Job Knowledge Test Design: A Cognitively-Oriented Approach

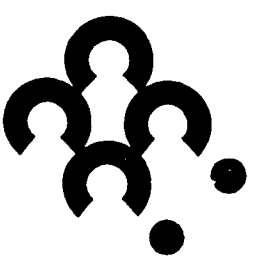
David DuBois
Personnel Decisions Research Institutes, Inc.
Minneapolis, Minnesota

Valerie L. Shalin
SUNY
Buffalo, New York

Keith R. Levi
MIU
Fairfield, Iowa

Walter C. Borman
Personnel Decisions Research Institutes, Inc., and
University of South Florida
Tampa, Florida

This research was sponsored by the Manpower, Personnel, and Training Program of the Office of the Chief of Naval Research (OCNR), Contract No. N00014-91-C-0224.



Personnel Decisions Research Institutes, Inc.
43 Main Street SE, Suite 405
Minneapolis, MN 55414
(612) 331-3680

Institute Report #241

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave Blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	July, 1993	Final, 3 Sept. 1991 to 31 May 1993

4. TITLE AND SUBTITLE
Job Knowledge Test Design:
A Cognitively-Oriented Approach

5. FUNDING NUMBERS
N00014-91-C-0224

6. AUTHOR(S)
David DuBois, Valerie Shalin, Keith Levi,
Walter Borman

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Personnel Decisions Research Institutes, Inc.
43 Main Street, SE
Riverplace, Suite 405
Minneapolis, MN 55414

8. PERFORMING ORGANIZATION REPORT NUMBER
Institute Report #241

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of Naval Research
Manpower, Personnel, and Training Programs
800 North Quincy
Arlington, VA 22217-5000

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

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

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

This report describes the application of cognitive methods to the measurement of performance using tests of job knowledge. The research goal is to improve the usefulness of job knowledge tests as a proxy for hands-on performance. The aPproach involves employing cognitive science methods to identify important knowledge content relevant to successful job performance, which may be missed by existing test development procedures. In an application to testing land navigation knowledge of U.S. Marines, the results suggest that cognitively-oriented job knowledge tests show improved correspondence with hands-on measures of performance, compared to existing content-oriented test development procedures. These results appear promising for the economical adaptation of cognitive methods to applications in performance measurement, training assessment, and training program evaluation.

14. SUBJECT TERMS
Job Knowledge, Performance, Personnel Testing,
Land Navigation

15. NUMBER OF PAGES

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	UL

Abstract

This report describes the application of cognitive methods to the measurement of performance using tests of job knowledge. The research goal is to improve the usefulness of job knowledge tests as a proxy for hands-on performance. The approach involves employing cognitive science methods to identify important knowledge content relevant to successful job performance, which may be missed by existing test development procedures. In an application to testing land navigation knowledge of U.S. Marines, the results suggest that cognitively-oriented job knowledge tests show improved correspondence with hands-on measures of performance, compared to existing content-oriented test development procedures. These results appear promising for the economical adaptation of cognitive methods to applications in performance measurement, training assessment, and training program evaluation.

Executive Summary

Introduction

The measurement of individual performance in the workplace is vital to numerous and important organizational objectives. It provides the basis for compensation and promotion decisions. Performance measures also serve as criteria for assessing the validity of personnel selection and placement programs, evaluating the effectiveness of training treatments, and determining the usefulness of job re-design efforts.

Job performance is traditionally assessed in one of three ways: by supervisory ratings, hands-on tests, or job knowledge tests. Hands-on tests are often considered to be the benchmark for measurement quality and user acceptance. Unfortunately, hands-on measures are very expensive and time consuming to develop and administer. As a proxy for performance, ratings can be affected by observer bias and provide only a modest correlation with hands-on measures. Given its economy and moderate correspondence with hands-on tests, job knowledge is a practical candidate to substitute for hands-on measures of performance. If the correspondence of job knowledge tests to other measures of performance can be sufficiently improved, the impact on organizational effectiveness could be substantial. Our research effort is aimed at the judicious integration of cognitive science methods to improve the relationship between tests of job knowledge and hands-on proficiency, while retaining the cost advantages of job knowledge tests.

In this report, we describe research on a job knowledge test for the task of land navigation by U.S. Marines. This task involves the use of a map, protractor and compass in planning and executing patrol routes. It is similar to the sport of orienteering. The land navigation task requires knowledge related to planning, perception, problem solving and the execution of actions in a physical task environment.

Strategy For Improvement

We propose that the moderate relationships between job knowledge tests and job performance are a result of limitations in existing test design methods. Specifically, we note four disadvantages of current test development procedures:

1) Job knowledge is inadequately defined by ratings obtained from job incumbents. Typically, task analyses are dominated by the rules and facts contained in the textbook literatures for that job. Developers of expert systems suggest that such knowledge is not operational for actually solving problems. Thus, the definition of the task and knowledge

domains may be deficient with respect to the content of knowledge required for successful performance.

2) Current methods for defining job knowledge provide little direct information on how this knowledge is organized and utilized for performance. The results of conventional task-oriented job analyses are the same whether experts or novices are utilized as job informants. Consequently, test specifications may also be deficient with respect to representing and assessing qualitative features of knowledge.

3) Existing methods for developing and selecting test content do not adequately identify the task and problem characteristics that most effectively discriminate among levels of performance. Existing methods design tests to sample representatively job knowledge content based on ratings of job importance and salience. However, these ratings used to guide the selection of test questions show little relationship to the diagnosticity of test questions suggest that specially designed problems are required for reliably identifying performance differences.

4) Job knowledge has traditionally been conceptualized as an accumulation of facts, principles, and procedures related to work performance. As a result, the psychological model underlying test scores is based on a single underlying trait of proficiency. This simple, unrealistic characterization of knowledge can provide only limited diagnostic information about knowledge and may be insufficient to capture important configural relationships between knowledge and performance. More recent work in cognitive and educational psychology provides a more complex and realistic model. This conceptualization is in terms of mental models; strategies for problem-solving, encoding, and retrieving information; and pattern recognition.

The approach that we propose rests upon the following premises about the potential contributions from cognitive science: a cognitively-oriented task analysis of actual performance will identify unique domain content with respect a) to the substantive domain and b) to the processes and types of information used; c) selection of test content based on its diagnosticity will improve test psychometrics; and e) the use of a more complex model of knowledge to score tests will improve the predictive efficiency and diagnostic usefulness of job knowledge tests. Specifically, we identified knowledge actually used during performance through the use of verbal protocol analyses and related methods. We employed a plan-goal graph representation to capture the knowledge content and goal structure of the studied task. We obtained diagnosticity ratings from task experts to identify the content categories and procedures that would best discriminate among levels of examinee performance and to specify the relative proportion of test questions to select. Finally, we used a probability-based inference network to score examinee responses and model a more complex pattern of relationships between knowledge and performance.

Results

The land navigation skills of 358 Marines were tested with the following performance measures: a 100 question knowledge test, 5 hands-on proficiency tests (addressing planning, location, distance, direction, and movement skills), and a work-sample performance test consisting of locating 4 stakes in a 5 square mile area, given the map coordinates. In this sample, job knowledge correlated .58 with the hands-on test and .42 with performance. By comparison, the average correlation between job knowledge and hands-on performance from previous studies was .38. For the subset of Marines in this sample who had recently been examined using existing tests, the correlation of job knowledge with work-sample performance was .08. The use of a probability-based inference network to score the job knowledge test provided similar results to conventional, total number correct scoring.

Conclusion

The evidence provided by this study supports the utility of cognitive methods for improving job knowledge tests. These preliminary findings suggest that cognitively-oriented test design increases the correspondence of job knowledge to hands-on and work-sample measures of performance. Of the several techniques adapted from cognitive science methods, those related to identifying unique domain content appear to have contributed most towards improving the predictive accuracy and interpretability of the job knowledge test. Furthermore, the incorporation of cognitively-oriented methods did not require substantially increased resources.

Other cognitive science methods, such as the use of a probability-based inference network to score the test presented potential advantages in diagnostic capabilities with no apparent loss to predictive efficiency. The capabilities to incorporate information from incorrect responses and to model configural patterns of knowledge-performance relationships offer the promise of substantial improvements in the integration of predictive and diagnostic uses of performance measurement. Additional research is needed to explore these interesting possibilities.

However, the generalizability of this research is limited. The study focused on a single task of the dozens performed within jobs in the Marine Corps. The results may not generalize to other areas of performance. On the other hand, estimates of the correspondence of job knowledge and performance may increase when cognitively-oriented methods are applied across all tasks of a job, due to increases in the stability of performance across tasks. Firm conclusions about the utility of cognitively-oriented approaches to job knowledge test design must wait for research which assesses the generality of the methods and results to entire jobs and to different types of performance domains.

Table of Contents

Introduction 1

Job Knowledge and Work Performance 2

Existing Methods For Job Knowledge Test Development 3
- Definition of the task domain 3
- Definition of the job knowledge domain 3
- Test content and format 4
- Test scoring and validation 4
- Limitations of current procedures 4

Contributions from Cognitive Science 5
- Definition of the task domain 5
- Definition of the knowledge domain 6
  - Methods for knowledge elicitation 6
- Knowledge representation: organization and structure 6
  - Computational cognitive architectures 7
  - Qualitative components of knowledge 8
- Test scoring. 9

Implementing Cognitively-Oriented Test Design 11
- Knowledge elicitation for land navigation 11
- Knowledge representation 14
  - Plan-goal graphs 14
  - Specifying test content 17
  - Developing test questions 18
- Test scoring. 21
- Summary of Cognitively-Oriented Test Design 26
- Methodological tradeoffs 27

Evaluation of Cognitively-Oriented Test Design 27
- Sample 28
- Measures 29
- Data collection 30
- Results 31
- Discussion 37

Conclusions 39

References 41

Tables

1 Specifications for a Cognitively-oriented Test of Land Navigation 18
2 Extract of a Knowledge Element Matrix 19
3 A Comparison of Methods for Eliciting and Representing Knowledge 26
4 Description of the Sample 28
5 Content Analysis of Land Navigation Tests 31
6 Correlations Between Job Knowledge, Experience, Proficiency, and Performance 32
7 Correlations Between Job Knowledge and Performance: A Comparison Between Cognitively-oriented and Existing Tests 33
8 Summary of Correlations Between Job Knowledge and Hands-on Proficiency . . . 34
9 Correlations Between Alternative Job Knowledge Scores and Performance 35
10 A Lens Model Analysis of Regression Components 37

Figures

1 A model of performance determinants 2
2 Portion of a plan-goal graph for U.S. Marine land navigation 15
3 Example test questions 20
4 A probability-based inference network for medical diagnosis. 22
5 A probability-based inference network of land navigation knowledge 25

1.0 Introduction

The measurement of individual performance in the workplace is vital to several important organizational and research objectives. It provides the basis for compensation and promotion decisions. Performance measures also serve as criteria for assessing the validity of personnel selection and placement programs, evaluating the effectiveness of training treatments, and determining the usefulness of job re-design efforts.

Job performance is traditionally assessed in one of three ways: by supervisory ratings, hands-on tests, or job knowledge tests. Each approach has advantages as well as certain disadvantages. For example, ratings can be affected by observer bias, hands-on testing is very expensive, and paper-and-pencil knowledge tests, while relatively inexpensive, correlate only moderately with actual performance.

Our research effort is aimed at the judicious integration of cognitive science methods to improve the relationship between tests of job knowledge and hands-on proficiency, while retaining the cost advantages of job knowledge tests. Hands-on proficiency testing is frequently cited as the benchmark for performance testing. The goal of our research is to sufficiently improve job knowledge measurement so that it can serve usefully as a substitute for hands-on measures of performance. The results of progress towards this goal will be substantially reduced costs for performance measurement and improved personnel decisions.

In this paper, we describe research on a job knowledge test for the task of land navigation by U.S. Marines. This task involves the use of a map, protractor and compass in planning and executing patrol routes. It is similar to the sport of orienteering. The land navigation task requires knowledge related to planning, perception, problem solving and the execution of actions in a physical task environment.

The major challenges of this effort involved adapting to the practical constraints of the project. The primary constraints involved requirements to use a written, multiple-choice format for testing and to maintain the cost advantages of job knowledge tests with respect to other performance measurement methods. That is, a cognitively-oriented approach must remain economical with respect to costs, time, and other resources (e.g., people). Guided by the practical nature of our goal and its corresponding constraints, numerous methodological adaptations were required to integrate cognitive with personnel psychology methods.

In order to describe our approach, we organized this report into the following framework. First, we depict the current conceptualization and measurement methods in job knowledge testing to provide a foundation for a contrast with methods from cognitive science. Next, we discuss three areas in which cognitive science can inform current practice in job knowledge testing and discuss how we implemented these methods in our project. Finally, we present some evidence which provides preliminary support for the usefulness of

this approach to job knowledge testing. For clarity and ease of reading, we will throughout the report refer to existing test development practices as "content-oriented". We will use the term "cognitively-oriented" to refer to the modifications of existing procedures which we adapted from cognitive science.

2.0 Job Knowledge and Work Performance

Job knowledge is central to most theories of work performance. While job knowledge is clearly distinct from performance, it is closely linked by guiding what, how, and when work performances are carried out. In reviews of the empirical literature, the mean correlations (corrected for criterion unreliability) of job knowledge were .41 with hands-on performance (Rumsey, Osborn, & Ford, 1985), .45 with supervisory ratings (Dye, Reck, & McDaniel, 1987) and .47 with training criteria (Dye, et al., 1987). Recent work in applied psychology to examine the causal determinants of performance appears to support this central role of job knowledge in performance (e.g., Borman, Hanson, Oppler, Pulakos, & White, in press; Borman, White, Pulakos, & Oppler, 1991; Hunter, 1983; McCloy, Campbell, & Cudeck, 1992; Schmidt, Hunter, & Outerbridge, 1986). For example, figure 1 displays the results of covariance structure analyses of empirical relationships among job knowledge and other determinants of performance. As a result, job knowledge tests appear to be the best candidate to proxy for the much more expensive hands-on measures of proficiency.

