48	.10	.59
Most Weighting[1]	8312	7.86	.20	.66
Least Weighting[1]	8312	6.50	.24	.69
Difference Weighting[1]	8312	1.36	.40	.76
Likert Format Scales				
Situational Judgment Test	8445			.80
Dinner Scenario	319			.50
Alcohol Abuse	421			.75

[1] These values reflect the 11-point scale.

The correlation between the mean ratings of the items across the two data collection efforts was very high ($r=.94$; $p<.001$) as was the correlation between the Army NCO ratings and the mean Air Force Phase 2 ratings ($r=.74$; $p<.001$).

Confirmatory Factor Analysis (CFA) Overview. Confirmatory factor analyses (CFA) were conducted both before and after correcting the correlation matrix with the multivariate correction for restriction of range (Lord & Novick, 1968). The multivariate correction was utilized to estimate the correlation matrix for the 13 variables had data been collected on a representative sample of the American youth population. It was expected that the two sets of analyses would be similar in most major respects with the exception that the loading of the first-order factors on the second-order factor, psychometric g, would be higher after correcting the sample matrix for restriction of

range.

The CFAs were structured to test the hypothesis that the experimental scales represent a factor that is not evidenced in the factor structure of the Armed Services Vocational Aptitude Battery. It was expected that all the first-order factors would be correlated and therefore have substantial loadings on psychometric g. The magnitude of these loadings were estimated by conducting a second-order factor analysis after determining the models that best describe the sample and corrected correlation matrices.

CFA Using the Corrected Correlation Matrix. Goodness of fit statistics are summarized in Table 5 for two 4-factor models and three 5-factor models based on analyses of the corrected correlation matrix. The factor structure of the two models that could not be rejected on the basis of the Chi-square test are summarized in Tables 6 and 7. As described below, these two models differ in the presence of a link between one of the ASVAB tests and the Social factor that was marginally significant, t=1.69.

The initial 4-factor and 5-factor models (Model 4I and Model 5I) were based on the factor structure reported by Kass, Mitchell, Grafton, and Wing (1983). Their analyses indicate four first order factors as follows: (1) Verbal composed of the General Science, Word Knowledge and Paragraph Comprehension tests; (2) Quantitative composed of the Arithmetic Reasoning and Math Knowledge tests; (3) Speed composed of the Numerical Operations and the Coding Speed tests; and (4) Technical composed of the Auto Shop, Mechanical Comprehension and Electronics tests. In the Kass model, only the General Science test loads on more than one factor.

With respect to the initial 4-factor model, Model 4I, the placement of the experimental scales on the four factors was based on modification indices provided by the LISREL program. For the initial 5-factor model, Model 5I, the experimental scales were hypothesized to load on a separate first-order factor termed Interpersonal Knowledge. The first-order factors were set free to correlate.

The structure of the Kass based 4-factor and 5-factor models were optimized on the basis of LISREL provided information, i.e., t-tests and modification indices, thereby producing Model 4-Optimized (Model 4O) and Model 5-Optimized (Model 5O). As implied by Table 5, it was not possible to produce a 4-factor model that could not be rejected on the basis of the Chi-square test. Model 5O could be rejected (p=.06, n.s.) and was therefore retained.

Table 5. Summary Statistics for the 13-variable CFAs.

Model	Root Mean Square Residual	Chi Square Statistics Value	DF	p	GFI	AGFI
		MODELS BASED ON KASS STRUCTURE				
4I Factor	.044	109.85	58	0.000047	0.92	0.87
5I Factor	.040	95.05	54	0.00048	0.93	0.88
		OPTIMIZED MODELS				
4O Factor	.035	83.22	56	0.011	0.94	0.90
5O Factor	.032	68.53	52	0.061 (ns)	0.95	0.91
5A Factor	.034	62.83	51	0.120 (ns)	0.95	0.91

The factor structure of Model 5O is contained in Table 6 and indicates that only the experimental scales have substantial positive loadings on the separate factor, identified as the Interpersonal Knowledge factor. The primary conceptual difference between the optimized 5-factor model and the model hypothesized on the basis of Kass is that the Auto Shop test, in addition to the General Science test, is allowed to load on two factors. The 5-factor models do not differ in the placement of the experimental scales.

