MICROCOPY RESOLUTION TEST CHART

NATIONAL BUREAU OF STANDARDS-1963-A

⑫

AIR FORCE

HUMAN RESOURCES

LABORATORY

AD-A164 083

DTIC
SELECTED
FEB 1 2 1986
S D

THEORY-BASED COGNITIVE ASSESSMENT

Patrick C. Kyllonen

MANPOWER AND PERSONNEL DIVISION
Brooks Air Force Base, Texas 78235-5601

January 1986

Approved for public release; distribution unlimited.

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235-5601

NOTICE

AD-A164083

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION / AVAILABILITY OF REPORT Approved for public release; distribution unlimited.
2b. DECLASSIFICATION / DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) AFHRL-TP-85-30	5. MONITORING ORGANIZATION REPORT NUMBER(S)

6a. NAME OF PERFORMING ORGANIZATION Manpower and Personnel Division	6b. OFFICE SYMBOL (If applicable) AFHRL/MOE	7a. NAME OF MONITORING ORGANIZATION
6c. ADDRESS (City, State, and ZIP Code) Air Force Human Resources Laboratory Brooks Air Force Base, Texas 78235-5601		7b. ADDRESS (City, State, and ZIP Code)

8a. NAME OF FUNDING / SPONSORING ORGANIZATION Air Force Human Resources Laboratory	8b. OFFICE SYMBOL (If applicable) HQ AFHRL	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER

8c. ADDRESS (City, State, and ZIP Code) Brooks Air Force Base, Texas 78235-5601	10. SOURCE OF FUNDING NUMBERS			
	PROGRAM ELEMENT NO. 61102F	PROJECT NO. 2313	TASK NO. T1	WORK UNIT ACCESSION NO. 33

11. TITLE (Include Security Classification)

Theory-Based Cognitive Assessment

12. PERSONAL AUTHOR(S)
Kyllonen, Patrick C.

13a. TYPE OF REPORT Interim	13b. TIME COVERED FROM Oct 84 TO Sep 85	14. DATE OF REPORT (Year, Month, Day) January 1986	15. PAGE COUNT 36

16. SUPPLEMENTARY NOTATION Also to appear in J. Zeidner (Ed.), Human productivity enhancement: Vol. 2. Acquisition and development of personnel. New York: Praeger.

17.	COSATI CODES		18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	cognition information processing
05	10		computerized testing learning
05	09		individual differences

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

This paper reviews the theoretical foundations, current issues, and a number of recent studies in the area of cognitive ability assessment, with particular focus on personnel selection and classification applications. A brief review of the history of aptitude testing is provided, with particular emphasis on the evaluation of the testing program within the military services. It is argued that new developments in cognitive psychology, along with the availability of modestly priced microcomputer systems, hold promise for improvements in ability assessment technology. Recent studies conducted at the Air Force Human Resources Laboratory as part of the Learning Abilities Measurement Program and similar projects are reported. The focus of these studies has been on determining (a) the utility of various cognitive tasks for providing meaningful information on individual cognitive skill levels, (b) whether such tasks tap abilities not measured by currently operational paper-and-pencil tests, and (c) how changes in processing efficiency can be analyzed. Cognitive diagnostic as well as selection-and-classification applications are discussed.

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT [X] UNCLASSIFIED/UNLIMITED [ ] SAME AS RPT. [ ] DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION Unclassified	
22a. NAME OF RESPONSIBLE INDIVIDUAL Nancy A. Perrigo, Chief, STINFO Office	22b. TELEPHONE (Include Area Code) (512) 536-3877	22c. OFFICE SYMBOL AFHRL/TSR

DD FORM 1473, 84 MAR   83 APR edition may be used until exhausted.
All other editions are obsolete.

THEORY-BASED COGNITIVE ASSESSMENT

Patrick C. Kyllonen

MANPOWER AND PERSONNEL DIVISION
Brooks Air Force Base, Texas 78235-5601

SUMMARY

The purpose of this paper was to review the foundations and current developments of psychological testing in the military and in other settings. In the first part of the paper, the issue of the value of psychological tests is addressed by reviewing a number of traditional validation studies. It is concluded that although tests may be quite useful in predicting standard outcome criteria, there is a need for developing new tests rooted in cognitive theory, and for developing richer validation data. In the second part of the paper, a number of recent studies that have employed testing methods based on cognitive psychology are reviewed. These studies suggest a number of areas in which cognitive psychology may contribute to the development of a new approach to ability testing. Such an approach should lead to a broader, more comprehensive system of ability assessment for personnel selection and classification purposes. A cognitive approach to ability assessment also suggests many possibilities for diagnosing particular learner deficiencies.

Accesion For		
NTIS CRA&I	✓	
DTIC TAB	☐	
Unannounced	☐	
Justification		
By		
Dist. ibution /		
Availability Codes		
Dist	Avail a.id / or Special	
A-1		

PREFACE

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

THEORY-BASED COGNITIVE ASSESSMENT

I. INTRODUCTION

In many large organizations, psychological tests are routinely administered to job applicants to obtain information about their likelihood of succeeding on the job or in prerequisite training. For historical and economic reasons, such tests typically consist of a variety of multiple-choice items administered in the paper-and-pencil format. Items are designed to probe the applicant's ability to reason logically, to think quantitatively, to comprehend verbal material and so forth. Traditionally, these abilities and other skills have been called aptitudes, reflecting the philosophical stance that such skills are properly characterized as personal traits, which are relatively impervious to training.