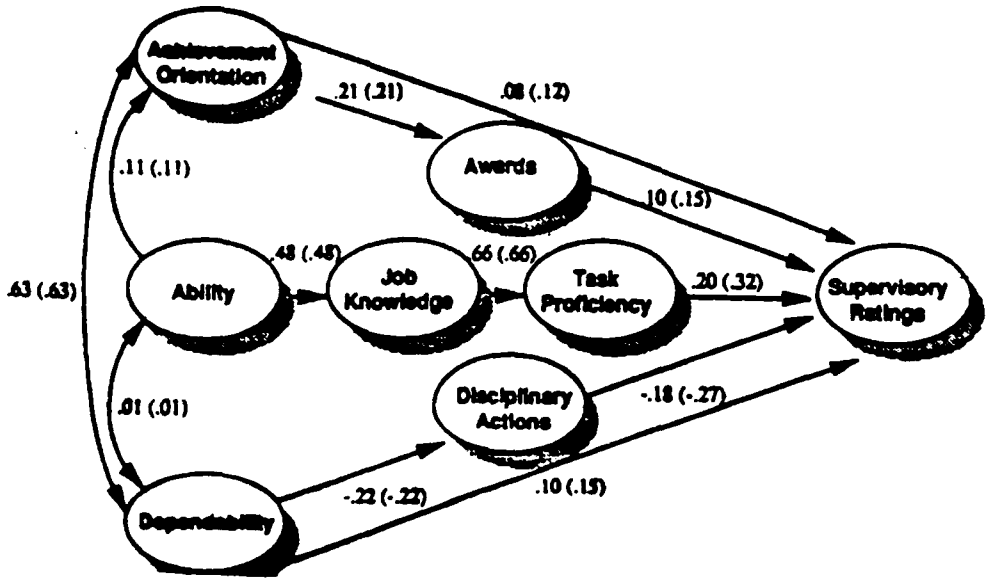


Figure 1. A MODEL OF PERFORMANCE DETERMINANTS (Borman et al., 1991)

However, while providing evidence for the centrality of job knowledge to performance, these data provide insufficient support for the use of job knowledge as a proxy for hands-on measures of performance. This project proposes that recent developments in cognitive science provide alternative conceptualizations and methods which potentially can improve the modest relationships currently reported in the literature. To provide a framework for evaluating the utility of cognitive methods for this purpose, we first describe existing methods for job knowledge test design.

3.0 Existing Methods For Job Knowledge Test Development

Job knowledge tests are frequently validated using a content validity approach, which substantiates the degree of overlap between the contents of the test and the job. A content validity strategy is employed because knowledge tests are considered to be work samples and because content validity is often used in settings where other validation strategies are difficult to employ (e.g., a criterion-related approach is not feasible due to small, available sample size). Operationally, content validity is established by having a panel of experts determine the relevance, representativeness, and fairness of test content with respect to successful job performance. The test construction methods used to develop content valid tests are known as *content-oriented* approaches to test design. As practiced by personnel psychologists, the development of job knowledge tests typically involves four major tasks: specification of the task domain; specification of the knowledge domain; development of the test content; and implementation of test scoring and validation.

3.1 Definition of the task domain. Test development begins with a systematic description of job tasks and behaviors. Methods for defining the tasks associated with a job typically consist of questionnaire-based ratings of the time spent and importance of tasks performed on the job. Within-job differences in tasks are defined by clustering these ratings into relatively independent dimensions of job performance. The relative importance of these dimensions provides a basis for representatively selecting and weighting test content.

3.2 Definition of the job knowledge domain. Definition of the knowledge associated with a job occurs in three steps. First, a comprehensive list of job knowledge elements and categories are generated through interviews with job experts, job observation, content analyses of documents and training materials, etc. As an example of the grain size of these elements, "arithmetic" might be identified as an element of job knowledge from the perspective of personnel psychology.

Next, these knowledge elements and categories are structured into a questionnaire for job incumbents, who rate each knowledge element for frequency and importance of use on the job. Knowledge elements that meet a minimum criticality threshold (combining frequency and importance) are then selected for inclusion on the test.

3.3 Test content and format. Content-oriented test development procedures establish test validity by demonstrating that the test representatively and fairly samples the job (Guion, 1978). Guion distinguishes between the test-content *universe* and the test-content *domain*. The test-content *universe* consists of all of the tasks, test conditions and measuring procedures that might be inferred from the job knowledge domain defined by the procedures discussed in the preceding paragraphs. The test-content *domain* consists of the actual specifications for test content, methods and format, depending upon the purpose and conditions for testing. Decision rules, influenced by the purpose of the test, identify elements of the test-content universe for the test-content domain. For example, test content may be selected by random sampling or representative sampling based on the relative importance of the knowledge categories.

3.4 Test scoring and validation. Inferences about test scores are based upon comparisons of the relevance and representativeness of test items to the knowledge requirements of the job. For example, Lammlein (1986) argued that content validity could be established by showing that 1) each item was relevant to at least one job performance requirement, 2) lack of knowledge would lead to some adverse consequence on the job, 3) each item was rated by job experts as being a necessary prerequisite to job performance and must be memorized, 4) each item had an appropriate level of difficulty for the job and 5) that the number of items in each knowledge category was proportionate to the salience of that knowledge category for the job.

3.5 Limitations of current procedures. We propose that the moderate relationships between job knowledge tests and job performance are a result of limitations in current practice related to the conceptualization and measurement of job knowledge. Of the several limitations in the above approach to the measurement of job knowledge, we focus on three. We discuss relevant work in cognitive science which address each of these limitations in the noted subsections:

1) The implied conceptualization of knowledge is based on an accumulation of facts, concepts, and skills which can be specified by ratings obtained from job incumbents. Typically, these task analyses are dominated by the rules and facts contained in the textbook literatures for that job. Yet, developers of expert systems suggest that such knowledge is not operational for actually solving problems (Breuker & Weilinga, 1987)— (in section 4.1). Thus, the definition of the task and knowledge domains may be deficient with respect to the content of knowledge required for successful performance.

2) Current methods for defining job knowledge provide little direct information on how this knowledge is organized and utilized for performance. Indeed, the results of conventional task-oriented job analyses are the same whether experts or novices are utilized as job informants (Conley & Sackett, 1987; Wexley & Silverman, 1978)—(in section 4.2). Consequently, the domain definition may also be deficient with respect to representing and

assessing qualitative features of knowledge.

3) Existing methods for developing and selecting test content do not adequately identify the task and problem characteristics that most effectively discriminate among levels of performance. Existing methods design tests to sample representatively job knowledge content based on ratings of job importance and salience. However, these ratings used to guide the selection of test questions show little relationship to the diagnosticity of test questions (Carrier, Dalessio, & Brown, 1990). Brown & VanLehn (1980) suggest that specially designed problems are required for reliably identifying performance differences—(in section 4.3). Existing methods for selecting test content may inadequately identify and sample the diagnostic job knowledge.

Thus, our approach to improving job knowledge test design rests upon the following three premises about the potential contributions from cognitive science: a cognitively-oriented task analysis of actual performance will identify unique domain content with respect a) to the substantive domain and b) to the processes and types of information used; and c) selection of test content based on its diagnosticity will improve test psychometrics.

4.0 Contributions from Cognitive Science

A substantial research base in cognitive science addresses these deficiencies in job knowledge measurement. We first describe an important finding of cognitive task analyses about differences between factual, textbook knowledge and what job incumbents know. Then we discuss cognitive methods for defining the knowledge domain and test specifications. Consistent with usage in the knowledge engineering community, these two activities are termed knowledge elicitation and knowledge representation.

4.1 Definition of the task domain. Textbook rules and facts constitute a limited approximation of the task knowledge underlying performance. Instructional materials, including both textual descriptions and diagrams, are naturally incomplete and require a great deal of inferencing by the learner in order to form executable procedures (Kieras, 1990). Indeed, the inherent incompleteness of instructed material provides the opportunity for learners to acquire incorrect "knowledge". For example, students acquire incorrect variations of a subtraction procedure, such as the smaller-from-larger bug, that were certainly not instructed (Van Lehn, 1983). This conclusion about the content of training materials casts suspicion on tests of performance knowledge based largely on an analysis of training materials.

Cognitive science research suggests that tests of knowledge and understanding should be drawn from what job incumbents actually know, rather than the training materials themselves. Of course, knowledge is not itself directly observable, and cannot be obtained

with accuracy by simply asking individuals to describe what they know. Subjects readily provide an account of task knowledge and reasoning that can differ substantially from the manner in which they actually perform the task, particularly when the account is obtained outside the context of the actual task (Ericsson & Simon, 1986; Nisbett & Wilson, 1977). In the following section we review some of the methods for uncovering task knowledge.

4.2 Definition of the knowledge domain. A fundamental strategy underlying many cognitive approaches to identifying subjects' knowledge is to capture the contents of current awareness during the performance of relevant tasks. Based on an information processing perspective of cognition, this approach asserts that subjects' verbal reports of the contents of current awareness and short term memory provide useful data about their cognitive processes (Ericsson & Simon, 1984). We next discuss three methods for eliciting subjects' knowledge.

4.2.1 Methods for knowledge elicitation. Cognitive science methods for capturing knowledge depend heavily on subject verbalization, but aim to preserve the task context in order to cue the relevant subject knowledge. In many cases, the cognitive scientist examines actual task performance while obtaining audio-taped verbal protocols (Newell and Simon, 1972). Using this method, subjects provide a continuous verbalization of their knowledge in the form of the goals, hypotheses, task objects, and inferences passing through working memory while executing the task of interest. This method has lead to significant scientific understanding of a great many cognitive tasks, including scientific problem solving, computer programming, and engineering design, etc.

A close relative to verbal protocols is "coaching" developed by Gelman and Gallistel (1978) and now incorporated into knowledge engineering techniques for expert systems (Hayes-Roth, Waterman & Lenat, 1983). According to this method, the subject coaches a collaborator in the experiment, thereby revealing how she thinks of the task while she executes it.

The analysis of ordinary communication within a team (Orasanu & Fischer, 1992) also provides access to the content of cognition. Here task knowledge is explicitly and naturally verbalized through an exchange with another participating agent.

All of these methods assume that the role of the experimenter is largely passive, serving to elicit verbalizations, but not to suggest interpretations of the knowledge. Yet, social interactions constitute an important influence on the data.

4.3 Knowledge representation: organization and structure. There are two lines of research within cognitive science that concern knowledge organization, and the manner in which knowledge is utilized during task performance. One line of research addresses computational cognitive architectures. A second line of research addresses qualitative differences in knowledge content associated with successful task performance. We discuss

each of these in turn, before addressing the manner in which we have applied this research to the development of a test of knowledge for land navigation.

4.3.1 Computational cognitive architectures. A "computational" model of task knowledge consists of computer code for actually performing the task in a symbolized task environment. The primary motivations for generating a computational model are completeness and explicitness; an executable specification of the task demonstrates a high degree of scientific understanding of the knowledge and processes involved. The model is usually based on one of a few theories of cognitive architecture. The theories are expressed as higher-level programming shells with components corresponding to familiar concepts from cognitive psychology including long term memory and attentional processes. There are a number of alternative architectures suitable for modeling knowledge-based performance and the changes in knowledge organization that occur with practice.

Production system architectures simulate the selection and sequential application of knowledge represented as a set of if-then inference rules. Accordingly, if a certain pattern of conditions occurs in a mental representation of the task, then a certain inference can be applied to modify the set of conditions. Often, a special condition element in the inference rule associates rules with particular tasks or task goals. In the two most commonly applied cognitive architectures, ACT* (Anderson, 1983) and Soar (Newell, 1991), the acquisition of expertise is reflected in predicted performance time, and in some cases, intermediate errors. Soar models expertise as the acquisition of chunked procedures, or groups of individual inference rules packaged together as the result of successful problem solving experience. ACT* distinguishes between proceduralized knowledge represented in a knowledge base of productions, and declarative knowledge represented as a network of concepts and relations. ACT* models expertise primarily as the transfer of declarative knowledge to procedural knowledge acquired with experience.

While production system architectures are perhaps the most dominant within the subset of cognitive science known as cognitive psychology, other computational architectures are also relevant to understanding the dimensions of knowledge pertinent to job knowledge testing. Frame-based architectures (Schank & Abelson, 1977; Bobrow & Winograd, 1977) package together a domain-specific task decomposition with complex patterns of task conditions and procedures. In some respects frame-based architectures avoid the need to search an extensive base of loosely organized rules for problem solving. Hierarchical planning systems (Sacerdoti, 1977) make use of different levels of abstraction in the representation of a planning problem. Planning problems often impose constraints on the *design* of a procedure that are only implicit in the procedure itself, for example constraints on the consumption of resources and the ordering of steps. A number of architectures and frameworks address the perceptual aspects of task performance, including neural network models (Hinton & Anderson, 1981; Rumelhart, McClelland, & the PDP research group, 1986), and models that accept analogue (rather than pre-processed/symbolic) task inputs

(Kosslyn, Flynn, Amsterdam, & Wang, 1990; Uttal, Bradshaw, Dayanand, Lovell, Shepherd, Kakarala, Skifsted, & Tupper, 1992).

Standard criteria for selecting between different architectures are not established. Many of the alternative architectures are actually modeling different aspects of cognitive behavior, and are candidates for merger (J.A. Anderson, 1990). A negative feature of all computational methods for describing task knowledge is that they are very time-consuming to program. We suggest below that the necessary level of analysis for job knowledge testing is more general than the code of a computational model. Rather than choose one architecture over another for constructing a limited computational model of land navigation, we borrowed general concepts freely from nearly all of them to address our broad applications concern.

4.3.2 Qualitative components of knowledge. The cognitive architectures identify broad characteristics of knowledge. But the contents of knowledge in a particular model for a particular task ultimately govern the manner in which the task is performed. Newell (1982; 1991) suggested that the knowledge content of a model is an abstraction of the computational architectures. That is, important differences between experts and novices may be more apparent at a knowledge level analysis of a task rather than a symbolic representation of task knowledge. The literatures on expertise and instruction address qualitative differences in the types of knowledge associated with successful task performance. We discuss four types of content differences below: 1) ontology, 2) explanatory knowledge, 3) tacit knowledge, and 4) goal recognition knowledge.