To further explore and verify the structure of the loadings of the variables in Model 5O, all the ASVAB tests were iteratively placed on the Interpersonal Knowledge factor. For all tests except Numerical Operations, this strategy resulted in a small loading for the ASVAB test on the Interpersonal Knowledge factor and a decrease in the quality of the overall fit statistics.

The inclusion of the Numerical Operations test on the Interpersonal Knowledge factor may have improved the overall fit of the model as is suggested by the Chi-square p-statistic, which increased to .12. (Refer to Model 5A in Table 5.) However, the t-value for this link is marginal, t=1.69. In any event, the Numerical Operations test had a substantial negative loading on the Social factor, -.49. This loading suggests a negative correlation between performance on the Numerical Operations test

and Interpersonal Knowledge.

Table 6. Model 5O (Corrected Correlation Matrix): CFA loadings For the 10 ASVAB Tests and the 3 Interpersonal Knowledge Scales.

Scale	Factors				
	Verbal	Speed	Quant	Interper	Tech
ASVAB TESTS					
General Science	.49	---	---	---	.46
Arithmetic Reasoning	---	---	.75	---	.21
Word Knowledge	.94	---	---	---	---
Paragraph Comprehension	.85	---	---	---	---
Numerical Operations	---	.88	---	---	---
Coding Speed	---	.80	---	---	---
Auto Shop	---	---	-.40	---	1.14
Math Knowledge	---	---	.91	---	---
Mechanical Comprehension	---	---	---	---	.86
Electronics	---	---	---	---	.88
EXPERIMENTAL SCALES					
Situational Judgment Scale	---	---	---	.85	---
Alcohol Abuse Indicators	---	---	---	.45	---
Dinner Behavior	---	---	---	.63	---

The loadings associated with the Models 5O and 5A (Optimized Alternate) are presented in Tables 6 and 7. As hypothesized, the loadings of the ASVAB tests were distributed over the 4 traditional ASVAB factors (Verbal, Quantitative, Technical and Speed), while the experimental variables loaded on the Interpersonal Knowledge factor.

Table 7. Model 5-A (Corrected Correlation Matrix): CFA loadings For the 10 ASVAB Tests and the 3 Interpersonal Knowledge Scales.

Scale	Factors				
	Verbal	Speed	Quant	Interper	Tech
ASVAB TESTS					
General Science	.49	---	---	---	.46
Arithmetic Reasoning	---	---	.76	---	.21
Word Knowledge	.95	---	---	---	---
Paragraph Comprehension	.85	---	---	---	---
Numerical Operations	---	1.32	---	-.49	---
Coding Speed	---	.75	---	---	---
Auto Shop	---	---	-.40	---	1.15
Math Knowledge	---	---	.91	---	---
Mechanical Comprehension	---	---	---	---	.86
Electronics	---	---	---	---	.88
EXPERIMENTAL SCALES					
Situational Judgment Scale	---	---	---	.86	---
Alcohol Abuse Indicators	---	---	---	.45	---
Dinner Behavior	---	---	---	.64	---

The higher order structure of Models 5A and 5O is summarized in Table 8. Across both models, the Verbal factor has the highest second-order loading followed by the other first-order factors. Note that the pattern of the loadings is very similar across the two retained models.

Table 8. Corrected Matrix: Second Order Loadings on g.

Factors	First-Order Factors					Second-Order
	Verb	Speed	Quant	IK	Tch	Loadings
MODEL 5O (5-FACTOR OPTIMIZED)						
Verbal	1.00	.74	.80	.85	.76	.94
Speed		1.00	.76	.68	.52	.80
Quantitative			1.00	.73	.74	.85
Interpersonal Knowledge				1.00	.73	.88
Technical					1.00	.81
MODEL 5A (5-FACTOR OPTIMIZED ALTERNATE)						
Verbal	1.00	.82	.79	.84	.76	.94
Speed		1.00	.74	.78	.71	.88
Quant			1.00	.76	.69	.85
Interpersonal Knowledge				1.00	.72	.89
Tech					1.00	.81

In general, the CFA of the corrected matrix verifies the hypothesized model to the extent that a separate factor, the Interpersonal Knowledge factor, is primarily composed of the three experimental scales. However, the factor structure of the ASVAB appears to be slightly more complex than hypothesized on the basis of Kass, Mitchell, Grafton and Wing (1983).