In recent years, many psychologists concerned with individual differences have begun questioning the foundations on which this kind of testing technology has developed. In particular, the model of intellectual functioning that characterizes individuals in terms of a small, finite set of relatively stable traits is fast giving way to a whole new conception of intellectual functioning based on the view of the person as information-processor. This reorientation has given rise to a "new look" in ability measurement that promises to overhaul the conventional way in which psychological tests are administered, along with the ways in which performance on such tests is interpreted. The new look substantially builds on traditional factor-analytic-based accounts of individual differences and borrows heavily from the discipline of experimental cognitive psychology. As such, the new research approach to individual differences represents the beginnings of a convergence between what Cronbach (1957) called the "two disciplines of scientific psychology": the experimental and the correlational. It seems quite likely at this time that the experimental-cognitive approach to individual differences will lead to the generation of a whole new system of psychometrics, based on experimentally studied process models instead of the traditional trait models. This paper presents a detailed account of what the new information processing approach is likely to bring in the way of new psychological tests and new ways of using test score data.

Before the new approaches to cognitive assessment are discussed, it is useful to review the role and status of a psychological testing program in an organization in general. Human ability assessment can play a broad and critical role in enhancing human productivity and thereby increasing organizational effectiveness. Consider four obvious ways in which any large organization can enhance the productivity of its workforce. The first, and often most expensive, is training. It has been estimated that organizations spend up to 10% of their total payroll on training (Gilbert, 1976), which can amount to a substantial sum. Training and assessment are integrally intertwined. An organization trains those most likely to benefit from training, and day-to-day decisions about what to train next depend on an assessment of what the student or trainee knows right now. The second, and in many ways the most elusive means of productivity enhancement, is to increase employees' motivation levels. The capability for assessing an individual's current motivation level is an obvious prerequisite to evaluating any policy designed to enhance motivation. The third method for productivity enhancement is to design the systems with which operators interact in such a way as to optimize the efficiency of the man-machine interaction. Witness the public attention given to conditions faced by air-traffic control operators, and note that many follow-up studies have been concerned with the redesign of systems in order to accommodate human factors (e.g., Hopkin, 1982). Improved human factoring of the workplace depends to a large degree on adequate assessment methods. Finally, organizational productivity can be enhanced with an appropriate selection and classification system.

Many managers do not fully appreciate the importance of initial personnel selection and classification decisions. Numerous research studies attest to the wide variation in the learning

and performance capabilities of individuals (Rimland & Larson, in press). Data collected at the Air Force Human Resources Laboratory (AFHRL) and elsewhere have shown that some individuals can acquire skills and knowledge 10 to 20 times faster than others (Payne & Tirre, 1984). Managers often incorrectly conclude that on-the-job performance deficiencies are due to motivation and training deficiencies when an equally plausible case could be made for identifying the source of the problem as selection and classification errors. It is highly unlikely that training can always overcome a serious talent deficiency; in fact, in high demand areas, individual differences more often than not are magnified by training and experience (Cronbach & Snow, 1977).

Unfortunately, present personnel measurement tests are not highly accurate in identifying before the fact who will be the fast and slow learners. Present tests do not measure many of the abilities required for acquiring skills demanded by diverse occupations. The importance of more accurate measurement of learning abilities is elevated by forecasts of manning problems in the next decade. The number of 18- to 21-year-olds in the national manpower pool will decrease by about 20% in the near future and will remain at this low level through the 1990s. At the same time, competition between the military and civilian sectors for these scarce manpower resources is expected to increase as a function of an improving and expanding economy. Clearly the importance of selecting the best people as a means of productivity enhancement will become increasingly important.

The purpose of this paper is to discuss the role that cognitive assessment techniques can play in enhancing human productivity. Although the most obvious role for assessment methods is in the area of personnel selection and classification, it can be argued that improved techniques can serve all four areas: training, motivation, human engineering, and selection and classification. The following section begins with a brief review of the history of aptitude testing, with particular emphasis on the evolution of the testing program within the military services. The section is concluded by pointing out that in recent years conventional notions of aptitude have come under attack from both within and outside the field, but that recent theoretical developments in cognitive psychology, coupled with the now almost ubiquitous microcomputer, promise to change the nature of ability testing and provide it with a firmer theoretical foundation.

It is important, in discussing a new theory-based approach to assessment, to provide at least the glimmerings of the theory that serves as the base. Thus, in the third section, a description of the human as an information processing system is outlined and how such a view might serve useful as a foundation for new cognitive-based testing research is discussed. This is followed with a description of new assessment techniques rooted in cognitive theory, and some of the studies that have attempted to determine the utility of these new measurement methods and approaches are reviewed. The studies can be divided naturally into clusters based on the focus of the investigation. One set of studies is concerned with the question of whether elementary cognitive tasks can supplement or even replace conventional tests. A second set of studies is concerned with how the ability to learn can be measured. A third set of studies addresses the issue of whether complex cognitive skills such as reading can be broken down into more elementary skills. In these study review sections, particular attention is given to studies that have been conducted as part of the AFHRL Learning Abilities Measurement Program (Project LAMP). Throughout this paper, the numerous possible applications of new methods for cognitive assessment are discussed, beyond the obvious ones in selection and classification contexts. These include applications in remedial diagnosis, the development of training systems, and the design of systems to accommodate human factors. Finally, in a summary section, these developments are reviewed and the cost effectiveness of some of the new forms of cognitive assessment is discussed.