Problem ontology refers to the objects, relations and attributes incorporated into the knowledge (Greeno, 1989). Chi, Feltovich & Glaser (1981) demonstrated that novices think about the diagrams for elementary physics problems in terms of everyday objects and relations (e.g., blocks on inclined planes). In contrast, experts think about the same diagram in terms of more abstract, but domain-specific, objects such as forces. Thus, even though we might make use of a production system architecture to model both expert and novice knowledge, the critical difference between the resulting models lies in the objects, relations, and attributes incorporated in each model, rather than the ratio of declarative to procedural knowledge, or the presence of chunked procedures.

Explanatory knowledge provides a rationale for why the procedure is so constructed, justifies the design of task procedures, or determines which principles must be respected for the procedure to remain valid. Greeno identified implicit knowledge of this kind for counting objects (Greeno, Riley & Gelman, 1984; Smith, Greeno & Vitolo, 1989). VanLehn, Jones, & Chi (1991) suggest that explicit explanatory knowledge is involved in the acquisition of new procedures from textbook examples.

A third type of knowledge content applies to decision-making in the task context. Sternberg and his colleagues refer to this as tacit knowledge (Sternberg, 1985; Wagner &

Sternberg, 1985). They state that this knowledge involves selective encoding to discern the relevant conditions for task goals; selective combination to integrate disparate pieces of tacit information; and selective comparison to decide when and how to bring past experience to bear in current situations. Tacit knowledge is thought to increase with job experience and is typically acquired from the task context and by modelling or simulation.

A fourth aspect of knowledge content supports the recognition of goal attainment. Greeno & Simon (1988) identified two different kinds of goal recognition problems. The first kind of problem associates criteria for success with the final *state* achieved by the problem, e.g., solving the 15-puzzle by sequencing the tiles in numerical order. Here, the final state is the sole criteria for determining success. The second kind of problem associates criteria for success with the *process* used to solve the problem, e.g., constructing a proof, or following the procedure for multi-digit subtraction. Here the final state is only partially informative about the correctness of the solution. In addition the process for arriving at the solution determines whether or not the problem was solved correctly. The knowledge required to ascertain goal attainment can be substantial. Even when the final state is the sole concern, the criteria for goal attainment can be ill-defined, as in the knowledge for identifying the attainment of a fugue (Reitman, 1965). And when goal attainment is based on having followed the correct procedure, knowledge is required to ensure that slight variations in response to a specific task environment have not violated the requirements of the procedure.

We have described four types of content differences associated with expert performance. Early studies of knowledge for multi-digit subtraction distinguished two models for distinguishing expert knowledge from novice knowledge. According to the overlay model, novice knowledge is a subset of expert knowledge. Young & O'Shea (1981) implemented incorrect multi-digit subtraction as the absence of condition elements on otherwise correct production rules. But VanLehn (Brown & VanLehn, 1980; VanLehn, 1983) suggests that novice knowledge includes the addition of repairs to incomplete knowledge, that is, it incorporates "bugs". Studies of novice knowledge in mechanics seem to exemplify the buggy model better than the overlay model, due to the presence of misconceptions (McCloskey, Carmazza, & Green, 1980). The availability of explanatory principles governing the design of correct procedures allows the learner to successfully adapt old knowledge to new and unusual constraints. However, uncovering the bugs requires time consuming empirical work since the correct model does not provide any guidance regarding possible misconceptions.

4.4 Test scoring. Job knowledge has traditionally been conceptualized as an accumulation of facts, principles, and procedures related to work performance. As a result, psychometric methods for scoring tests model knowledge as a single, continuous variable. In contrast, recent work in cognitive and educational psychology conceptualizes knowledge in terms of mental models; strategies for problem-solving, encoding and retrieving information; pattern

recognition; etc. This view is aptly expressed as follows:

> *Learners increase their competence not by simply accumulating new facts and skills, but by reconfiguring their knowledge structures, by automating procedures and chunking information to reduce memory loads, and by developing strategies and models that tell them when and how facts and skills are relevant.*
> (Mislevy, 1989, p. 1)

Such a view of the psychological model of knowledge underlying performance points out the limitations of existing psychometric procedures for modelling the knowledge states of individuals. Both classical test theory and item response theory are statistical frameworks which yield test scores based on a psychological model of overall proficiency. Such test scores are very useful for selecting the best group of individuals from an applicant pool or assigning entry personnel to training programs which require increasing levels of ability. That is, characterizing examinees in terms of their likelihood of responding correctly provides useful information when decisions are linearly ordered. However, these methods are less well suited to modelling and informing decisions when a range of non-linear options are available, for example, about placing individuals in training or tailoring instruction .

One example of this difficulty for existing psychometric methods can be expressed in terms of instructional options for two persons who receive the same score on a test. One person may possess superior knowledge and the other may have employed a superior problem-solving strategy to achieve the same score. Optimal instruction for each person would differ if this information were known.

Similarly, in a proportional reasoning task Béland (1992) found that sometimes an improved strategy can lead to getting items wrong that were previously answered correctly. That is, sub-optimal strategies can sometimes lead to correct responses for the wrong reasons. Conversely, optimal strategies can be imperfectly executed. Neither classical test theory or item response theory are capable of modelling these configural response patterns.

Another implication of cognitive models of knowledge is that even response errors can be informative about an individual's level of task understanding. Yet classical test theory does not readily provide a framework for interpreting and integrating information from incorrect answers. Other response qualities can also provide useful information about a person's level of expertise. For example, the speed of response is often taken as an indicator of proficiency based on improved skill automaticity found in more skilled performers.

When knowledge can be adequately characterized by a single underlying dimension, item response theory (IRT) provides an adequate statistical model for inferring the unobserved ability variable from the observed pattern of responses to test questions. IRT

models the probability of a correct response to a given test question as a function of an examinee's ability or performance proficiency (usually denoted by θ) and the measurement properties of the test question (parameters for θ, discrimination, guessing, etc.). Item parameters specify the degree and structure of the relationship between observable item responses and the latent variable θ by quantifying the conditional probabilities of item responses, given θ. Thus, θ is statistically inferred from the observable responses in the form of a likelihood function. The coupling of probability-based inference with a simple model for overall proficiency provides the foundation for test construction, equating, adaptive testing and validation research (Béland & Mislevy, 1992).

However, when decision alternatives are based on more complex models of knowledge, then an alternative approach to statistical inference is required. To improve measurement of job knowledge, a framework that characterizes knowledge in terms of strategies, task understanding, and procedural knowledge was needed. Following previous work in artificial intelligence (e.g., Pearl, 1988) and psychometrics (Béland & Mislevy, 1992; Masters & Mislevy, 1991; Mislevy, 1989, 1991; Mislevy, Yamamoto, & Anacker, 1991), we propose to examine probability-based inference networks to model multiple unobserved variables as an alternative to total number correct or IRT methods for scoring tests.

5.0 Implementing Cognitively-Oriented Test Design

Based on our understanding of the cognitive science literatures, we adapted cognitive methods to modify content-oriented test design in three areas: knowledge elicitation, knowledge representation/test specifications, and test scoring. These adaptations of cognitive methods address each of the limitations of content-oriented methods cited in section 3.5. In this section, we describe the methods we employed and discuss the modifications that were made to achieve the application objective.

5.1 Knowledge elicitation for land navigation. This project incorporated all three protocol methods for obtaining insights into subjects' knowledge of land navigation, but also introduced some adaptations to standard practice. Most studies employing protocol analysis typically record with only an audio recorder. In a relatively minor adaptation of standard practice, we employed a video camera to record the visual aspect of the land navigation task. Video recording captures the steady variation in the task environment, documents the environmental referents in subjects verbal accounts and provides "notes" for later reference.

The video protocols were obtained in eight, 3-hour (approximately) sessions. Each session consisted of presenting a two man team with the task of planning, then executing a navigational route. The routes, located in wilderness areas of the base, covered a distance of about 3 miles. Most of the subjects were nominal experts. Some were instructors of land

navigation, others had several years experience as the navigator for their unit. By utilizing a variety of experts, individual differences in successful performance strategies, skills, and knowledge could be observed. We also observed two novice and two "decayed" experts (i.e., a couple of years had lapsed since performing land navigation tasks) to obtain information about knowledge at several levels of proficiency. Additionally, this knowledge elicitation phase was conducted at two different sites (North Carolina and California) to identify differences in knowledge required for navigating in distinctly different environments: a flat forested terrain and a barren, mountainous terrain.

The following excerpt illustrates a typical protocol of communication between patrol team members at the start of a route through several checkpoints in the woods.

> Staff Sergeant: When we get to 200 meters, approximately uh, turn to the right, cross the stream there, we should see a small draw, a small draw to our right, and a big draw to our left, and a small finger sticking out, sloping down toward us, and approximately say 100 meters, 150 meters from the bend in that stream, [the checkpoint] should be located on the top of those two draws.
> Sergeant: 300 meters from here to here.
> Staff Sergeant: Its 300 meters approximately from one point to the next point, but I'm saying that, from the stream it should be about 100 meters down, down to the bend.....
> Sergeant: Ok.
> Staff Sergeant: .....it should be a 100 meter shot, it should be a 100 meter shot over. But even if we go 300 meters down the stream, we should be able to go to our right, to do it by terrain association, we should still get it. We've got two options, we'll have to decide once we get down there. [After we find the first checkpoint,] to get to our next checkpoint, then uh, approximately, by looking at this, we should be able to head like, uh, west. West, due west, to the checkpoint, which will put us back on the track of about how many meters?
> Sergeant: 5.
> Staff Sergeant: Ok, 500 meters (to the track), 500 meters and we should be able to come down the side of uh, down the hill once again, down the draw, the hill slopes down towards the stream, it should be right inside the fork here, that should be the side we should be coming to (motions behind him)
> Sergeant: Yeah.
> Staff Sergeant: Right inside the fork. Ok, we cross that, we approximately get to that uh, shape coming between the small draw, leading up to uh the next box. It should be on the top of the hill and to our right (motions to his right). So that's uh, basically what we are doing. We should be able to find the [checkpoint] with no problem.

Perhaps the most surprising aspect of this interaction is the availability of terminology for the perceptual aspects of the task. We expected the knowledge regarding such aspects of the task to be tacit. However, we note that accepted navigation methods govern task performance in this domain, and that all team members must share the same model of terrain differences in general, and the terrain for the specific case in question. Because the domain requires coordinated team performance, we suggest that a vocabulary has evolved to correspond to those aspects of the task model that must be shared. Thus, while verbal protocols may provide limited insight into the manner in which perceptual knowledge is represented or processed, protocols like this one are sufficient to support the design of test items that tap the circumstances of knowledge use.

Ultimately, the standard methods described above were insufficient for our task in several respects. The amount of cognitive activity per unit time in a multi-dimensional, dynamic task precludes a complete determination of all of the cognitive content by concurrent verbalization, simply because it takes time to speak. Thus, some amount of retrospective protocol was required. Further, the typical interventions by the experimenter were somewhat more intrusive than standard, and included the following domain-specific questions, designed to elicit specific, retrospective protocols, for example:

How do you know you are at the checkpoint?
How do you maintain your direction while crossing the stream?
What points are you using for the resection?
Why are you using those points?
Where are we now?

Real world task domains provide an enormous variety of task contexts, each one potentially associated with unique knowledge. Consequently, we decided to employ the use of directed probes to more fully explore task understanding and the variety of task situations. However, by introducing such questions, we depart from optimal protocol procedures. In addition to eliciting a post hoc protocol, the questions themselves assume a mutual understanding about current tasks goals, and potentially bias subject responses. These departures from standard methodology involve a tradeoff between accuracy and the efficiency and comprehensiveness of coverage.

Additionally, we employed the critical incident methodology (Flanagan, 1954; Smith & Kendall, 1963) to obtain reports from current instructors and from participants in recent military operations requiring land navigation (e.g., Operation Desert Storm). In both cases, the primary focus of discussion was to describe the kind of errors they have observed.

This completes our discussion of the cognitive science methods adapted for uncovering task knowledge. The job knowledge tests we develop will only be successful to the extent that our liberal methodological adaptations to the task domain and applications goals in fact lead to appropriate inferences regarding job knowledge.

5.2 Knowledge representation. The next step involved encoding the protocols into a systematic representation of land navigation knowledge. We selected the plan-goal graph (Geddes, 1989; inspired by work from Schank & Abelson, 1977) as a framework to organize our observations of land navigation. This framework was sufficiently flexible to incorporate a model of knowledge adapted from the cognitive literature (as discussed in section 4.2). To assist in test construction, we transformed the plan-goal graph into tabular or matrix form. In this section, we describe the procedures employed for these two forms of knowledge representation.

5.2.1 Plan-goal graphs. A goal corresponds to a desired state of the world. A goal is satisfied by any one of its child plans. A plan specifies a method for satisfying a goal. The same goal may be satisfied by very different methods. Domain plans and goals at various levels of abstraction are organized by a plan-goal graph (Rouse, Geddes, & Hammer, 1990; Sewell & Geddes, 1990). The plan-goal graph decomposes the most abstract purposes of the task into increasingly resolved descriptions, until the descriptions are sufficiently detailed and complete to be executable. The internal nodes specify the relationship between the executable procedures and the overall goal structure of the domain; they justify the execution of child procedures in terms of the specific purposes they achieve, and the task features they accommodate.

Figure 2 contains a portion of the plan-goal graph for the domain of land navigation. The plans are contained in boxes and the goals are contained in ovals. Thus, the "follow the road plan" and the "distant features plan" constitute two of the nine different methods for achieving the goal of "traveling in the direction of the next point". The different methods are potentially disjunctive; executing anyone of them will satisfy the goal of traveling in the direction of the next point. The "compass at night plan" contains two conjunctive goals. Both "knowing the bezel ring setting" and "setting the bezel ring" must be accomplished to achieve the parent plan.

The plan-goal graph has two advantages for the applications problem under consideration. First, the plan-goal graph clearly illustrates the domain-specific goal structure of performance, an important element of job knowledge. Second, by requiring knowledge to be linked to task goals, it justifies the claim that the tested knowledge is directly relevant to task performance.