CFA Using the Sample Correlation Matrix. The analyses conducted on the sample correlation matrix are similar to those conducted on the corrected correlation matrix. Initially, a 4-factor and a 5-factor model were hypothesized on the basis of the factor structure reported by Kass, Mitchell, Grafton, and Wing (1983). As described above, the placement of the experimental scales on the factors were based on the modification indices for the 4-factor model and were placed on a separate first-order factor for the 5-factor model. Two additional models, Model 4OS (4-factor optimized sample) and Model 5OS (5-factor optimized

sample), were produced by modifying the Kass-based models on the basis of the LISREL provided modification information.

Goodness of fit statistics are summarized in Table 9 for the four models, i.e., the two 4-factor models and the two 5-factor models, for the sample correlation matrix. Although none of the models could be rejected, the Chi-square statistic for the Model 5OS approaches the .05 level (p=.026) and appears to be the best model on the basis of the Goodness of Fit statistics.

Table 9. Sample Matrix: Summary Statistics for the 13-variable CFAs.

Model	Root Mean Square Residual	Chi Square Value	DF	p	GFI	AGFI
MODELS BASED ON KASS STRUCTURE						
4 Factor	.064	118.33	58	0.0000051	0.91	0.86
5 Factor	.051	90.94	54	0.0012	0.93	0.89
OPTIMIZED MODELS						
4OS Factor	.059	101.98	57	0.00023	0.92	0.88
5OS Factor	.046	74.82	53	0.026	0.94	0.91

The factor structure of Model 5OS is presented in Table 10. The CFA distributed the loadings of the ASVAB tests over the four hypothesized ASVAB factors while the experimental variables loaded on the Interpersonal Knowledge factor.

The primary conceptual difference between the optimized 5-factor model and the model hypothesized on the basis of Kass is that one additional ASVAB test, the Mechanical Comprehension test, loaded on two factors. Thus the hypothesized model was confirmed except that the placement of the ASVAB tests on the first-order factors was slightly more complex than expected.

Table 10. Model 5OS (Sample Correlation Matrix): CFA loadings For the 10 ASVAB Tests and the 3 Interpersonal Knowledge Scales.

Scale	Factors				
	Verbal	Speed	Quant	Interpers	Tech
ASVAB TESTS					
General Science	.58	---	---	---	.26
Arithmetic Reasoning	---	---	.95	---	---
Word Knowledge	.89	---	---	---	---
Paragraph Comprehension	.61	---	---	---	---
Numerical Operations	---	.81	---	---	---
Coding Speed	---	.78	---	---	---
Auto Shop	---	---	---	---	.66
Math Knowledge	---	---	.71	---	---
Mechanical Comprehension	---	---	.30	---	.60
Electronics	---	---	---	---	.69
EXPERIMENTAL SCALES					
Situational Judgment Scale	---	---	---	.71	---
Alcohol Abuse Indicators	---	---	---	.36	---
Dinner Behavior	---	---	---	.47	---

The second order loadings of the factors on psychometric g are contained in Table 11. As can be seen, the Verbal factor has the highest loading on psychometric g followed by the other first-order factors.

Table 11. Sample Matrix: Second Order Loadings on g.

Factors	First-Order Factors					Second-Order
	Verb	Speed	Quant	IK	Tch	Loadings
MODEL 5O (5-FACTOR OPTIMIZED)						
Verbal	1.00	.11	.48	.57	.55	.87
Speed		1.00	.36	.25	-.23	.38
Quantitative			1.00	.47	.32	.46
Interpersonal Knowledge				1.00	.40	.62
Technical					1.00	.60

Integration of the Factor Analyses. The two models retained by the CFA of the corrected and sample correlation matrices, i.e. Model 5O and Model 5OS, are extremely consistent. The two models differ only in the presence of one link between the Technical factor and the Arithmetic Reasoning test. Furthermore, the presence or absence of this link has no direct impact on the principal question of interest, which is the hypothesized existence of an Interpersonal Knowledge factor.