2

II. HISTORY OF COGNITIVE ASSESSMENT

Up until the mid-1970s, the predominant form of theorizing about cognitive abilities employed factor-analytic concepts and methods. Perhaps the first theory of cognitive ability was the two-factor theory proposed by Spearman (1905), who noted that correlation matrices of cognitive test scores exhibited what he termed "positive manifold," a condition characterized by the absence of zero or negative correlations. That is, if a person outperforms others on one cognitive test, that person is likely to do better than others on any other cognitive test. Spearman attempted to explain this phenomenon by proposing the concept of "general ability" which itself was defined as the level of "mental energy" available to the person. Spearman assumed that differences in mental energy level were responsible both for differences in success in schooling and success on the cognitive tests. Later, Thurstone (1938) modified the Spearman proposal by postulating seven relatively independent primary mental abilities such as verbal, spatial, deductive reasoning, and memory abilities. Thurstone contended that to predict a person's score on any cognitive test, it was not sufficient to know that person's general ability level because different kinds of abilities played different roles on various tests. More formally, Thurstone proposed that relative standing on one test could be predicted from relative standing on other tests by the common factor equation:

$$y_{ij} = w_{j1}F_{1i} + w_{j2}F_{2i} + w_{j3}F_{3i} + \ldots + w_{jk}F_{ki} + S_{ij} + e_{ij}$$

In this equation, y_{ij} is relative standing of the ith person on the jth test; the F's represent the relative level of ability of the person on the 1..k abilities, and the w_j's represent the relative importance of each ability in predicting relative standing on the jth test. In effect, Thurstone was advancing a kind of "mental chemistry" theory of learning and cognition, in which any learning or performance activity could be characterized by an ability requirements (or importance) profile, and any person could be characterized by an ability level profile. The importance of Thurstone's system for classifying people and tasks was realized by military psychologists, and it might not be too unfair to assert that the form of even present-day selection and classification systems in industry and education, as well as in the military services, is a fairly direct result of Spearman's and Thurstone's contributions.

Consider, for example, Guilford and Lacey's (1946) monumental work, Printed Classification Tests, which is widely credited with establishing the groundwork for virtually all subsequent military selection and classification testing (Weeks, Mullins, & Vitola, 1975). In that report, the authors divide the presentation into a more-or-less subjective task analysis of aircrew operators' jobs followed by an evaluation of tests of a number of general and specific abilities. According to Guilford, in his Preface, one of the key features of the work was "...the inclusion of analysis of job criteria by the factorial methods. It is believed that in this direction lies an economical, systematic, and dependable procedure for coverage of aptitudes and for fitting tests to vocations" (p. iii). That is, the utility of personnel classification tests was made obvious by the inclusion in the basic factor equation of both abilities tapped by tests and abilities demonstrated in training or in jobs.

In one of the earliest validation studies following Guilford and Lacey's report, Dailey (1948) emphasized the Thurstonian underpinnings of military testing by declaring that "...a fundamental postulate has been that each airman specialty requires a different combination of specific aptitudes for success. A further postulate is that basic airmen entering the Air Force have greatly different patterns of the specific aptitudes essential to success in various specialties," and further stating that in developing tests "...heavy emphasis is placed upon the techniques of factor analysis" (p. 1). Since that time, a similar theoretical rationale is routinely stated in the introductory remarks of validation studies (e.g., Gragg & Gordon, 1950, p. 3). As evidence that even today, Spearman's and Thurstone's influence is felt in classification battery development efforts, consider the following quote from Weeks et al. (1975).

3

The fundamental postulate, which has served as the basis for the development of the classification batteries, is that each Air Force job specialty requires a specific pattern of aptitudes for success. If the major aptitudes common to the various specialties can be separately measured, it would be possible to predict each applicant's probable success in any job specialty by means of an empirically weighted composite score based on those tests measuring aptitudes necessary for that specialty (p. 7).

It is useful at this point to review the degree of success to which such a testing philosophy has led. Despite widespread public attention and some confusion over the matter (e.g., Gould, 1982; Nairn, 1980), it is the case that relatively short (less than 3 hours total administration time) batteries of psychological tests are remarkably accurate at predicting future learning and performance criteria. As many have pointed out, one of the problems in some of the public criticisms is the failure to take into account the range restriction phenomenon. An inspection of the validity coefficients of the College Entrance Examination Boards's Graduate Record Exam, for example, gives a depressing picture of the utility of psychological tests. Wilson (1982, p. 16) found a validity coefficient of 0.27 using the GRE-V to predict first-year graduate grade point average (GPA) in verbal fields (N = 620) and a coefficient of 0.28 using the GRE-Q to predict GPA in quantitative fields (N = 529). Neither coefficient seems large enough at first glance to place a great deal of confidence in the tests' abilities to select candidates for graduate study. However, it is highly likely that the reason for the apparent modesty in the magnitude of the coefficients is to be found in the severely restricted range of ability characteristic of samples of students engaged in graduate study. Samples of military enlistees, on the other hand, offer a much greater, though still not completely representative, degree of heterogeneity, and countless validity studies conducted in military settings over the past 40 years attest to the utility of psychological tests. Table 1 summarizes results of a number of validity studies conducted using evolving versions of military selection and classification tests over the last 40 years. The validities shown in the table are correlations between weighted composites of test scores from the particular test battery with final technical school GPA. Across a wide variety of courses and a wide variety of batteries, validity coefficients are consistently high.