In addition, the plan-goal graph highlights important modeling decisions associated with task decomposition that are not usually emphasized by the more traditional cognitive architectures applied to basic research task domains. One issue is the arbitrariness of the leaf node level of analysis on task knowledge; the decomposition proceeds until the modeler determines that the plans are operational. A second issue is the criterion for distinguishing two different plans for the same goal. The criterion we use for distinguishing two different plans is when the candidates involve qualitatively different concepts that cannot be captured

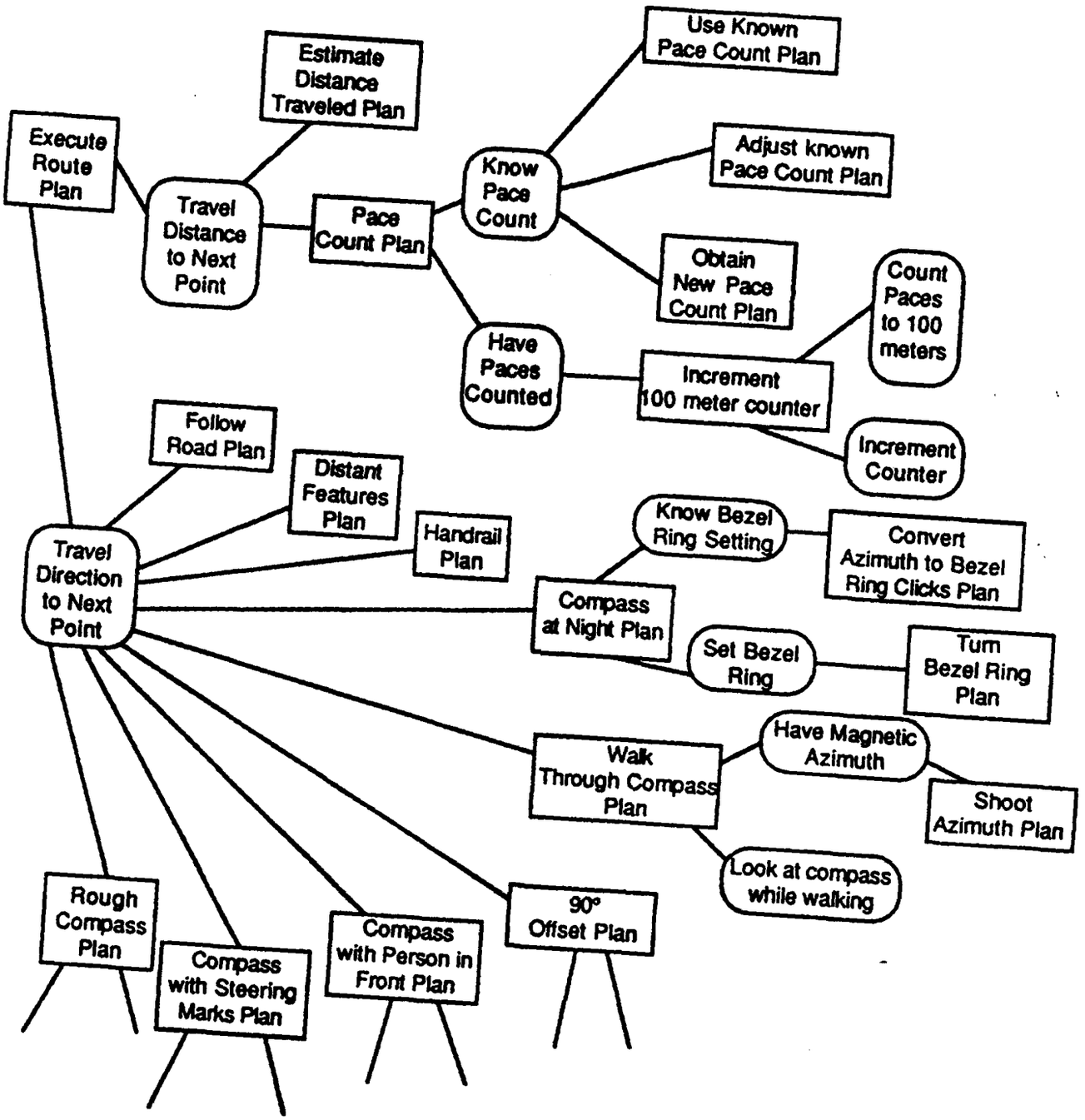


Figure 2. PORTION OF A PLAN-GOAL GRAPH FOR U.S. MARINE LAND NAVIGATION.

by adjusting the range of a quantitative parameter (Geddes, 1989). For example, the seven different plans for knowing the direction of the next checkpoint involve different tools, and different features of the task environment. When two plans do share knowledge, it is indicated by having them point to the same lower-level goal and plan in the decomposition.

To accommodate the various characteristics of knowledge identified previously, we annotated each plan in the plan and goal graph for land navigation with the following information, derived in part from the qualitative components of knowledge described in section 4.3.2:

1. General Concepts and Principles: The explanation for why the plan works. If these concepts or principles are violated, the plan will not be successful. For example, following a map of a terrain works because there is a one-to-one correspondence between the location of features of the map and the location of features in the environment. If this weren't true, following a map would not facilitate navigation. Knowledge of the essential principles behind the design of task procedures allows for the adaptation of task knowledge to unusual circumstances.

2. *Procedure Selection:* Knowledge guiding the choice of alternative procedures for a given situation.

3. Procedure Execution: Knowledge of the steps or sub-goals for executing a procedure.

4. Goal Attainment: Knowledge for determining the priority, sequencing and standards governing plan completion.

5. Pattern Recognition: Specific physical and conceptual information that must be present and are used to support the execution of the plan.

These five components of knowledge formed the basis for our model of land navigation performance. In addition to these five components, we also annotated the plan-goal graph with candidate test questions and descriptions of typical mistakes.

The annotated plans acknowledge a number of the issues raised in the previous section. The specific annotations are compatible with a recent presentation by Estes (1992). The most inclusive published reference for the basic components of this annotation is Smith, Greeno, and Vitolo (1989). Smith et al. distinguished between specific procedures for executing a goal, the perceptually-based knowledge for interpreting the world, and the abstract knowledge for generating a specific procedure for a specific task environment.

The plan-goal graph suited the intermediate level of analyses that we adopted to meet project goals Thus, we did not formally encode or analyze the protocol data, nor did we implement a computational cognitive model. Rather, we informally associated the protocol data with the plan-goal graph, guided qualitatively by a Bayesian framework of successively revised expectations, developed first from the scientific literature and land navigation training materials, and refined through several protocol gathering sessions. This approach very substantially reduced the time, personnel, and other costs that would incur from more formal data analytic methods.

5.2.2 Specifying test content. We next transformed the plan-goal graph into matrix form in order to develop specifications for the number and content of test questions. The matrix is defined by five content areas and five qualitative components of knowledge. The content areas consist of a complete list of land navigation procedures organized into five content categories: planning, location, distance, direction and movement. The columns of the matrix specify the five qualitative components of knowledge that were used to annotate the plan-goal graph.

Using this matrix representation of job knowledge, we implemented two changes to content-oriented test development procedures. First, specifying and sampling test content across the different components of knowledge is an addition to content-oriented test procedures that focus solely on the categories and elements of domain content. Consequently, content-oriented tests appear to consist predominantly of questions which tap knowledge of how to execute procedures and declarative knowledge. The addition of questions that address other components of knowledge will be important to the extent of their contribution to predicting performance, distinguishing levels of expertise, and diagnosing training deficiencies.

The second modification to content-oriented procedures for specifying test content involves the method for how job knowledge is sampled to construct the test. A test based on representative sampling of job knowledge ensures that test content closely resembles the job. Consistent with content-oriented test development procedures, we used this matrix organization to develop a sampling plan for specifying test content based on the job knowledge domain. In contrast to content-oriented procedures, the job knowledge domain was representatively sampled based on experts' ratings of relative diagnosticity rather than ratings of the importance or frequency of use of the knowledge. Content selected on the basis of importance does not ensure that a test will effectively discriminate among levels of performance--jobs frequently contain tasks which are important but are performed well by most incumbents. Thus, we convened a panel of expert navigators to rate the relative diagnosticity of the categories and procedures of the matrix representation of knowledge. The resulting test specifications are displayed in Table 1.

Table 1

SPECIFICATIONS FOR A COGNITIVELY-ORIENTED TEST OF LAND NAVIGATION

		Component Categories					
	Content Categories	Principles Concepts A	Procedure Selection B	Procedure Execution C	Goal Attainment D	Pattern Recognition E	Row Totals
1	Planning	4	3	6	3	4	20
2	Location	6	4	8	4	6	28
3	Distance	4	2	4	2	4	16
4	Direction	4	2	4	2	4	16
5	Moving	4	4	6	2	4	20
Column Totals		22	15	28	13	22	100

Note: Numbers represent percentage of the total number of test questions.

We used this structure to specify the number and type of questions to be developed for the test. The numbers in Table 1 represent percentages of total test content. The matrix defined the job knowledge test in terms of content and structure at three levels of analysis—knowledge categories, procedures, and elements. The matrix categories were used to determine the relative distribution of test questions across the content categories and components of knowledge, based on ratings of diagnosticity.

A complete list of land navigation procedures constituted the second level of analysis. Having previously determined the number of test questions to be selected from each content category, we specified which procedures to include in the test based on the relative diagnosticity of each procedure within each category. Thus, at the procedure level, the plan-goal graph guided the writing of test questions by specifying which procedures to test and by identifying the tasks, conditions and standards required for each land navigation procedure.

5.2.3 Developing test questions. In addition to providing a rationale for test specifications, representing job knowledge in matrix form also provided explicit guidance for the development of test questions. At the element level of analysis, the matrix informed the construction of response alternatives by describing the components of knowledge which support each procedure and by identifying the types and distribution of errors committed for each land navigation task. An excerpt of the matrix at the element level is presented in Table 2.

Table 2

EXTRACT OF A KNOWLEDGE ELEMENT MATRIX

Content Category: Location Model Level: Knowledge Element Components/Steps	Procedure Determine Grid Coordinates	 Determine Position By Terrain Association
Procedure Execution		
Identify correct grid zone	X	
Align protractor scale to grid line	X	
Read right, then up for grid coordinate	X	
Read digits to correct precision	X	
Orient the map		X
Scan the ground		X
Identify major & unique features		X
Compare shape, size, orientation, slope		X
Concepts & Principles		
Properties of identifiable location		X
Grid representation of geography	X	
Procedure Selection		
Select location finding method		
Select major, unique features		X
Goal Knowledge		
Read coordinates at center of point	X	
Confirm location using 3+ features		X
Pattern Recognition		
Use grid zone designator	X	
Must identify recognizable features		X
Features must be on map		X
Map symbols, legend info	X	
Terrain features on ground		X
Terrain features on map		X
Errors		
Use wrong grid scale	X	
Insufficent precision	X	
Place protractor incorrectly	X	
Reads up, then right	X	
Fails to use grid zone designator	X	

To further guide the design of test items, we used the simple overlay heuristic that any one of these aspects of knowledge might be missing in a given student. For example, students may know a procedure quite well, but not know how to execute it with a particular environmental feature or examinees may lack important knowledge behind why the procedure works. To test for understanding of principles, we developed transfer problems (Wertheimer, 1945) that varied specific task details while depending upon the same explanatory knowledge. Because cognitive science stresses the importance of uncovering job knowledge within the context of the real task, we also reasoned that a test of job knowledge should maintain the task context to whatever extent possible and that test questions should reflect the information processes used on the job. Example test questions for the knowledge components of principles, procedure selection and goal recognition are presented in Figure 3.

Principles

You are in a foreign country with your 1:50,000 scale map and a compass, but without your protractor. The host country has four defective protractors. Which single protractor will be the most useful to you?

1. protractor is missing the mile and degree scales
2. protractor is missing the 1:50,000 coordinate scale
3. protractor is missing the index line
4. protractor is missing the 1:100,000 coordinate scale

Procedure Selection

Corporal Fellows is conducting a security patrol. He has just entered a marshy area and lost sight of his steering mark. What is the best course of action for him to follow?

1. Get through the marshy area as quickly as possible so that he can find his steering mark again.
2. Continue in the general direction of his objective.
3. Continue forward in a straight line until his steering mark comes back into view.
4. Immediately stop and, using his compass, select another steering mark.

Goal Attainment

You are planning a route to cross hilly, forested terrain. About how far apart should the checkpoints be?

1. one checkpoint every 200 meters
2. one checkpoint every 600 meters
3. one checkpoint every 1200 meters
4. one checkpoint every 2000 meters

Figure 3. EXAMPLE TEST QUESTIONS.

5.3 Test scoring. The research goal of the project was to improve the correspondence of job knowledge tests with hands-on tests of performance. Based on the results of the task analyses, it was apparent that task strategies guided when and how information was utilized for performance. The task analysis identified three such strategies: terrain association, map and compass, and a mixed strategy. Although there is some overlap in knowledge and skills required for successful performance of each navigational strategy, there remains a substantial amount of unique knowledge required to perform each strategy effectively. The cognitive task analysis suggests not only the importance of strategy to land navigation knowledge, but that strategy may interact with specific subsets of knowledge to produce configural patterns of responses that could further improve performance predictions. That is, if only certain subsets of knowledge are relevant to each strategy and if the strategy employed for performance could be known, then scoring only the subset of knowledge related to the strategy employed by the examinee could improve performance predictions. Based on Thorndyke's "identical elements" approach to transfer of training, this would extend performance measurement from modelling group to individual level performance. Specifically, we hypothesized that an explicit account of task strategy in test scoring procedures would improve the correspondence of knowledge and performance scores.

Thus, in order to score the job knowledge test results, we needed to be able to characterize knowledge in terms of several strategy by content knowledge states. This contrasts with item-correct scoring and item response theory (IRT), which both model knowledge in terms of a single underlying trait.

In order to represent and evaluate the effects of task strategies on performance, we made use of recent developments in student modelling from the field of artificial intelligence (Lauritzen & Spiegelhalter, 1988; Mislevy, Yamamoto, & Anacker, 1991). They employed probability-based inference networks to model the many-to-many relationships found among concepts in learning. To describe how this approach might be useful for scoring job knowledge tests, we first present an application to medical diagnosis.