To address the possibility that the results from the CFA might reflect an inadequate model, an exploratory factor analysis (EFA) was conducted and is reported in Appendix B. In general, the results of the EFA are consistent with the two CFAs in that the EFA demonstrated a separate first-order factor that corresponds to the Interpersonal Knowledge factor. The EFA, however, led to somewhat lower estimates of the loadings of the first-order factors on psychometric g. This finding may reflect the imperfect fit of the models retained on the basis of the CFA. In other words, if most of the ASVAB tests and exploratory scales have low loadings on a number of factors, then the factors extracted by the EFA will be slightly less correlated than those extracted by the CFA. The lower factor correlations would then result in lower second-order loadings. On the other hand, the low loadings may also capitalize on chance relationships, a possibility that would lend more credence to the results from the CFA.

Discussion

In general, the data collected during Phase 2 replicated Phase 1 conclusions. Specifically, the SJT could be used to collect individual difference data given a population with little military experience and the Likert response format could be used to improve the reliability of an existing situational judgment scale.

The confirmatory and exploratory factor analyses are consistent with the conclusion that the experimental scales load on a separate factor, identified as Interpersonal Knowledge, that has a second-order loading on g that is equivalent to that of the other factors. In other words, these data lend as much credence to the existence of a separate Interpersonal Knowledge factor as they do for the other first-order factors. This suggests that additional research is needed to refine the measurement and the understanding of this factor and the corresponding scales.

One limitation with the present design is that content is confounded with method across the three experimental scales analyzed in the factor analyses. It is possible that the experimental factor corresponds to the ability to process and understand the probabilistic relationships referenced by the experimental scales than to Interpersonal Knowledge per se. In any event, this issue cannot be further explored without additional data collection. Incidently, the demonstration of a "probabilistic" factor would be important because the demonstration would suggest a general ability to manage (learn, understand and process) the probabilistic information and uncertain relationships that are intrinsic to many tasks, including those involving interpersonal interactions.

General Discussion

Methodological Implications for Existing Scales

One concern with using a scale developed to measure NCO supervisory skills to collect data with Air Force recruits is the possibility that the scale may not be sensitive to individual differences given a population with little military experience. This could happen if the performance of many of the Air Force recruits approached a random level, i.e., a floor effect occurred. However, the data indicate only a minor mean difference in performance between the Air Force recruits and the Project A soldiers. Furthermore, the reliability parameters for the forced choice version of the Situational Judgment Test were very similar to those estimated with the Project A population. These findings indicate that the psychometric properties of the Situational Judgment Test were not greatly effected by utilizing

a recruit population.

With respect to the question of response format, the reliability estimates indicate a more reliable scale can be obtained by substituting the Likert format for the forced choice format. This was expected because much more data are collected with the Likert format; increasing the amount of data should, according to the Spearman-Brown prophecy formula, result in a more reliable measure. This demonstration would be trivial if the increase in reliability was simply due to increasing the length of the scale in a manner that increases administration time, but the additional data were collected with only a minor impact on administration time. Because time limitations often restrict the amount of data that can be collected to support psychological research and personnel selection, any procedure that improves the reliability of a scale without increasing its administration time is important.

The reason that this format has not been utilized in the past is not clear; the procedure is not particularly complex, although it can be surprisingly difficult to explain and justify. One limitation to the Likert scoring procedure is that z-score transformations and distance calculations require a substantial amount of computing time per subject. Although this procedure could be performed manually, it is unlikely that research with the Likert format would be conducted by individuals who are not familiar with a programming language. In this research, all the z-score transformations were computed by Pascal programs that were developed for this purpose.

In any event, the improvement in reliability is notable because this issue has not been empirically explored or reported. This method could be used to improve the psychometric properties of a number of existing scales that may address interpersonal domains, i.e., telephone sales representative skills (Phillips, 1992), collection agency negotiation skills (Phillips, 1993), administrative and interpersonal skills of educators (Ostroff, 1991) and managerial skills (Motowidlo, Dunnette & Carter, 1990). These scales have never been assembled and administered to a common group of subjects with the purpose of describing the underlying factor structure that such a battery would possess. Future research could easily address this question.

Measurement of Non-Traditional Domains

One pleasantly surprising outcome of this research was the ease with which the two additional Interpersonal Knowledge scales were developed. Although identifying appropriate content domains was comparatively difficult, developing the items was remarkably straightforward. Most of the items were generated though informal conversations and the reference patterns used to score the scales were based on subject data thereby requiring

only a single round of data collection[5].