Table 1. Validity Coefficients of Air Force Test Batteries

Test Battery	Year	Number of Courses	Range of Validities	Median Validity	Median Sample Size
AC1-A	1951	29	.32 -- .77	.61	264
AC1-B	1956	21	.34 -- .77	.60	402
AC2-A	1959	46	.11 -- .80	.57	124
AQE-D	1958	3	.46 -- .50[a]	.47	182
AQE-F	1963	41	.29 -- .90	.63	433
AQE-62	1962	4	.75 -- .81[a]	.79	1493
AQE-64	1968	57	.38 -- .87	.64	410
AQE-66	1973	46	.18 -- .90	.68	115
AQE-J	1971	4	.69 -- .84[a]	.82	3396
ASVAB-3	1968	46	.29 -- .87[a]	.68	---[c]
AQE/AFQT (1)	1974	42	.16 -- .63[b]	.42	1000
AQE/AFQT (2)	1974	43	.16 -- .65[b]	.44	823
AQE/AFQT (3)	1974	57	.22 -- .68[b]	.53	890

Notes. The first ten rows are adapted from Weeks, Mullins, and Vitola (1975). The last three rows are adapted from Christal (1976).

[a]Inferred validities from test relationships with previous batteries for which actual validity studies were conducted.

[b]Not corrected for restriction of range.

[c]Unknown.

4

One of the problems with the data in Table 1, as well as data from most validity studies, is in the criterion of final technical school GPA. Weeks et al. (1975) realized the lack of an empirical job performance criterion as a critical limitation in these studies, but the problem until recently has been in the lack of any large scale efforts to develop satisfactory criteria. It seems likely now with pressure being applied both by the United States Congress and the military services that valid job performance measures are on the horizon (Eaton & Shields, 1985; Gould & Hedges, 1983). Christal (1976) viewed the problem not only in terms of the lack of good performance measures, but also in the inherent difficulty of "selling" the utility of aptitude tests by using only the grade point average criterion. He suggested considering other criterion variables, such as time to acquire skills, rate of skills decay, and time for skills reacquisition. He also reported data from a number of studies that showed that such criteria are as predictable by aptitude test scores as is the usual GPA. Nevertheless, there is an obvious need for validation studies to include criteria that are both more valid in their relationship to the target performance and more easily translated to dollars and cents utility.

If indeed it can be recognized that theoretical developments paved the way for subsequent selection and classification applications, it is useful to consider the directions that ability theory has moved since the days of Thurstone, in order to forecast future possible changes in actual selection and classification systems. One avenue of theoretical advance has been in what might be called factor theory. Much of this work may for the most part be viewed either as extensions or syntheses of the Spearman and Thurstone proposals. Thus, for example, Vernon (1961) and Burt (1940) have combined the two-factor and primary abilities models by proposing hierarchical models of ability organization, with general ability at the top and more specific abilities arranged in orderly fashion beneath. A currently popular variation on this scheme is that there are two general abilities: general fluid-analytic (Gf) and general crystallized (Gc) (Cattell, 1971; Horn, 1968). Gf is believed to be close, if not identical, to Spearman's g (Gustaffson, 1984) and is said to drive the development of Gc, which represents the product of accumulated learning experiences. Some developmental data would appear to support the distinction: Gf level peaks during early adulthood, while Gc level rises continuously with age (Cattell, 1971; Snow & Lohman, 1981).

Work on extending Thurstone's proposal is exemplified by Guilford's structure-of-the-intellect model, which proposes 120 abilities but, more importantly, specifies the dimensions of product, operation, and content along which mental tests can be classified. However, even Guilford (1982) admits that these specific abilities may be correlated, and as Gustaffson (1984) points out, this opens the door to a unifying hierarchical theory of ability organization. In fact, Gustaffson has proposed such a theory, the HILI model (Hierarchical, LISREL-based model), which takes advantage of recent developments in linear structural equation modeling techniques. Gustaffson proposes a unifying synthesis of the Thurstonian primary-factor model with the Cattell-Horn fluid-crystallized model. It is likely that such a unifying hierarchical model brings with it advantages for practical application. For the most general decisions about an individual's cognitive status, a test or battery of tests designed to tap the highest order factor might be administered. With more subtle requirements for classification decisions, coupled with the luxury of more available testing time, the samples might be obtained from the lower strata of ability levels.

Despite these potentially important recent developments in factor theories and methods, much of the theoretical interest in individual differences in learning and intelligence for the most part waned during the 1960s. This lack of interest was due primarily to disaffection for the method of factor analysis and its associated theories of learning and intelligence. Applied psychologists noted a stagnation in the field from a utility standpoint. Despite increases in the mathematical sophistication of factoring methods, a concomitant increase in occupational and academic validities was not demonstrated (Christal, 1981), which is apparent from the absence of any upward trend in the magnitude of coefficients in Table 1 over the years. At the same time,

others expressed a renewed dissatisfaction due to not understanding what it was that intelligence tests measured (Hunt, Frost, & Lunneborg, 1973; Sternberg, 1977). An important change in perspective occurred with the development of a psychology of cognition, starting in the late 1960s (see especially Neisser, 1967) and flourishing in the 1970s. The new cognitive psychology provided a set of methodological tools for exploring mental processes within an experimental approach, and beginning in the mid-1970s, the study of individual differences in mental processes was once again an active area of investigation.