Andreassen and colleagues (Andreassen, Woldbye, Falck, and Andersen, 1987) developed an inference network to model diagnostic knowledge in the domain of electromyography--the relations among nerves and muscles. A graphic representation of this network is presented in Figure 4. The variables are presented as nodes and the associations between them are represented by links. In this application, the inference of interest is the diagnosis of one of several possible disease states from observable information, such as symptoms, tests, patient history, and the like.

Three possible diseases are represented in the figure: chronic axonal neuropathy, myopathy, and myotonic dystrophy (shown on the left). These diseases are each associated with a pattern of observed symptoms, or syndromes (shown in the middle of the figure). The observable symptoms and test results are represented on the right hand side of the

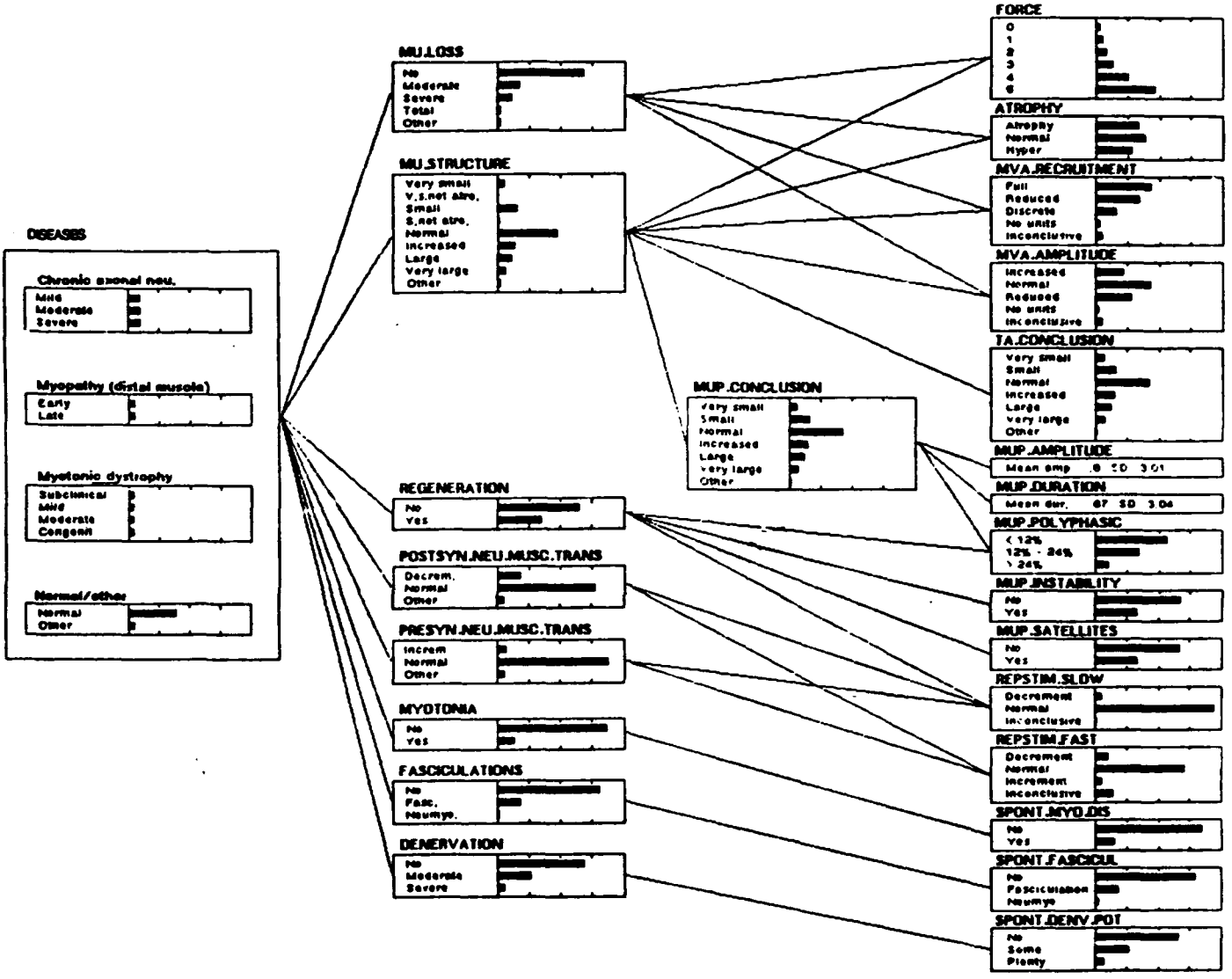


Figure 4. A PROBABILITY-BASED INFERENCE NETWORK FOR MEDICAL DIAGNOSIS.

figure. The links between the nodes in the network indicate the conditional probability relationships existing between the disease states, syndromes, and observable information.

The state of knowledge about a patient before a diagnosis begins is expressed by population base rates for each variable and the conditional probability relationships between variables defined by the links. The conditional probabilities may be determined logically, empirically, rationally, or through model-based statistical estimation (Mislevy, et al., 1991). In constructing the network, the flow of reasoning is from diseases (on the left) to syndromes (in the middle) to observed states. That is, given a disease state, the probabilities for the relevant syndromes are specified. Also, given a syndrome, the probabilities for the observable variables (tests, etc.) are entered into the network.

Actual medical diagnoses proceed in the opposite direction, reasoning from observable symptoms and tests to infer a possible disease state. Within the network, this is represented by propagating the evidence from the observable variable nodes throughout the network. The computations involve algorithms which are generalizations of Bayes' Theorem. As each new piece of evidence becomes available, the network is updated to represent the current state of evidence.

The inference network explicitly handles uncertainty in inferring unobserved variables from observed ones by modelling it as a probability distribution (Mislevy, 1989). Thus, it is not necessary to have absolute knowledge or a causal theory in order to construct a network and model important decision applications (Charniak, 1991).

A key concept for developing and implementing inference networks involves the notion of conditional independence. Conditional independence is defined as a set of variables that may be interrelated in a population, but independent given the values of another set of variables. For example, in item response theory, observed item responses are caused by and explained in terms of the unobserved ability variable θ. For tests of the same construct, item responses are typically highly intercorrelated. However, given a specified value of θ they are not substantially correlated--that is, they are independent conditioned on ability.

Conditional independence is utilized in inference networks to vastly improve the efficiency of computing the relevant probability distribution. For example, to completely specify a distribution involving 10 binary random variables, the total number of probabilities to specify would be 1023 (i.e., 2^n-1). As the number of variables increased, the number of probabilities to specify would increase exponentially, making most realistic applications unfeasible to develop or compute. For example, a multiple choice test consisting of 100 items scored right or wrong would involve specifying 1.26×10^{30} probabilities. However, by making use of conditional independence to specify the links between variables into a network, the number of probabilities to specify is very substantially reduced. In the case of the 100 item test, 200 probabilities would be required. Furthermore, recent advances (e.g.,

Lauritzen & Spiegelhalter, 1988) have greatly improved the efficiency of the computations required to update large networks of nodes by making use of local operations on small subsets of interrelated variables (i.e., grouped through conditional independence). In sum, the theory and knowledge of a substantive area can be applied to structure an efficient inference network.

Inference networks have been applied to medical diagnoses (Andreassen et al 1987; Heckerman, 1990; Spiegelhalter, Franklin, and Bull, 1989), language understanding (Charniak & Goldman, 1989), proportional reasoning (Béland & Mislevy, 1992), vision (Levitt, Mullin, & Binford, 1989), and heuristic search (Hansson & Mayer, 1989). As noted by Béland and Mislevy (1992), inference network representations are similar in structure to other efforts to explain observed variables from latent ones, for example, the use of factor analysis to identify the latent structure of cognitive abilities, path analysis in sociological and economic applications, analysis of genotypes in animal husbandry. Further, these methods share the inferential logic of conditional probability relationships found in Spearman's (e.g.,1907) early work with latent variables, Wright's (1934) path analysis, Lazarfeld's (1950) latent class models, and Jörskog and Sörbom's (1989) LISREL diagrams.

There are additional motivations for considering an inference network approach to test scoring. This approach can accommodate different types of test information that are difficult to incorporate into total correct or IRT scoring. For example, information from incorrect responses can be informative about an examinee's level of understanding of the task and the speed of response can indicate the degree of skill acquisition. Additionally, the more detailed representation of knowledge available in an inference network is informative to training decisions as well as performance measurement or selection uses, thereby increasing the potential utility of this approach.

5.3.1 A Land Navigation Inference Network. Defining task strategy as the primary feature of a model of land navigation knowledge affected the topology of the inference network representation in two ways. First, as can be seen in Figure 5, the selection of which strategy to employ for the terrain is included as a node (top left node in the figure) in the network which directly impacts expertise. Second, the 100 land navigation multiple choice questions were divided into one of four scales according to which navigational strategy they were related: terrain, compass, mixed, or all three. The terrain strategy involved using major terrain features to locate one's position and guide one's movements. The compass strategy consists of locating position and directing movements by reading compass azimuths and by tracking distances moved along azimuths. Some Marines flexibly made use of both strategies, switching back and forth as needed. We labelled this a mixed strategy. Corresponding to effective performance of the mixed strategy was the knowledge of conditions for selecting the optimal strategy for each situation. Thus, we labelled the subset of knowledge related to mixed strategy use, "JKSelect". Finally, some knowledge was common to all 3 strategies, such as basic map reading skills. Based on the content, we

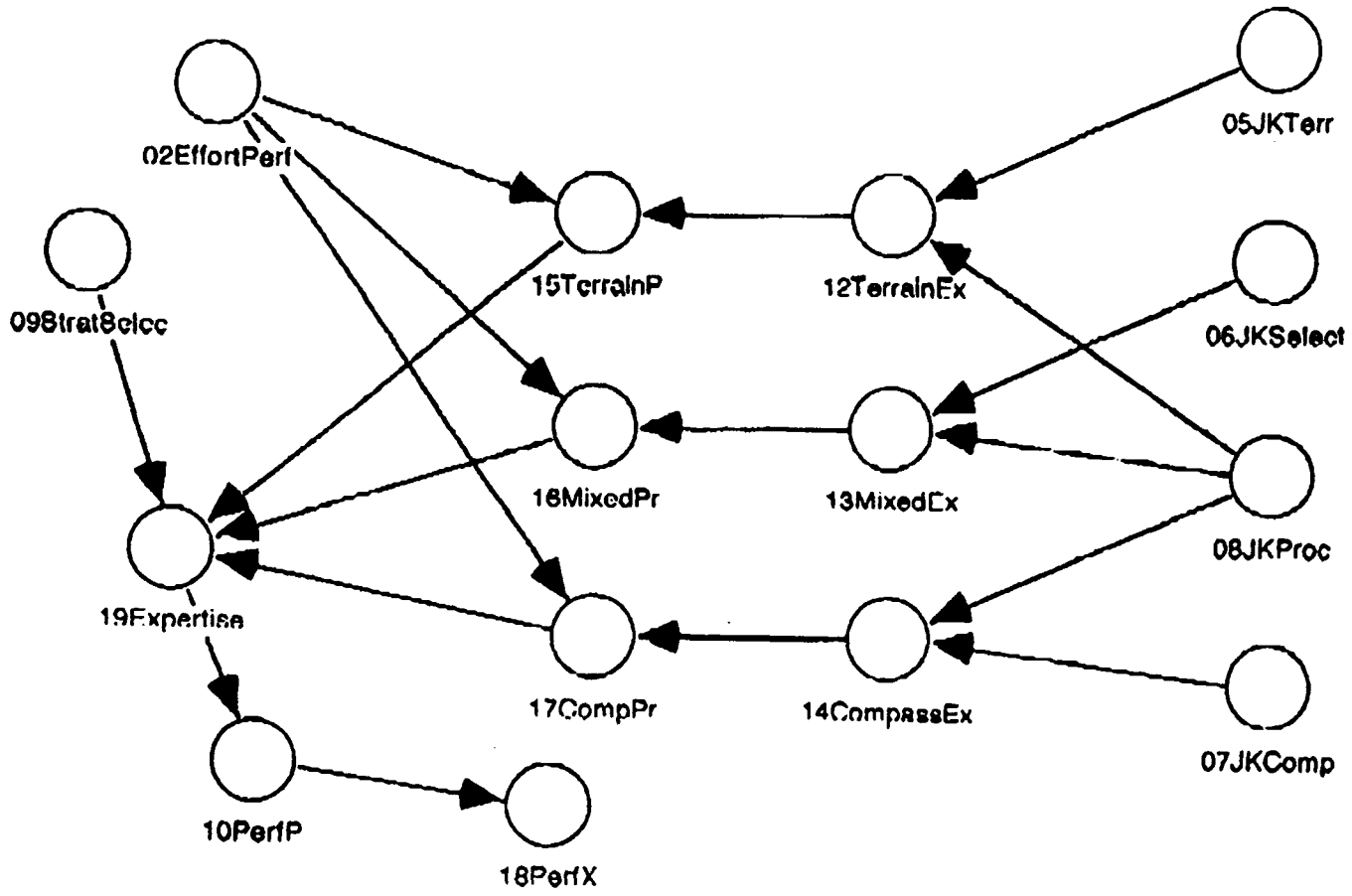


Figure 5. A PROBABILITY-BASED INFERENCE NETWORK MODEL OF LAND NAVIGATION KNOWLEDGE.

labelled this knowledge scale "procedure execution" (the JKProc node in Figure 5). Accordingly, this node is shown with arrows to each of the nodes representing strategy execution. Taken together, the representation of knowledge displayed by the inference network in Figure 5 indicates that successful land navigation performance is a funetion of the selection of a navigational strategy and the effective execution of that strategy.

Although it is too large to display in the report, we constructed a second network to model the relationships of job knowledge to hands-on measures of performance proficiency (in contrast to the work-sample measure modelled in the previous network). The topology of this network was based on modelling the performance on each of the 26 steps involved in the total of the 5 hands-on tests. Nodes for the relevant subset of knowledge (typically about 3 to 10 knowledge items per hands-on step) were linked to each hands-on step, then aggregated across all of the hands-on steps. Selection of items from the knowledge test related to each hands-on test was accomplished based on ratings made by one of the project staff. Some knowledge items were related to more than one hands-on step and a few were rated as not directly related to any hands-on step.