The most important implication of this work is the demonstration that this method may be used to measure individual differences in knowledge domains that traditional testing formats are unable to effectively address. For example, the Likert method could potentially be used to develop tests oriented towards knowledge domains that are associated with specific personality traits. Thus individual differences in emotional stability might correlate with knowledge of either the relative effectiveness of strategies that lessen feelings of emotional distress or knowledge of the extent to which specific situations are likely to result in emotional distress. Emotionally stable individuals would be expected to perform better on such a scale. In a similar vein, individuals who are high in assertiveness or dominance would be expected to know more about being assertive or dominant.

An important advantage to a personality scale based on the Likert format over existing personality inventories is that faking is not an issue with such a scale. This is because the Likert format explicitly requires subjects to estimate the objective appropriateness of various actions, i.e., the most correct responses are also the most socially desirable. Existing personality inventories usually require subjects to describe their personalities through agreement with various statements and many of these statements vary in social desirability. As a result, existing personality inventories are often highly fakeable, e.g., instructing subjects to "fake good" on a personality inventory resulted in a 1.7 standard deviation unit increase on the Assessment of Background Life Experiences (ABLE) composites (Young, White & Oppler, 1991; Young, White & Oppler, 1992). The important point is that faking would not be an issue with a knowledge scale developed to measure a personality trait because it would be a maximum performance measure. This type of scale could be directly used to predict performance and support personnel selection and classification.

In this research, all the scales were scored with a common reference pattern. However, it is important to realize that Likert-based scales can be scored with multiple reference patterns corresponding to different groups. One method to explore group differences would then be based on the use and analysis of the different scoring patterns. (For example, males and females may disagree as to the implications of specific actions for some content areas.) Analysis of these patterns, i.e., both the individual difference estimates and the scoring reference patterns, would allow insight into the social

---

[5] With respect to scale development, Dr. Beatrice Farr deserves special thanks for her suggestions and insights.

interactions and altercations that can develop between members of differing and sometimes competing groups (c.f., Kochman, 1983).

Interpersonal Knowledge and Psychometric g

The various factor analyses are consistent in that they indicate the existence of a separate factor, labelled Interpersonal Knowledge, that has a substantial correlation with the other first-order factors. The most important finding is that the experimental scales could not be subsumed under any of the traditional ASVAB (cognitive) factors. The second-order factor analysis is important because it indicates that the loading of the Interpersonal Knowledge factor on psychometric g is comparable to the loadings of the other first-order factors. In summary, the factor analyses indicate the existence of a separate factor with a second-order loading on psychometric g that is roughly equivalent to that of the other first-order factors.

The demonstration of a separate Interpersonal Knowledge factor has a direct bearing on the debate over the hypothesized existence of practical and social intelligence as separate from general intelligence (Sternberg & Wagner, 1993; Jensen, 1993). These data support the assertion that Social Intelligence is distinct from other cognitive ability factors, a finding consistent with Sternberg and Wagner's assertions, while also indicating a substantial loading on psychometric g as argued by Jensen (1993). In other words, these positions are not mutually exclusive and both may be correct. This interpretation assumes that the factor corresponds to the test content as opposed to the test format.

That the second-order factor analysis indicates that interpersonal knowledge has a high loading on psychometric g does not trivialize the importance of this factor. All the first-order factors have substantial second-order loadings and the factor correlations indicate that the Interpersonal Knowledge construct correlates at about the same level with Verbal and Quantitative Intelligence as Verbal and Quantitative Intelligence correlate with each other. It is important to note that classification efficiency (i.e., allocation efficiency) is dependent on the existence of separate factors (Johnson & Zeidner, 1991). The practical value of such a factor would be realized by developing a battery of interpersonal knowledge scales that predict tasks requiring interpersonal skills.

Intuitively, this finding seems reasonable in that while "high verbal-low social ability" individuals may be located, they seem about as uncommon as "high verbal - low quantitative ability" individuals. A substantial correlation between interpersonal knowledge and verbal or quantitative intelligence is entirely reasonable. In fact, a low loading on psychometric g

would imply a near-zero correlation between social intelligence and general intelligence and would be unbelievable in that very low IQ scoring individuals (e.g., the mentally retarded) would be predicted to perform almost as well on Interpersonal Knowledge scales as high IQ individuals.