With the shift from the traditional differential and correlational methods of investigating cognition to experimental methods came a shift in emphasis. While it is always true that individual differences research is concerned with the ways in which people differ and the underlying sources that lead to those differences, the newer approaches are at least equally concerned with specifying the necessary prerequisite knowledge and cognitive skills that allow any intelligent act, including learning, to occur. Thus, compared to the factor-analytic approach, the experimental approach is characterized to a considerably greater degree by the method of testing competing models of intelligent performance and only then examining individual differences in the parameters of the appropriate model. This general approach actually has taken two forms, labeled by Pellegrino and Glaser (1979) as the <u>cognitive correlates</u> and <u>cognitive components</u> approaches. The goal of investigations conducted within the cognitive correlates framework is to determine the cognitive skills and knowledge structures underlying observed differences between high and low skilled individuals in broad ability (e.g., verbal, spatial, numerical) or skill domains (e.g., physics, chess, electronics troubleshooting, geometry, computer programming). In this research, high and low ability individuals (or experts and novices) are identified, then administered a series of cognitive tasks, each of which is presumed to be well understood in its cognitive requirements. Tasks are typically selected so as to tap specific cognitive mechanisms which enables the researcher to test competing hypotheses about the sources of between-skill-group differences.

The goal of research conducted within the cognitive components framework is similar to cognitive correlates research in that it too seeks to explain the cognitive mechanisms underlying broad ability differences. The difference is that the components approach is to investigate directly the usually more complex tasks on which the ability differences are observed, an approach originally suggested by Estes (1974). In typical cognitive components investigations, individuals attempt tasks that bear a strong resemblance to intelligence test items. By systematically varying features of the task, the researcher allows various mathematical models to be fit to the latency or error data for individual subject. Each model is normally an embodiment of a particular theory of the tasks performed. Parameters of the models usually represent various psychological processes, and thus an individual's process execution times or probabilities are estimated directly in the parameters.

In the past 10 years, a considerable amount of individual differences work has used these two approaches or variations, and this has led to the crystallization of different views on the nature of individual differences in intelligence and learning ability. The remainder of this paper is devoted to consideration of this recent work.

III. THEORETICAL FOUNDATIONS

Typically, two criteria are used when determining which or what kinds of tests to include in a battery for selection and classification decisions. The first, which might be designated the <u>job sample criterion,</u> rests on the evaluation of the content validity of the candidate test. If one must select a secretary, for example, a battery that includes tests of typing skill and that samples other typical secretarial duties would satisfy this criterion. A second criterion, which might be labeled simply the <u>empirical criterion,</u> is satisfied to the degree that performance on

6

candidate tests correlate positively with performance on the target job. There are problems with both of these criteria. A logistic problem with the job sample criterion is that there may be too many jobs from which to extract work samples. If the test battery is to be used for classification decisions, and there are many possible jobs into which an applicant might be classified (as is the case in large organizations such as the Air Force), then the amount of testing time required to sample all possible jobs is prohibitive. However, an even more devastating problem with the job sample criterion is that job requirements are constantly changing. For example, it is likely that the skills involved in successful manuscript production on a mechanical typewriter are different from those employed in using a powerful word processing system.

There are also problems with the empirical criterion. First, the criterion can only be applied after a decision about what test to try has been made. It can be determined whether the currently existing battery of secretarial tests does an adequate job, but some other means for selecting those tests in the first place is required. Second, the empirical criterion does not provide an absolute standard against which to measure a test's success; if a validity coefficient of 0.30 is found, does that mean a test is a good one or not? The empirical criterion, applied after the fact, never provides information about whether some other test might prove to be more valid than the existing one. Finally, a third problem is that validity studies themselves can be quite expensive.

One of the premises this paper is based on is that these problems may be alleviated through the application of a technology of psychological testing derived from cognitive theory, in what might generically be called the cognitive skill assessment approach. Such an approach, in principle, would require (a) the determination of what cognitive skills are required in training and in the work place, (b) the determination of what cognitive skills are involved in taking psychological tests, and (c) the matching of training/job skills with cognitive task skills and thereby logically deriving training/job skills requirements. This would amount to a kind of decomposition analysis in which aptitudes would be redefined as sets of cognitive skills and jobs would be defined as sets of cognitive requirements. Such an approach would provide a different perspective on the person-job-match system (Holtz & Schmitz, 1985) and would serve as a flexible, adaptive system for specifying job skills and person skills for all kinds of training situations--computer-assisted instruction, on-the-job training, and even lockstep classroom instruction. Also, such a system offers promise as a test construction tool. One can imagine specifying in advance what cognitive skills will be measured when various facets of a complex cognitive task are systematically manipulated.

A system such as the one outlined, however, requires the foundation of a solid theory of individual differences in cognition. Unfortunately, such a theory does not yet exist, and much of the remainder of this paper will be devoted to an assessment of how far along is the development of such a theory. The first consideration is what such a theory should do:

1. A theory of individual differences in cognition should specify the cognitive processes and knowledge structures that underlie individual differences in the ability to acquire and apply knowledge and skills in a broad variety of contexts. Call the underlying attributes the sources of individual differences.

2. The theory should specify how these sources can be assessed at the level of the individual.

3. The assessment techniques should yield quantifiable indicators that serve both to provide an account of how well the theory fits the data (i.e., how well the source measurements predict performance in learning and intelligent-performance contexts) and be used in principle as ability measurements in an operational context.

7

In the last few years a number of theoreticians have applied some of the ideas emanating from cognitive psychology in speculating on the form a theory of individual differences in cognition might take. Snow (1978; 1980) has proposed that individuals might differ either in the efficiency with which they are able to execute elementary information processes, or in their approach to or their general strategy for attacking problems. Hunt (1978) has suggested that differences in cognitive abilities might be reduced to differences in knowledge, strategies, or mechanistic processes. And Sternberg (1977; 1980) has proposed an elaborate component hierarchy in which individuals differ in meta-components, performance components, acquisition components, retention components, and transfer components. All these proposals must be viewed as somewhat speculative at the present, but nevertheless they may be useful in proposing research directions.