As just described, the topologies of the two networks were defined by the task analysis. The second important element of the inference network is the specification of the strength of relationships denoted by the arrows, or links, in the network. These relationships are expressed as conditional probabilities for the "child" nodes and base rates for the "parent" nodes. (Child nodes have incoming arrows, parent nodes have outgoing arrows.) We specified these conditional probabilities and base rates in two ways--rationally and empirically. We employed only empirical specifications for the first network. The rational specifications were made by the principal investigator, without recourse to the data (i.e., before data analyses were conducted). Empirical specifications were employed for the second network. The network specifications were implemented in the Ergo software program (Noetic Systems, Inc., 1989) which uses an optimized Lauritzen-Spiegelhalter algorithm.

6.0 Summary of Cognitively-Oriented Test Design

We adopted a flexible approach in modifying standard cognitive methods to serve the application objective of improving the correspondence of job knowledge and hands-on measures of performance. Specifically, we implemented three changes to a content-oriented approach to job knowledge test design. First, we employed protocol analyses to identify the information actually used during task performance. Second, we used a model of knowledge consisting of five qualitatively distinct components to ensure complete coverage of the information required for successful task performance. This model was incorporated into the knowledge representation (i.e., plan-goal graph) and the formal test specifications. Third, we selected content from the job knowledge domain based on how well it differentiates among levels of performance. Fourth, we proposed use of probability-based inference network to score the job knowledge test. A comparison of content, cognitive, and the cognitively-oriented methods implemented in this project are summarized in Table 3.

Table 3

A COMPARISON OF METHODS FOR ELICITING AND REPRESENTING KNOWLEDGE

		Methods		
	Activity	Content-Oriented	Cognitive Science	Cognitively-Oriented
1	Task Analysis	Importance Ratings	Verbal Protocols	Video Protocols
2	Knowledge Representation	List of Knowledge Elements/Categories	Computational Model	Plan-Goal Graph
3	Basis For Test Specifications	Importance Ratings	Experimental Analysis	Diagnosticity Ratings

6.1 Methodological tradeoffs. As is evident from Table 3, each adaptation of cognitive methods to job knowledge testing involved some tradeoffs. The general nature of these modifications involved employing a level of analysis which is intermediate to that of personnel and cognitive science. These methods appear to extend and refine the relevance, completeness, and detail of knowledge which can be specified via content-oriented methods alone. However, achieving these improvements while retaining the cost efficiency of job knowledge testing involved several methodological tradeoffs. For example, compared to the number of procedures involved in land navigation (we identified 65), very few protocols were collected. To extend our coverage of the domain, we also made use of retrospective protocols and critical incidents. Additionally, we did not analyze protocols in detail, conduct experimental analyses of expertise, or construct computational model of performance. This approach substantially reduced the time, personnel and other costs that would incur from more formal data analytic methods. On the other hand, it is probable that the inclusion of these more precise procedures would have resulted in additional improvements in the job knowledge test and its correspondence with performance.

An advantage of a cognitively-oriented approach to test design is that these methods appear to be more informative to training needs analyses, diagnoses, and curriculum design than are conventional approaches to performance measurement. By eliciting, representing, and scoring knowledge in terms of a more rich conceptualization of expert knowledge, we can inform a wider and more precise range of treatment options. Cognitive task analysis appears promising as an integrated, cost effective method which potentially could support the development of measurement instruments for performance, selection, and training purposes. To achieve this goal, future research and improvements in technology (e.g., for computer data collection and evaluation of performance protocols) will be required.

7.0 Evaluation of Cognitively-Oriented Test Design

To evaluate the efficacy of a cognitively-oriented approach to test design, we administered the job knowledge test and hands-on performance measures to a sample of Marines. Corresponding to the four primary methodological changes we implemented, we examined three hypotheses. First, we hypothesized that a cognitive approach to defining the knowledge actually used for performance would result in the identification of unique content compared to existing tests of land navigation. That is, we expect that cognitively-oriented methods for task and knowledge analyses would reveal land navigation goals and procedures that are under-represented or not represented at all in existing tests. Second, we proposed that the use of a cognitive model of the qualitative components of knowledge would result in a unique structure of knowledge in the representation and sampling of test content. By formally including knowledge components (i.e., principles, procedure selection, procedure execution, goal knowledge, and pattern recognition) in the knowledge representation and sampling of test content, we expect a cognitively-oriented test to more completely and

representatively sample these components of knowledge. Third, we hypothesized that basing the selection of test content on the diagnosticity rather than importance of the content would result in improved psychometric properties of the test. Fourth, we proposed that the use of probability-based inference network to score responses on the would improve the predictive power of the test. Finally, we hypothesized that these changes to test development methods would collectively result in an improved correspondence to hands-on measures of job performance. We now review the study and its results.

7.1 Sample. The sample consisted of 408 Marines stationed at Camp Pendleton, California. All study participants were trainees in one of four training schools at the base. The rank and level of experience of Marines varied between each training school, as is shown in Table 4. Usable data were obtained from 358 of these participants.

Table 4

DESCRIPTION OF THE SAMPLE

	School/Group				Total
Characteristics	MCT	ITB	AIT-SLS	AIT-PS	Sample
Number of Subjects	128	187	63	31	409
Average Age	19	19	22	30	21
Average Months in Service	4	5	36	138	22
Median Rank	PVT	PFC	LCPL	SSGT	PFC
Median No. of LandNav Classes	2	3	4	5	3

Notes:
MCT - Marine Combat Training, the first school after basic training.
ITB - Infantry Training Battalion, the first training after MCT for infantry MOS.
AIT-SLS - Advanced Infantry Training, Squad Leaders School.
AIT-PS - Advanced Infantry Training, Platoon Sergeants.

7.2 Measures

Job Knowledge. The written job knowledge test consisted of 100 multiple choice questions covering all areas of land navigation. Some test questions referred to one of three map segments provided to the examinees. Examinees also were provided with protractors. The job knowledge score consisted of the total number of correct responses, except as described below for scoring based on the use of a probability-based inference network.

Background and Experience. Land navigation training, experience, and demographic information was obtained using a three page form administered at the same time as the job knowledge test. This form assessed the type, recency, and frequency of land navigation skills, the types of terrains navigated and the amount of previous and related navigational experience. The background form consisted of rationally developed scales covering 6 aspects of experience: quality/variety, duration, frequency, recency, feedback, and previous/related experience. An overall experience variable was constructed by converting each scale to z-scores and summing across the 6 scales.

Hands-on Proficiency. Proficiencies in individual land navigation skills were assessed by administering five hands-on tests, one each for planning, location, distance, direction and movement. Each hands-on test consisted of 4-6 steps, with each step scored GO or NO-GO. Scoring consisted of the sum of correct (i.e., GO) responses over the 26 steps of the five tests.

Performance. The performance test consisted of locating 4 stakes in a 3 square mile area, given the map coordinates for each stake. There were a total of 30 stakes in the area, with 6 stakes used as possible starting points. Each examinee within a group was assigned a different route. The time to complete the route was recorded. The performance score consisted of summing the z-scores for the number of stakes correctly identified and the amount of time to complete the route.

Compared to hands-on tests, this test more closely resembles the actual task of land navigation required on the job. That is, it involves deciding which skills and strategies to employ for an objective, given the terrain, weather, and other conditions, as well as the execution and integration of those skills in order to successfully complete the task. Hence, we refer to this measure as a "work sample" measure of performance, in contrast to the previous measure which we denote as a "hands-on" measure of proficiency.

Effort , Test Attitudes, and Task Strategy. One page de-brief forms were administered following the job knowledge and performance tests to assess examinee effort, attitude toward the tests, and task strategies. The debrief form for the written test consisted of 10 Likert-type items assessing examinee' effort, fatigue, preparation, etc., based generally on Arvey, Strickland, Drauden, and Martin's (1990) motivational components of test taking. The

debrief form for the performance test consisted of 15 items: 3 assessed effort, 2 items asked for self reports of performance scores, and 10 items asked examinees to identify which strategies and procedures were used during performance.

Marine Corps Measures. In addition to the research-only measures, written and performance land navigation tests developed by the Marine Corps had been recently administered as part of training evaluation for the two groups from Advanced Infantry Training school. These test scores were made available for our research. The written test consisted of 15 multiple choice questions. There were four Marine performance tests, each involving locating stakes in a wilderness area of the base under a different set of conditions. One test consisted of locating 4 stakes during daylight with a map and a compass. A second daylight test involved use of a map without a compass to locate 3 stakes. The third test consisted of locating 3 stakes at night with a map and compass. The fourth performance test consisted of determining the grid coordinates of 3 locations using the methods of resection and intersection.

7.3 Data collection. Data collection was conducted at Camp Pendleton, California, over a three day period. The written tests were administered in two hours by one of the research staff. The background and experience form was administered first, then the job knowledge test, followed by the 1 page debrief from. Together, the written forms involved two hours of examinee time.

Hands-on testing was conducted by 16 non-commissioned officers. For the hands-on tests, stations were set-up in a large open field and were administered to a few (1-6) examinees at a time. Completion of the five hands-on tests required about two hours of examinee time.

The performance tests for 3 of the 4 groups were administered by a Marine land navigation instructor the week before data collection for the remaining tests. The test for the remaining group was administered during the data collection period under more difficult conditions. The base was officially closed the entire week of testing due to flooding, and the weather had some effect on testing conditions. Group 2 (ITB) completed the performance test on terrain that was soft and muddy—undoubtedly more so than for the preceding three groups. Further, test administrators (NCOs) were pulled in and out of testing to respond to various emergencies and duties created by the weather conditions. This probably reduced the reliability of administration and scoring for the performance test for group 2 (ITB) and the hands-on tests for all groups. The muddy conditions had two additional effects on group 2. It probably slowed the performance times for group 2. Second, it reduced the time available for completion of hands-on resulting in a relatively high proportion of missing data (i.e., completion of only 3 or 4 of the 5 hands-on tests). Hence, values for missing data on the hands-on tests for group 2 were imputed by using the mean of the total sample (i.e., all four groups) of each hands-on test.

7.4 Results. We hypothesized that a cognitively-oriented job knowledge test would differ from traditional tests with respect to knowledge content, and that these content differences would correspond to increases in correlations of job knowledge with performance. To examine differences in content between existing tests and our cognitively-oriented test, content analyses were conducted for the seven land navigation tests that could be located . The results of the content analyses are shown in Table 5. The contents of the land navigation tests were analyzed according to the two dimensions of our cognitive model of knowledge--substantive content and cognitive components. That is, each question from each test was sorted twice: first, into one of the five content categories of land navigation and second, into one of the six cognitive components.

Table 5

CONTENT ANALYSIS OF LAND NAVIGATION TESTS

Content Categories	Existing Land Navigation Tests							Cognitively-Oriented Test
	AIT	TBS	MCI	Army JPM	Marine JPM	EST Average		
Content								
Planning	0	0	10	0	0	0	2	17
Location	34	56	42	66	62	72	55	38
Distance	6	18	15	22	21	28	18	16
Direction	54	26	31	12	17	0	23	16
Movement	6	0	2	0	0	0	1	13
Total Percent	100	100	100	100	100	100	100	100
Components								
Principles/Concepts	7	0	6	0	0	0	2	8
Procedure Selection	0	0	3	0	0	0	1	12
Procedure Execution	33	55	25	67	67	46	49	32
Goal Knowledge	0	0	0	0	0	0	0	4
Pattern Recognition	0	42	35	33	29	18	26	38
Facts/Labels	60	3	31	0	4	36	22	6
Total Percent	100	100	100	100	100	100	100	100
Test Length	15	30	128	9	24	11	36	100

Note: Numbers represent percentage of test content.

By inspection, the cognitively-oriented job knowledge test clearly differs from existing tests. In terms of content areas, the cognitively-oriented test includes substantially more questions which assess planning and movement. In fact, 5 of the other 6 tests contain no questions addressing planning and 4 tests contain no questions addressing movement. With respect to the cognitive components, the cognitively-oriented test contains substantially more content assessing procedure selection, goal structure, principles/concepts, and fewer questions which assess declarative knowledge.

The second set of hypotheses involved the relationships of job knowledge to other performance related constructs and measures. Descriptive statistics for each measure are displayed in Table 6. Correlations above .10 and .14 are significant at $\alpha = .05$ and .01, respectively (for n = 359). The diagonal of the table displays estimates of internal consistency reliability. As shown in Table 6, job knowledge as conceptualized and measured by a cognitively-oriented approach, is significantly correlated with both hands-on proficiency and work sample measures of land navigation performance.

Table 6

CORRELATIONS BETWEEN JOB KNOWLEDGE, EXPERIENCE, PROFICIENCY, AND PERFORMANCE

	Mean	SD	Correlations							
			1	2	3	4	5	6	7	8
1 Job Knowledge	43	12.0	.87							
2 Strategy	2	.9	.21	—						
3 Experience	0	2.2	.50	.25	—					
4 Hands-on Proficiency	15	4.7	.58	.25	.44	.74				
5 Performance	2	1.3	.33	.23	.32	.31	.61	-		
6 Performance Composite	.05	1.6	.42	.31	.44	.40	.79	—		
7 Effort (Written Test)	3	.6	.32	.02	.38	.12	.05	.13	.76	
8 Effort (Performance Test)	7	1.8	-.02	-.08	.12	-.06	.26	.06	.19	—

Notes:

1. Reliabilities for the job knowledge and hands-on tests are alpha coefficients.
2. Performance test reliability was estimated by correlating the PDRI and AIT performance test scores, both of which consisted of a 4 stake landnav course.
3. The performance score is an equally weighted composite of the number of stakes found and the time it took to complete the test (time is reversed scored).
4. The correlations reported are uncorrected for statistical artifacts (e.g., unreliability).

To evaluate whether these results represent an improvement in job knowledge test design, two comparisons are available. First, these results can be compared directly to results obtained from using an existing land navigation test with a sub-sample of the same subjects. Second, comparisons can be made to previous results cited in the scientific literature. We first compare the results to an existing test of land navigation.