The approach adopted in this research is entirely consistent with the general practice of measuring intelligence by estimating knowledge for a very circumscribed domain. For example, vocabulary scales typically have a very high loading on verbal intelligence scales despite the fact that verbal intelligence is generally conceptualized to be broader in scope than simply being able to define terms. In an analogous manner, although the experimental scales may be measuring only a minor aspect of the realm of social functioning and social intelligence, this type of scale may have substantial potential to estimate individual differences in social intelligence.

Future Research

Additional data collection is not currently planned. However, substantial research could easily be undertaken and would likely yield theoretically important results. One logical extension is the exploration of the factor structure of the domains measured by these interpersonal scales and other situational judgment tests, i.e., telephone sales representative skills (Phillips, 1992), collection agency negotiation skills (Phillips, 1993), administrative and interpersonal skills of educators (Ostroff, 1991), and managerial skills (Motowidlo, Dunnette & Carter, 1990). These scales have never been administered to a common group of subjects with the purpose of exploring the underlying factor space. The resultant factor analysis would yield interesting regardless of its consistency with those described in this paper. From a practical perspective, the factor analysis might identify scales to be included in a battery of Social Intelligence tests.

Conceptually, the Likert-based method has substantial potential to measure individual differences in non-traditional content domains, such as personality. Emerging personality theory (Mayer, in press; Mayer, DiPaola & Salovey, 1990) relates personality traits to experience, knowledge and abilities. For example, empathy and extroversion are hypothesized to be related to emotional intelligence, which is loosely defined as the ability to infer the emotions of individuals; presumably this ability is gained through experience. Consistent with this notion is the expectation that specific personalty types will be more familiar with the associated knowledge if only because specific experiences are associated with the specific personality profiles. If this expectation is correct, then it follows that

personality could be measured with corresponding knowledge scales. Resulting scales would have the tremendous advantage of eliminating faking as a issue and the methodological implications would be of broad practical and theoretical importance.

References

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. Educational Researcher, 13, 4-16.

Campbell, J. P., & Zook, L. M. (1991). Improving the selection, classification and utilization of Army enlisted personnel: Final report on Project A (Research Report 1597). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A242 921)

Carroll, J. B. (1993). Human cognitive abilities: A survey of factor analytic studies. New York: Cambridge University Press.

Hallahan, D. P., & Kauffman, K. U. (1991). Exceptional children. Needham Heights, MA: Allan.

Hanson, M. A., & Borman, W. C. (1992). Development and construct validation of the situational judgment test (SJT) (PDRI Report #230). Minneapolis, MN: Personnel Decisions Research Institute.

Jensen, A. R. (1993). Test validity: g versus "tacit knowledge." Current Directions in Psychological Science, 2, 9-10.

Johnson, C. D., & Zeidner, J. (1991). The economic benefits of predicting job performance. New York: Praeger.

Kass, R. A., Mitchell, K. J., Grafton, F. C., & Wing, H. (1983). Factorial validity of the Armed Services Vocational Aptitude Battery (ASVAB), Forms 8, 9 and 10: 1981 Army applicant sample. Education and Psychological Measurement, 43, 1077-1087.

Keating, D. P. (1978). A search for social intelligence. Journal of Educational Psychology, 70, 218-223.

Kochman, T. (1983). Black and white: Styles in conflict. Chicago, IL: University of Chicago Press.

Legree, P. J. (1994). The effect of response format on reliability estimates for tacit knowledge scales (Research Note 94-25). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences. (AD A283 547)

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley Publishing.

Marlowe, H. A. (1986). Social intelligence: Evidence for multidimensionality and construct independence. Journal of Educational Psychology, 78, 52-58.

Mayer, J. D. (in press). The system-topics framework and the structural arrangement of systems within and around personality. Journal of Personality.

Mayer, J. D., DiPaolo, M., & Salovey, P. (1990). Perceiving affective stimuli in ambiguous visual stimuli: A component of emotional intelligence. Journal of Personality Assessment, 54, 772-781.

McNemar, Q. (1969) Psychological statistics. New York: Wiley.