In the LAMP project, a slightly different framework has been adopted, not only as a heuristic for guiding research but also as a way of organizing and classifying the existing and now burgeoning literature on individual differences in cognitive abilities. The framework is derived from a critical review of the existing cognitive-differential literature (Kyllonen, 1985a). From the review, three general conclusions can be drawn. First, whatever it is that underlies intelligent performance also underlies the ability to learn. This is consistent with both the empirical evidence and theoretical considerations derived from an analysis of current cognitive theory. Second, four sources can be tentatively identified as underlying the ability to learn and to perform intelligently. These are (a) working memory capacity, (b) information processing speed, (c) the declarative or factual knowledge base, and (d) the procedural or strategic knowledge base. Currently, these sources are merely taxonomic categories for variables that in principle could be measured on individual subjects. Nevertheless, such a taxonomic delineation is a useful first step. The third conclusion to be drawn is that the sources do not contribute additively to proficiency in learning and intelligent behavior--they interact. In particular, the extent of an individual's declarative and procedural knowledge base in a particular domain affects both the individual's effective working memory capacity and his or her speed of processing information related to that domain. These relationships are depicted graphically in Figure 1 in what is termed here the interactive common sources framework.
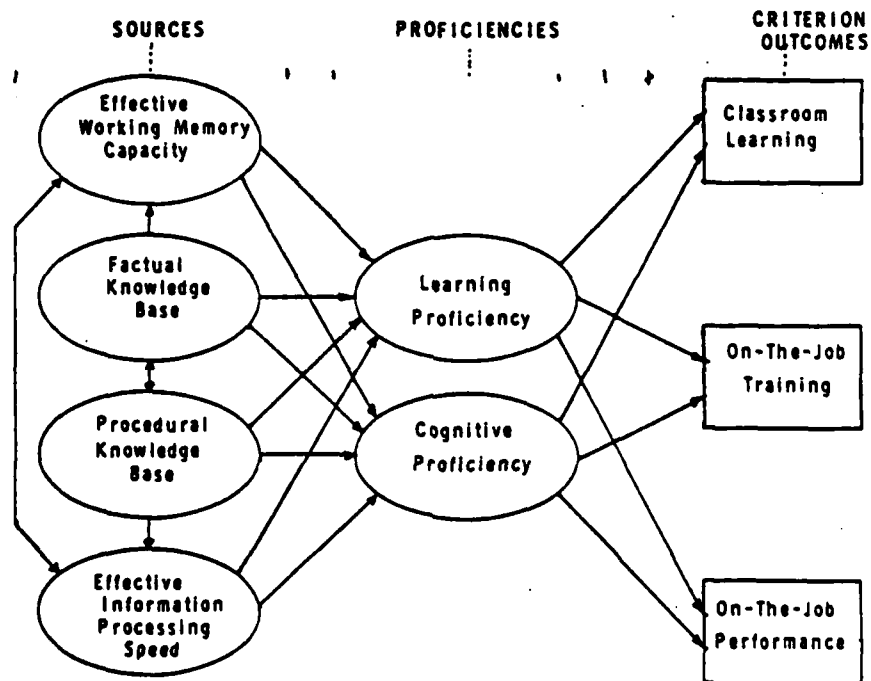

Figure 1. Interactive Common Sources Framework.

8

According to the framework, the success that an individual experiences in classroom learning activities, on-the-job training, and on-the-job performance is determined by the individual's level of cognitive and learning proficiency. Cognitive proficiency refers here to an individual's ability to remember, to make decisions and choices, and to solve problems. Learning proficiency refers to the individual's ability to acquire new facts and cognitive skills. This distinction is meant to align roughly with the classical distinction between learning and performance. The guidance provided by this framework also permits the tentative assumption that differences between people in learning and cognitive proficiency levels result from differences in the more fundamental sources of processing speed, memory capacity, declarative (factual), and procedural (rule-based) knowledge. The common sources view is that these components are what underlie both learning and cognitive differences. The interactive view is that the sources interact with each other in determining proficiency levels. For example, extensive factual knowledge in a particular domain (e.g., chess) can enlarge an individual's effective memory capacity (e.g., to memorize a complex board configuration) and effective processing speed (e.g., to select the best next move). Techniques for measuring these variables are discussed in the following section.

IV. ISSUES AND TECHNIQUES FOR COGNITIVE ASSESSMENT

The framework presented in the previous section can serve as a useful guide for research; nevertheless, important details have been left unspecified. Before the framework can evolve into a model or theory of individual differences in cognition, three classes of issues must be considered: assessment methodology, analysis of complex cognitive skill, and analysis of learning.

Methods for Cognitive Assessment

It is critical to address the issue of how the underlying sources of knowledge and the information processing parameters can be assessed. More precisely, the questions to be asked are as follows: How can processing speed be measured? How can an individual's working memory capacity be determined? How can the extent and quality of an individual's knowledge base be assessed? In each case, the obvious questions must be answered: Can the target source (processing speed, memory capacity, knowledge) be measured reliably? Is it already measured by conventional tests, or are new techniques required? Can the source be considered a unidimensional construct, or is there more than one dimension involved? And finally, does source capability change with practice, and if so, by how much?

These issues may be considered in the context of a number of cognitive correlates studies that have been conducted at AFHRL and elsewhere in the last few years. Much of this work has been driven by the general consideration of whether elementary cognitive tasks might someday supplement or even replace conventional tests as aptitude and performance measures. Cognitive psychologists have been remarkably successful in developing mathematical models that account for patterns of error and latency data across a large number of tasks by positing various mental processes and knowledge structures. It has occurred to a number of individual-differences researchers that to the degree such models are valid representations of psychological processing, the parameters of such models can serve as direct indicators of the speed or accuracy with which an individual can execute a particular psychological process.