Table 7

CORRELATIONS BETWEEN JOB KNOWLEDGE AND PERFORMANCE:
A COMPARISON BETWEEN COGNITIVELY-ORIENTED AND EXISTING TESTS

	Performance Measures							
	Marine Corps				Research			Combined
	Day	Night	Terrain	Total	Day	Time	Total	Total
Total Tests								
Cognitive Test	0.26	0.37	0.32	0.51	0.33	-0.42	0.40	0.48
Marine Test	-0.20	0.31	0.34	0.15	-0.06	0.08	0.01	0.08
Content scales: Cognitively-oriented test								
1 Planning	0.46	-0.01	0.30	0.01	0.25	-0.06	-0.07	-0.07
2 Location	0.22	0.36	0.31	0.53	0.29	-0.39	0.47	0.55
3 Distance	0.39	0.25	0.24	0.34	0.20	-0.40	0.31	0.35
4 Direction	0.28	0.40	0.13	0.42	0.23	-0.39	0.28	0.35
5 Movement	0.33	0.40	0.29	0.57	0.28	-0.34	0.39	0.51
Component scales: Cognitively-oriented test								
A Principles	0.01	0.19	0.13	0.17	0.07	-0.31	0.23	0.21
B Selection	0.30	0.15	0.23	0.36	0.14	-0.15	0.12	0.22
C Execution	0.25	0.42	0.31	0.54	0.31	-0.43	0.39	0.49
D Goals	0.19	0.43	0.41	0.52	0.42	-0.27	0.42	0.52
E Patterns	0.20	0.22	0.26	0.35	0.18	-0.35	0.29	0.33
F Declarative	0.22	0.27	0.15	0.37	0.36	-0.23	0.36	0.40
Composite Tests								
Marine + 1, 5	0.34	0.40	0.34	0.59	0.36	-0.32	0.35	0.49
Marine + ABD	0.36	0.43	0.30	0.62	0.39	-0.36	0.42	0.55

Thirty-one subjects participated in both the Marine and the research land navigation tests. The results are displayed in Table 7. The performance measure for the "Combined Total" column was constructed by converting the Marine Total and the Research Total to z-scores, then summing the two. The first two rows of correlations provide a direct

comparison between the two tests. The cognitively-oriented test is consistently superior across all but one of the performance measures. The last two rows show the correlations resulting from adding scores from specific sub-scales of the cognitively-oriented test to the score from the Marine written test. The sub-scales added consist of content not typically found in existing land navigation tests (planning [1] and movement [5]; principles [A], procedure selection [B], and goal attainment [D]). No tests corresponding to the hands-on proficiency tests are routinely administered by the Marine Corps, and so were not available for comparison.

The primary objective of this study is to determine if tests of job knowledge could be sufficiently improved to replace the much more costly hands-on tests of proficiency. Thus, a comparison to correlations of hands-on tests with job knowledge tests developed using existing methods is of particular interest. A search of the literature produced only one study which contained correlations of job knowledge with hands-on proficiency. This study reports the results from the U.S. Army's Project A research on performance measurement.

These results are combined with the data from the present study and are summarized in Table 8. The results show the correspondence of job knowledge with hands-on proficiency measured at two levels of analysis. The first three rows show the average correlations of job knowledge tests with hands-on tests for individual job tasks (e.g., perform CPR). For the content-oriented tests (i.e., the Army data), the values represent the average correlation of job knowledge and hands-on test scores across 15 different tasks for each of the nine jobs. There were approximately 7 items (a range from about 3 to 15 items) for each knowledge test and about 10 to 15 items for each hands-on test (a range from about 5 to 50 items).

Table 8

SUMMARY OF CORRELATIONS BETWEEN JOB KNOWLEDGE AND HANDS-ON PROFICIENCY

	Sample Size	Number Of r's	Mean r	Mean ρ	SD r	SD ρ
Task Level						
Cognitively-oriented Test	358	1	.28	.36	–	–
Content-oriented Tests	1227	9	.16	.26	.043	.056
Combined Total	1585	10	.18	.27	.050	.063
Job Level						
Cognitively-oriented Test	358	1	.57	.68	–	–
Content-oriented Tests	1227	9	.38	.59	.149	.207
Combined Total	1585	10	.40	.60	.152	.198

The second three rows report correlations at the job level of analysis. These correlations represent the relationships between job knowledge and hands-on proficiency when scores are first summed across tasks within each job (15 hands-on tests are aggregated for the correlations shown here), and then correlated. The columns labelled ρ display the results for correlations corrected for unreliability in both variables. The corrections were made using split-half estimates of reliability for all tests shown. The Army job knowledge and hands-on tests both consist of about 100 items. The Marine land navigation knowledge test consists of 100 items and the combined hands-on tests contain 26 items.

Using the significance test proposed by Cohen (p. 175, 1988), the differences between correlations for the cognitively-oriented job knowledge test and the average content-oriented are significant. However, as noted by Hunter and Schmidt (1990), for the sample size and variation in observed correlations found here, such tests are prone to Type I errors. Much additional data, especially including constructive replications of the job knowledge/hands-on test correlations for cognitively-oriented tests, will be needed before a clear assessment of the potential improvement of cognitively-oriented tests can be made.

7.3.1 Analysis of task strategy. The first set of analyses were directed towards comparing the results of scores derived from the inference network to scores from adding the total number correct of job knowledge items. Correlations were computed between job knowledge, as scored by the inference network and job knowledge-total number correct; and both work-sample and hands-on measures of land navigation proficiency. These results are displayed in Table 9. The inference network job knowledge score was derived from the expected values of expert performance.

Table 9

CORRELATIONS BETWEEN ALTERNATIVE JOB KNOWLEDGE SCORES AND PERFORMANCE

Job Knowledge Measure	Work-Sample Performance	Hands-On Proficiency
Inference Network, Model 1	0.42	0.59
Total Number Correct	0.42	0.58

As the table shows, there were no real differences in the degree of correspondence of job knowledge with hands-on proficiency or work sample measures of performance whether scored by total number correct or the inference network.

To better understand the performance of the inference network, we performed a lens model analysis. This analysis decomposes the components of the correlation to identify whether the inference network was capturing different information than number correct scoring (e.g., configural response patterns or the use of different cues; Levi, 1985, 1989). The formula for this analysis was taken from Tucker (1964) and is as follows:

$$r_a = GR_cR_s + C\,(1 - R_c^2)^{1/2}\,(1 - R_s^2)^{1/2}$$

where

r_a = correlation of forecasts and outcomes
$\hat{Y}_c$, $\hat{Y}_s$ = predicted values from a linear regression of outcomes and forecasts on cues, respectively
G = correlation of $\hat{Y}_c$ and $\hat{Y}_s$
R_c = multiple correlation of cues and performance outcomes
R_s = multiple correlation of cues and forecasts
C = correlation of residuals corresponding to $\hat{Y}_c$ and $\hat{Y}_s$

In order for the inference network to do better than a linear model (i.e., $r_a > R_c$), the *C component must be large and positive. C is a measure of the predictable variance* <u>not</u> contained in the linear model. Thus, C can be large and positive if the network captures valid configural relationships among the given cues or makes use of cues not available in the regression. Further, the network will be successful in predicting performance to the extent that the network predictions also capture the linear components of the cues (i.e., as R_s and G approach 1).

To summarize, there are three multiple regression components of interest. First, the nonlinear component of the forecast must be significant in order for the network model to outpredict a regression model. Second, given C, the network model will do well to the extent that there is a large linear component predicted by the network model, R_s. That is, the network model must be able to capture the valid linear variance as well as the nonlinear variance. And third, the network will perform well to the extent that its modelling of the linear component of the cues agrees with the linear regression model (i.e., G is close to 1).

The values for C, R_s, and G for the three network models are presented in Table 10. The first network model, "Performance, Model 1", was constructed to model the effects of strategy and knowledge on performance. This model is the one depicted in Figure 5. The second inference network, "Performance, Model 2", disaggregates knowledge into the specific elements (i.e., test questions) related to each proficiency (i.e., each hands-on step) and then related to performance. This same network is also employed to determine an

expected value for hands-on proficiency (i.e., the "Proficiency" model). The results indicate that the inference networks capture a similar linear component of the cues as do regression models (i.e., G approaches 1.0), but only the "hands-on" network captures unique, possibly configural, information in comparison to regression models (i.e., $C > 0$).

TABLE 10

A LENS MODEL ANALYSIS OF REGRESSION COMPONENTS

Inference Network Models	Regression Components				
	C	G	Ra	Rs	Re
Proficiency, Hands-on	0.353	0.967	0.586	0.835	0.523
Performance, Model 1	0.064	0.938	0.418	0.833	0.496
Performance, Model 2	0.095	0.990	0.424	0.801	0.474

Model 1 = Aggregated network with 4 knowledge scales.
Model 2 = Disaggregated network relating knowledge items to hands-on steps.

7.5 Discussion. The data gathered by this study provide preliminary support for the utility of cognitive methods for improving job knowledge tests and their correspondence to other measures of performance. A basic premise of this study was that improvements in job knowledge testing would result from a more complete and relevant assessment of this construct. The results of the content analyses of existing tests show that the cognitive task and knowledge analyses identified unique domain content. Importantly, this content was shown to significantly contribute to the correspondence of job knowledge with hands-on proficiency tests. Additionally, the correlation of a cognitively-oriented job knowledge test with hands-on performance was higher than similar relationships found in previous research.

At the job level of analysis, this study may represent an underestimate of the relationship between job knowledge and hands-on proficiency. Land navigation is one of dozens of tasks that Marines are responsible for knowing, even though most Marines rarely, if ever, perform this task on the job. Consequently, the reliability and stability of measurement for an isolated, infrequently performed task is probably less than that found across the aggregate of essential, frequently performed job duties. Assessing job knowledge and performance across most or all important job duties may further increase the relationship obtained in this study. Thus, firm conclusions about the efficacy of cognitively-oriented

approaches to job knowledge test design must wait for research which assesses the generality of the methods and results to entire jobs and to different types of performance domains.

The contributions of modelling job knowledge via a probability-based inference network are less straightforward, thus requiring additional discussion. There are two potential advantages of an inference network approach to modelling knowledge representations and their relationships to performance. First, it could lead to improved prediction of performance by better capturing configural components of performance. Second, it could provide useful diagnostic information for a broader array of organizational decisions, such as training as well as selection and placement.

With respect to improvements in prediction, these initial results indicate that the use of probability-based inference networks provide no significant gains, at least for this measurement context (i.e., Marine land navigation). There are several possible explanations for these results which future research will need to address. The explanations fall into two categories: limitations with respect to measurement and to theory.

Limitations of Measurement. For example, the criterion measure sampled performance in only one type of terrain. However, the only configural patterns predicted for performance involved a terrain by strategy by knowledge interaction and a strategy by knowledge interaction. Hence, one explanation is that there was insufficient configurality in the criterion performance to be predicted relative to on-the-job performance. According to this hypothesis, the usefulness of the inference network would be evident if greater configurality in performance had been present as a result of assessing performance in two or more different terrain types (desert, mountains, hilly forests, etc.). Unfortunately, this was not possible due to cost considerations. Additionally, potential configurality in the criterion was further reduced by the moderately strong linear relationships between strategy, knowledge, and performance.

A second explanation for the lack of improved accuracy is that the test scales designed to assess knowledge relevant to successful performance of each strategy possessed insufficient construct or discriminant validity. Insufficient construct validity would result from a deficient or contaminated operationalization of the knowledge related to the execution of each strategy. For example, from the task analysis it was apparent that pattern recognition is a key component of terrain strategy execution. If the test questions assessing pattern recognition were insufficiently sensitive to differences in examinees' pattern recognition knowledge, then no strategy by knowledge interactions would be detected. Consequently, a reduction in construct and/or discriminant validity between the knowledge scales would reduce any advantage of configural scoring methods.

Limitations of Theory. This finding could also result from an incorrect theory about how knowledge affects performance. For example, if those who select a particular strategy

have similar levels of knowledge in how to execute the strategy, then the correlation between that knowledge scale and performance would be smaller rather than larger for the subgroup using that strategy compared to the corresponding correlation across all groups (i.e., including those who used other strategies). Some support for this hypothesis can be derived from an examination of the correlations of the knowledge scales with performance within each performance strategy. Contrary to expectations, the compass knowledge scale had the lowest correlation with performance within the group of examinees who used the compass strategy. Similarly, the terrain knowledge scale produced the smallest correlation with performance within the group of examinees who used the terrain strategy. This finding could be a result of a restriction in range for these knowledge scales within the groups using the corresponding strategy. That is, if only those selecting a strategy were similarly proficient in that strategy, then restriction of range and a corresponding decrease in the correlation would result. This would in turn reduce the size of a strategy by knowledge interaction.

Irrespective of contributions to prediction, the use of inference networks appear to offer some advantages for diagnosis. By modelling individual differences in performance processes, such as strategies, choice of procedural tactics, etc., this more detailed information can be very useful for providing feedback on alternative strategies and tutoring to improve performance. Additionally, the flexibility of inference networks for incorporating various sources and types of information, such as the type of incorrect responses to questions, makes it a potentially useful tool for diagnosing training or performance problems.

8.0 Conclusions

The evidence provided by this study supports the utility of cognitive methods for improving job knowledge tests. These preliminary findings indicate that cognitively-oriented test design increases the correspondence of job knowledge to hands-on and work-sample measures of performance. Of the several techniques adapted from cognitive science methods, those related to identifying unique domain content appear to have contributed most towards improving the predictive accuracy and interpretability of the job knowledge test. Specifically, the use of knowledge elicitation techniques such as verbal protocol analysis and coaching, and representing the content in terms of a plan-goal graph contributed to the development of test questions that more closely correspond to actual performance. Furthermore, the incorporation of cognitively-oriented methods did not require substantially increased resources.

Other cognitive science methods, such as the use of a probability-based inference network to score the test presented potential advantages in diagnostic capabilities with no apparent loss to predictive efficiency. The capabilities to incorporate information from incorrect responses and to model configural patterns of knowledge-performance relationships offer the promise of substantial improvements in the integration of predictive and diagnostic

uses of performance measurement. Additional research is needed to explore these interesting possibilities.