Motowidlo, S. J., Dunnette, M. D., & Carter, G. W. (1990). An alternative selection procedure: The low-fidelity simulation. Journal of Applied Psychology, 75, 640-647.

Ostroff, C. (1991). Training effectiveness measures and scoring schemes: A comparison. Personnel Psychology, 44, 353-374.

Phillips, J. F. (1992). Predicting sales skills. Journal of Business and Psychology, 7, 151-160.

Phillips, J. F. (1993). Predicting negotiation skills. Journal of Business and Psychology, 7, 403-411.

Sternberg, R. J., Conway, B. E., Ketron, J. L., & Bernstein, M. (1981). People's conceptions of intelligence. Journal of Personality and Social Psychology, 41, 37-55.

Sternberg, R. J., & Wagner, R. K. (1993). The g-ocentric view of intelligence and job performance is wrong. Current Directions in Psychological Science, 2, 1-4.

Tisak, M. S., & Ford, M. E. (1983). A further search for social intelligence. Journal of Educational Psychology, 75, 196-206.

Wagner, R. K., & Sternberg, R. J. (1986). Tacit knowledge and intelligence in the everyday world. In R. Sternberg & R. Wagner (Eds.), Practical intelligence: Nature and origins of competence in the everyday world (pp. 51-83). New York: Cambridge University Press.

Walker, R. E., & Foley, J. M. (1973). Social intelligence: Its history and measurement. Psychological Reports, 33, 839-864.

Young, M. C., White, L. A., & Oppler, S. H. (1992, October). Effects of coaching on validity of a self-report temperament measure. Paper presented at the meeting of the Military Testing Association, San Diego, CA.

Young, M. C., White, L. A., & Oppler, S. H. (1991, October). Coaching effects on the assessment of background and life experiences (ABLE). Paper presented at the meeting of the Military Testing Association, San Antonio, TX.

Appendix A

Instructions for the Two Situational Judgment Test Conditions

Page A-2 contains the instructions for the Likert version of the SJT; an example of a scenario with ratings is contained in the instructions. Page A-3 contains the instructions and the example used for the forced choice version of the SJT. The example was modified for the Air Force subjects by replacing the Army term, Platoon, with the equivalent Air Force term, Flight.

SITUATIONAL JUDGMENT TEST

In this booklet, you will be presented with a series of supervisory situations. These are situations in which a first line supervisor might find him/herself. After each situation several possible responses to that situation are listed. To insure realistic scenarios, the situations and responses are based on the experiences and statements of senior NCOs.

Your task is to read each situation and the responses listed. Then rate the appropriateness of each of the actions on the 11 point scale. Be sure to rate all the actions.

Below is an example of an item that has been completed properly.

---

1 2 3 4 5 6 7 8 9 10 11

Extremely Inappropriate — **Neither Appropriate Nor Inappropriate** — **Extremely Appropriate**

You are a Work Center NCOIC. Over the past several months you have noticed that one of the other Work Center NCOICs in your Flight hasn't been conducting his Common Task Training (CTT) correctly. Although this hasn't seemed to affect the Flight yet, it looks like the Flight's marks for CTT will go down if he continues to conduct CTT training incorrectly. How appropriate are the following actions.

2 a. Do nothing since performance hasn't yet been affected.

7 b. Have the Work Center NCOIC meeting and tell the Work Center NCOIC who has been conducting training improperly that you have noticed some problems with the way he is training his troops.

8 c. Tell your Flight sergeant about the problem.

10 d. Privately pull the Work Center NCOIC aside, inform him of the problem, and offer to work with him if he doesn't know the proper CTT training procedure.

You may not agree with the ratings for this item, but this example shows you how these items should be completed.

Be sure to rate each item on the 11 point scale and be sure to use the entire scale.

SITUATIONAL JUDGMENT TEST

In this booklet, you will be presented with a series of supervisory situations. These are situations in which a first line supervisor might find him/herself. After each situation several possible responses to that situation are listed. To insure realistic scenarios, the situations and responses are based on the experiences and statements of senior NCOs.

Read each situation and the responses listed. Then decide which of these possible responses would be the most effective. Place an "M" in the box next to the most effective response.

Next decide which of these possible responses is the least effective. Place an "L" in the box next to the least effective response. The boxes in front of the remaining response alternatives should be left blank.