9

In one of the first large scale efforts constructed with this general philosophy in mind, Rose and Fernandez (1977) described:

...a program of research dealing with the development and validation of a comprehensive standardized test battery that can be used as an assessment device for the evaluation of performance in a wide variety of situations.... Equally important, the battery is being designed to include tests that possess construct validity: there will be a firm theoretical and empirical base for inferring the information processing structures and functions that the tests purport to measure. It is expected that such a battery will permit improved personnel management decisions to be made for a wider variety of Navy-relevant jobs than is currently possible using existing techniques. (from the abstract).

With equal enthusiasm, Carroll (1980), after an extensive review of the then existing literature, proclaimed that the new approach of investigating individual differences with elementary cognitive tasks offered considerable promise not only in supplementing conventional tests but also as a means for assessing the effects of physiological changes and of aging. Of particular interest to Carroll was the possibility that the absolute measurement afforded by the analysis of cognitive tasks, as contrasted with relative measurement given by conventional correlational methods, might ultimately result in a <u>Systeme Internationale</u> of experimental psychology.

Individual Differences on Cognitive Tasks

It is instructive, then, to consider just how promising is the approach of supplementing traditional ability measures with scores from elementary cognitive tasks. The first issue *concerns whether there are reliable individual differences in various scores* that can be computed from such tasks. If the scores or parameters are not reliable, it does not mean that such scores are imprecise (Rogosa & Willett, 1983), but it does mean that there are no individual differences to speak of in the task or parameter of interest. Thus the establishment of reliability can be viewed as a central issue in the determination of whether a particular score is a good measure of individual differences.

In the Rose and Fernandez (1977) study, 54 college students were administered a battery of nine cognitive tasks, each presented on a computer. Between 36 and 38 subjects (depending on the task) were readministered the tasks on a second day, thereby allowing the computation of test-retest reliabilities. Tasks were selected to represent the domains of memory, psycholinguistics, and visual information processing. Further, tasks selected (a) had a history of published support, (b) had an adequate theoretical rationale, (c) were adaptable to paper-and-pencil or computerized administration, and (d) indicated that reliable individual differences were present on the task. Table 2 presents descriptive statistics on various scores computed from the tasks.

10

Table 2. Descriptive Statistics from Rose-Fernandez Study

	M1	M2	SD1	SD2	$r_{xx'}$
Posner and Mitchell's Letter Classification Task (Response Latencies)					
Physical Match (PI)	585	547	64	57	.57
Name Match (NI)	684	629	100	71	.58
Category Match (CI)	849	771	173	121	.78
Different	761	693	104	83	.81
NI-PI	99	81	62	45	.29
CI-NI	164	137	131	102	.69
Meyer's Lexical Decision Task (Response Latencies)					
Word Recognition	736	647	112	74	.66
NonWord Recognition	916	756	252	113	.53
Encoding Facilitation	975	958	78	75	.42
Baron's Graphemic/Phonemic Analysis Task (Response Latencies)					
Sense-Nonsense	1205	1193	246	197	.83
Homophone-sense	1289	1187	300	241	.90
Homophone-nonsense	1579	1423	306	235	.47
SH/HN	.81	.83	.09	.09	.37
Sternberg's Digit Scanning Task (Response Latencies)					
(Slope, Positive)	75	49	32	21	.60
(Intercept, Positive)	442	425	88	78	.52
(Slope Negative)	48	47	28	15	.45
(Intercept, Negative)	536	464	98	59	.51
Joula's Word Scanning Task (Response Latencies)					
(Slope Positive)	56	52	32	24	.19
(Intercept, Positive)	483	446	102	89	.46
(Slope, Negative)	47	53	32	31	.00
(Intercept, Negative)	544	446	145	67	.40
Joula's Category Scanning Task (Response Latencies)					
(Slope, Positive)	122	93	85	65	.31
(Intercept, Positive)	611	637	245	216	.68
(Slope, Negative)	214	140	96	56	.32
(Intercept, Negative)	575	595	238	176	.36
Clark's Sentence-Picture Verification Task (Response Latencies)					
"below"	136	110	155	149	-.06
"negate"	829	685	354	319	.81
"comparison"	200	146	183	200	.28
"encode"	1735	1489	404	330	.59

Table 2 (Concluded)

	M1	M2	SD1	SD2	r_{xx}'
Collins and Quillian's Fact Verification Task (Response Latencies)					
Slope, superset relation	63	42	57	77	.21
Intercept, superset relation	1035	1017	205	220	.69
Slope, property relation	67	53	89	75	.16
Intercept, property relation	1118	1121	257	248	.73
Shepard & Teghtsoonian's Recognition Memory Task (Probability Correct)					
Proportion Correct	.73	.73	.07	.07	.56
exponent on decay function (lag)	-.07	-.10	.19	.08	.31
intercept on decay function (lag)	.86	.93	.15	.12	.21
Probability (hit)	.73	.77	.11	.12	.56
Probability (false alarm)	.28	.31	.12	.12	.67
d' (sensitivity)	1.28	1.34	.41	.45	.62
Beta (bias)	1.09	.93	.62	.50	.38

Notes. Response latencies are in milliseconds; M1: Mean on day 1, M2: Mean on day 2, SD1: Standard deviation over subjects on day 1, SD2: Standard deviation on day 2, R_{xx}': day 1, day 2 correlation.