The generalizability of this research is limited by several factors. The study focused on a single task of the dozens performed within jobs in the Marine Corps. The results may not generalize to other areas of performance. On the other hand, estimates of the correspondence of job knowledge and performance may increase when cognitively-oriented methods are applied across all tasks of a job, due to increases in the stability of performance across tasks. Replication of this work in other jobs is warranted to confirm these results and to more precisely quantify the improvement in correspondence to other measures of performance. Firm conclusions about the utility of cognitively-oriented approaches to job knowledge test design must wait for research which assesses the generality of the methods and results to entire jobs and to different types of performance domains.

References

Andreassen, S., Woldbye, M., Falck, & Andersen, S.K. (1987). MUNIN: A causal probabilistic network for interpretation of electromyographic findings. *Proceedings of the 10th International Joint Conference on Artificial Intelligence.*

Anderson, J. A. (1990). Hybrid computation in cognitive science: Neural networks and symbols. Applied Cognitive Psychology, 4, 337-347.

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review, 89*, 369-406.

Anderson, J.R. (1983). *The architecture of cognition.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Arvey, R. D., Strickland, W., Drauden, G., & Martin, C. (1990). Motivational components of test taking. Personnel Psychology, 43, 695-716.

Béland, A. & Mislevy, R.J. (1992). Probability-based inference in a domain of proportional reasoning tasks. Princeton, NJ: Educational Testing Service.

Bobrow, D. G., & Winograd, T. (1977). An overview of KRL, a knowledge representation language. Cognitive Science, 1, 3-46.

Borman, W.C., Hanson, M.A., Oppler, S.H., Pulakos, E.D. & White, L.A. (in press). The role of early supervisory experience in supervisor performance. *Journal of Applied Psychology.*

Borman, W.C., White, L. A., Pulakos, E. D., & Oppler, S. H. (1991). Models evaluating the effects of ratee ability, knowledge, proficiency, temperament, awards, and problem behavior on supervisor ratings. *Journal of Applied Psychology, 76*, 863-872.

Borman, W.C., Pulakos, E.D., & Oppler, S.H. (1992). *Models evaluating the effects of rater characteristics on peer job performance ratings.* Paper presented as part of a symposium, "Modeling and Operationalizing Implicit Performance Theories", Joy Hazucha, Chair, Society of Industrial-Organizational Psychology, Montreal.

Breuker, J. & Weilinga, F. (1987). Use of models in the interpretation of verbal data. In A. L. Kidd (Ed.), *Knowledge Acquisition for Expert Systems: A Practical Handbook.* New York: Plenum Press.

Brown, J., & Van Lehn, K. (1980). Repair theory: A generative theory of bugs in procedural skills. *Cognitive Science, 4*, 379-426.

Carrier, M. R., Dalessio, A. T., & Brown, S. H. (1990). Correspondence between estimates of content and criterion-related validity values. *Personnel Psychology, 43,* 85-100.

Charniak, E. (1991). Bayesian networks without tears. *AI Magazine*, Winter, 50-63.

Charniak, E. & Goldman, (1991). A probabilistic model of plan recognition. In *Proceedings of the Ninth National Conference on Artificial Intelligence*, 160-165. Menlo Park, CA: American Association for Artificial Intelligence.

Chi, M.T.H., Feltovich, P., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. Cognitive Science, 5, 121-152.

Conley, P. R. & Sackett, P. R. (1987). Effects of using high versus low performing job incumbents as sources of job analysis information. *Journal of Applied Psychology, 72*, 3, 434-437.

Dunnette, M. D. (1963). A note on *the* criterion. *Journal of Applied Psychology, 47*, 251-254.

Ericsson, K. A. & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data.* Cambridge, MA: MIT Press.

Ericsson, K. A. & Simon, H. A. (1986). Verbal reports as data. *Psychological Review, 87*, 215-251.

Flanagan, J. C. (1954). The critical incident technique. *Psychological Bulletin, 51*, 327-358.

Geddes, N.D. (1989). *Understanding human operator's intentions in complex systems.* Unpublished doctoral dissertation. Georgia Institute of Technology, Atlanta, GA.

Gelman, R. & Gallistel, R. (1978). *The child's understanding of number.* Cambridge, MA: Harvard University Press.

Greeno, J. G. (1989). A perspective on thinking. *American Psychologist, 44,* 134-141.

Greeno, J. G. & Simon, H. (1988). Problem solving and reasoning. In R.C. Atkinson, R.J. Hernstein, G. Lindzey, & R.D. Luce (Eds.) Stevens' *Handbook of Experimental Psychology*.

Greeno, J.G., Riley, M. S. & Gelman, R. (1984). Conceptual competence and young children's counting. Cognitive Psychology, 16, 94-143.

Guion, (1978). *Principles of work sample testing: III. Construction and evaluation of work sample tests* (TR-79-A10). Alexandria, VA: US Army Research Institute for the Social and Behavioral Sciences.

Hanson, M., & Borman, W. C. (1989). *Development and construct validation of a situational judgment test as a performance measure for first line supervisors.* Symposium presentation at the 4th Annual Convention of the Society for Industrial/Organizational Psychology, Boston, MA.

Hanson, M., DuBois, D., Johnson, S., Carter, G., & Peterson, N. (1988). *A catalogue of alternatives to hands-on job performance measurement.* (Report No. 163) Minneapolis: Personnel Decisions Research Institute.

Hansson, O., & Mayer, A. (1989). Heuristic search as evidential reasoning. In *Proceedings of the Fifth Workshop on Uncertainty in Artificial Intelligence*, 152-161. Mountain View, CA: Association for Uncertainty in Artificial Intelligence.

Hayes-Roth, F., Waterman, D.A., & Lenat, D. B. (eds.) (1983). Building expert systems. Reading, MA: Addison-Wesley.

Heckerman, D. (1990). *Probabilistic similarity networks* (Technical Report STAN-CS-1316). Palo Alto, CA: Stanford University, Departments of Computer Science and Medicine.

Hinton, G. E., & Anderson, J.A. (1981). Parallel models of associative memory. Hillsdale, NJ: Lawrence Erlbaum Press.

Hunter, J. E. (1983). A causal analysis of cognitive ability, job knowledge, job performance, and supervisor ratings. In F. Landy, S. Zedeck, & J. Cleveland (Eds.), *Performance Measurement and Theory*. Hillsdale, NJ: Lawrence Erlbaum & Associates.

Jörskog, K.G. & Sörbom, D. (1989). *LISREL 7: User's Reference Guide.* Mooresville, IN: Scientific Software, Inc.

Kieras, D.E. (1990). *The role of cognitive simulation models in the development of advanced training and testing systems.* In N. Frederiksen, R. Glaser, A. Lesgold and M.G. Shafto (Eds.). Diagnostic modeling of skill and knowledge acquisition. Lawrence Erlbaum Associates, NJ.

Kosslyn, S.M., Flynn, R.A., Amsterdam, J.B. & Wang, G. (1990). Components of high-level vision: A cognitive neuroscience analysis and accounts of neurological syndromes. *Cognition*, 34, 203-277.

Lammlein, S. E. (1986). *Proposal and evaluation of a model for job knowledge testing.* Unpublished doctoral dissertation, University of Minnesota, Minneapolis, Minnesota.

Lauritzen, S.L., & Spiegelhalter, D.J. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society, Series B, 50*, 157-224.

Lazarfeld, P.F. (1950). The logical and mathematical foundation of latent structure analysis. In S.A. Stouffer, L. Guttman, E.A. Suchman, P.F. Lazarfeld, S.A. Star, & J.A. Clausen, *Studies in Social Psychology in World War II, Volume 4: Measurement and Prediction.* Princeton, NJ: Princeton University Press.

Levi, K.R. (1989). Expert systems should be more accurate than human experts: Evaluation procedures from human judgment and decision-making. *IEEE Transactions on Systems, Man, and Cybernetics, 19*, 3, 647-656.

Levi, K.R. (1985). Likelihood judgments from physicians and linear models. Doctoral dissertation, Department of Psychology, University of Michigan.

Levitt, T., Mullin, J., & Binford, T. (1989). Model-based influence diagrams for machine vision. In *Proceedings of the Fifth Workshop on Uncertainty in Artificial Intelligence*, 233-244. Mountain View, CA: Association for Uncertainty in Artificial Intelligence.

Lin, Z. & Carley, K.M. (1992). *Maydays and Murphies: A study of the effect of organizational design, task, and stress on organizational performance* (Technical Report No. UPITT/LRDC/ONR-URI-HGD-2). Pittsburgh, PA: Learning Research and Development Center.

Masters, G., & Mislevy, R.J. (1991). *New views of student learning: Implications for educational measurement* (RR-91-24-ONR). Princeton, NJ: Educational Testing Service.

McCloy, R.A., Campbell, J.P., & Cudeck, R. (1992). A confirmatory test of a model of

performance determinants. Unpublished manuscript.

McCloskey, M., Carmazza, A., & Green, B. (1980). Curvilinear motion in the absence of external forces: naive beliefs about the motion of objects, Science, 20, 1130-1141.

Mislevy, R.J. (1992). *Foundations of a new test theory* (RR-89-52-ONR). Princeton, NJ: Educational Testing Service.

Mislevy, R.J. (1990). Modeling item responses when different subjects follow different solution strategies. *Psychometrika, 55*, 195-215.

Mislevy, R.J., Yamamoto, K. & Anacker, S. *Toward a test theory for assessing student understanding* (RR-91-32-ONR). Princeton, NJ: Educational Testing Service.

Newell, A. (1991). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

Newell, A. (1982). The knowledge level. *Artificial Intelligence, 18*, 87-127.

Newell, A., & Simon, H. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

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

Orasanu, J. (1990). Shared mental models and crew decision making (CSL Report 46). Cognitive Science Laboratory, Princeton University, Princeton, NJ.

Orasanu, J. & Fischer, U. (1992). *Team cognition in the cockpit: Linguistic control of shared problem solving*. In Proceedings of the 14th Annual Meeting of the Cognitive Science Society. Hillsdale, NJ: Erlbaum.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. San Matco, CA: Kaufmann.

Reitman, W.R. (1965). Cognition and thought: An information processing approach. New York, NY: Wiley.

Rouse, W.B., Geddes, N.D. & Hammer, J.M. (1990). Computer aided fighter pilots. *IEEE Spectrum*, March, 388-41.

Rumelhart, D.E., McClelland, J.L., & the PDP Research Group. (1986). Parallel distributed processing: Explorations in the microstructure of cognition (Vols. I & II). Cambridge, MA: MIT Press.

Sacerdoti, E.D. (1977). A structure for plans and behavior. New York, NY: Elsevier-North Holland.

Schank, R.C. and Abelson, R.P. (1977). *Scripts, plans, goals and understanding.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Schmidt, F. L., Hunter, J. E., & Outerbridge, A. N. (1986). Impact of job experience and ability on job knowledge, work sample performance, and supervisory ratings of job performance. *Journal of Applied Psychology, 71*, 432-439.

Schmidt, F. L., & Kaplan, L. B. (1971). Composite versus multiple criteria: A review and resolution of the controversy. *Personnel Psychology, 24,* 419-434.

Sewell, D.R. & Geddes, N.D. (1990). A plan and goal based method for computer-human system design. *Human Computer Interaction*, INTERACT 90, New York: North-Holland, 283-288.

Shalin, V. L., Sczepkowski, M., & Bertram, D. (1993). *Physician expertise and workload in the medical intensive care unit.* Unpublished manuscript. Buffalo, NY: Department of Industrial Engineering, SUNY at Buffalo.

Smith, P.C., & Kendall, L.M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology*, 47, 149-155.

Smith, D.A., Greeno, J. G. & Vitolo, T.M. (1989). A model of competence for counting. Cognitive Science, 13(2), 183-212.

Spearman, C. (1907). Demonstration of formulae for true measurement of correlation. *American Journal of Psychology*, 18, 161-169.

Spiegelhalter, D., Franklin, R., & Bull, K. (1989). Assessment criticism and improvement of imprecise subjective probabilities for a medical expert system. In *Proceedings of the Fifth Workshop on Uncertainty in Artificial Intelligence*, 345-342. Mountain View, CA: Association for Uncertainty in Artificial Intelligence.

Spirtes, P. Scheines, R. & Glymour, C. (1990). Simulation studies of the reliability of computer-aided model specification using the Tetrad II, EQS, and LISREL programs. *Sociological Methods and Research, 19*, 3-66.

Sternberg, R.J. (1985). *Beyond IQ: A triarchic theory of human intelligence*. New York: Cambridge University Press.

Tucker, L.R. (1964). A suggested alternative formulation in the developments by Hursch, Hammond, and Hursch, and by Hammond, Hursch and Todd. *Psychological Review*, 71, 528-530.

Uttal, B.R., Bradshaw, G., Dayanand, S., Lovell, R., Shepherd, T., Kakarala, R. Skifsted, K., Tupper, G. (1992). THE SWIMMER: An integrated computational model of a perceptual motor system. Hillsdale, NJ: Lawrence Erlbaum Associates.

Van Lehn, K. (1983). On the representation of procedures in repair theory. In H. P. Ginsburg (Ed.), *The Development of Mathematical Thinking*. New York : Academic Press.

Van Lehn, K., Jones, R.M. & Chi, M.T.H. (1991). Modeling the self-explanation effect. *The Journal of the Learning Sciences, 2*(1), 1-59.

Wagner, R. K. & Sternberg, R. J. (1985). Practical intelligence in real world pursuits: The role of tacit knowledge. *Journal of Personality and Social Psychology*, 49, 436-458.

Wexley, K. N. & Silverman, S. B. (1978). An examination of differences between managerial effectiveness and response patterns on a structured job analysis questionnaire. *Journal of Applied Psychology, 63*, 5, 646-649.

Wertheimer, M. (1945). *Productive thinking*. New York, NY: Harper & Row.

Wright, S. (1934). The method of path coefficients. *Annals of Mathematical Statistics, 5*, 161-215.

Young, R.M. & O'Shea, T. (1981). Errors in children's subtraction. Cognitive Science, 5, 153-177.