Below is an example of an item which has been completed properly.

You are a Work Center NCOIC. Over the past several months you have noticed that one of the other Work Center NCOICs in your Flight hasn't been conducting his Common Task Training (CTT) correctly. Although this hasn't seemed to affect the Flight yet, it looks like the Flight's marks for CTT will go down if he continues to conduct CTT training incorrectly. What should you do?

__**L**__ a. Do nothing since performance hasn't yet been affected.

___ b. Have the Work Center NCOIC meeting and tell the Work Center NCOIC who has been conducting training improperly that you have noticed some problems with the way he is training his troops.

___ c. Tell your Flight sergeant about the problem.

__**M**__ d. Privately pull the Work Center NCOIC aside, inform him of the problem, and offer to work with him if he doesn't know the proper CTT training procedure.

You may not agree with the placement of the "**M**" and the "**L**" for this item, but this example shows you how these items should be completed.

In summary, for each item you will place an "**M**" for Most effective next to one response alternative, and an "**L**" for Least effective next to another response alternative. The boxes in front of the rest of the response alternatives will be left blank. Please use only one "**M**" and only one "**L**" per item.

Appendix B

Exploratory Factor Analysis

In order to gain additional insight into the factor structure underlying the correlation matrix, an exploratory principal axis factor analysis with oblique factor rotation was conducted on the corrected correlation matrix. The pattern matrix produced for a five factor solution, which was specified on the basis of the scree plot test, is reported in Table B1. A four factor solution was explored, but only the 5 factor solution was interpretable and is reported.

Table B1. Pattern Matrix: Oblique Rotation of the Principal Axis Factors of the 10 ASVAB and 3 Interpersonal Knowledge Tests.

Scale	Factors				
	1	2	3	4	5
ASVAB Tests	(Pattern Matrix)				
General Science	.34	.01	.12	.40	.24
Arithmetic Reasoning	.05	.10	.05	.15	.69
Word Knowledge	.56	.11	.22	.18	.09
Paragraph Comprehension	.66	.19	.01	.08	.09
Numerical Operations	.04	.85	-.12	.06	.11
Coding	.04	.78	.15	-.07	-.08
Auto Shop	.04	.04	.00	.90	-.07
Math Knowledge	.02	.04	.03	-.06	.92
Mechanical Comprehension	-.02	.02	.06	.65	.29
Electronics	.20	-.02	.08	.65	.13
Interpersonal Knowledge Tests					
Situation Judgment Scale	.22	.08	.53	.09	.08
Alcohol Abuse Indicators	-.09	.05	.37	.23	.02
Dinner Behavior	.14	.07	.55	-.11	.10

Inspection of the pattern matrix in Table 5 indicates that the 5 factor solution resulted in the Interpersonal Knowledge scales loading on Factor 3. The Interpersonal Knowledge scales had only low loadings on the other factors and the other ASVAB scales had only low loadings on the 4th factor. Examination of the other 4 factors indicates that the ASVAB scales loaded in the typical manner: Factor 1 contains verbal tests, Factor 2 represents the Speed factor, Factors 4 corresponds to the Technical factor, and Factor 5 is the Quantitative factor.

The Factor Correlation Matrix is contained in Table B2 and indicates that substantial correlations exist between all the factors. In accordance with the method used by Carroll (1993), the factor matrix was included in a second order principal components analysis to estimate the second order loadings (on psychometric g) of the 5 first order factors. These estimates are contained in Table 6 and it can be seen that the g-loading of Factor 3, which was composed of the Interpersonal Knowledge scales, is of a magnitude that is consistent with the loadings of the other first-order factors.

Table B2. Factor Correlation Matrix.

Factors		(Factor Correlations)					Second-Order
		1	2	3	4	5	g-Loadings
1	Verbal	1.00					.78
2	Speed	.58	1.00				.80
3	Quantitative	.61	.66	1.00			.87
4	Interpersonal	.46	.51	.49	1.00		.75
5	Technical	.38	.30	.55	.48	1.00	.68

To summarize the exploratory factor analysis, the ASVAB tests and the Interpersonal Knowledge scales load on separate factor and the second order principal components analysis indicates moderately high g-loading for all the first-order factors.