The various parameters for each task are described in detail by Rose and Fernandez (1977). It may nevertheless be useful to consider the top listed task, Posner and Mitchell's (1967) Letter Classification Task, in more detail, both to provide a sense for the scores that are derived from a single task, and because this task in particular has received considerable attention in the individual-differences *literature*. In this task the subject is shown two letters side by side. The task is to determine whether the two letters are (a) physically identical (e.g., A-A), (b) identical in name (e.g., A-a), or (c) are the same in terms of vowel-consonant category (e.g., A-E), depending on individual task instructions. If the letters did match according to task instructions, the subject was to respond by pressing one of two keys on a panel, but if the letters were not the same, the subject was instructed to press the other key. In addition to the mean response times for each of these three tasks (for "same" trials), three other scores could be computed. A "Difference" score was computed from response times on trials for which the letters did not match. A name-identity (NI) minus physical-identity (PI) response time score was computed to reflect the speed with which an individual can access information from long-term memory. The rationale behind this computation is that to make the PI judgment, an individual can respond on the basis of the displayed physical information, but to make an NI judgment, the individual must retrieve the name of both letters from long-term memory before these abstract name codes can be compared to one another.

Table 2 presents means and standard deviations, in milliseconds, for the day 1 and day 2 statistics from the Rose and Fernandez study, in the columns marked M1, M2, SD1, and SD2. Also, the test-retest correlations are displayed as reliability indices in the column marked r_{xx}'. Note that in all but a very few cases, there was a considerable decrement in response latency over days, and also a corresponding decrement in variability. The test-retest reliability data show that in many cases the ordering of individuals changed substantially from one day to the next. This can be interpreted in one of two ways. The conventional wisdom is that this indicates such scores are unstable and therefore not good candidates for a performance test battery. Alternatively, if both day 1 and day 2 internal reliabilities are high, but the test-retest reliability is low, it can mean that some individuals are benefiting from practice and others are not, which itself could be an important individual difference variable. Unfortunately, Rose and Fernandez did not provide internal consistency data.

In any event, caution should be applied before taking any of Rose and Fernandez' reliabilities too seriously, because they are based on an extremely small sample. Yet the pattern of reliabilities may still be informative, and one pattern result apparent from Table 2 is that derived scores are generally less reliable than scores that represent the duration of performing a complete task, which is consistent with many other studies in the literature. For example, while both the NI and PI match scores are highly reliable, the NI-PI difference score is not. Carter and Krause (1983) reported data on some of the tasks in Table 2, along with some others, and found that, in all cases, slope scores (a kind of difference score) were less reliable than were the mean response times from which the slopes were computed. From this result, they argued that slope scores should not be used as performance measures, but rather the mean response times by themselves are sufficient for answering most questions the applied researcher might be interested in asking. However, the Carter and Krause argument is at odds with the stated philosophy of Rose and Fernandez, who argued that total scores are often less meaningful than scores derived from total scores insofar as they reflect combinations of psychological processes rather than a single process, such as "memory comparison." The meaning of this discrepancy is discussed in the following paragraphs.

The topic of difference scores has been a highly controversial one in the psychological literature for at least the last 25 years, but a recent analysis of the topic by Rogosa and his colleagues (Rogosa, Brandt, & Zimowski, 1982; Rogosa & Willett, 1983) is clarifying. Rogosa et al. argued that considering the statistical, as well as psychometric, properties of change measures leads to an evaluation of the reliability of the difference score that is at odds with widely accepted notions. In particular, Rogosa and colleagues showed that the difference score is unreliable when individual differences in change do not exist, but that it can be highly reliable if in fact individual differences in change do exist. In many of the tasks inspected by Carter and Krause and others, there is a high correlation between two of the scores from which the slope is computed. Also, in many of the reported studies, there is an extremely high correlation between response time on the NI and PI match tasks. What these high correlations indicate is that there is very little in the way of individual differences in the change between performance on the two tasks. Thus, it is not merely a statistical artifact that produces low change score reliabilities. Rather, the low reliabilities (or conversely, high correlations between task 1 and task 2) are interesting empirical results that establish the lack of individual differences in the change variable.

It should be pointed out that Carroll (1980) has presented analyses of data on many studies that have appeared in the literature and are similar to the Rose and Fernandez study (1977). Unfortunately, the vast majority of those studies, like the Rose and Fernandez study, also suffered from a small sample size. More recently, the Naval Biodynamics Laboratory has supported a number of studies that have investigated the psychometric properties of a large number of cognitive tasks (Carter & Krause, 1983; Kennedy, Bittner, Carter, Krause, Harbeson, McCafferty, Pepper, & Wiker, 1981; Kennedy, Bittner, Harbeson, & Jones, 1981). These studies too have suffered from small sample sizes.

Do Cognitive Tasks Measure Unique Abilities?

If it turns out that there are reliable individual differences on many elementary cognitive tasks, then the question of whether these scores represent previously unidentified abilities can be addressed. A recently completed study at AFHRL approached this question by comparing cognitive task scores to aptitude test scores. Fairbank, Tirre, and Anderson (1984) administered 30 different cognitive tasks, divided into six task batteries, to six independent samples of Air Force enlistees. Tasks were selected from a taxonomy to reflect verbal, spatial, and quantitative processing and to yield meaningful error and latency scores. That is, not all the tasks were designed solely to measure some form of processing speed. For each task, standard

13

mathematical models of errors and solution latency were fit to the data, then certain parameters of these models were extracted as indicators of various aspects of processing efficiency. The rationale behind the different models will not be discussed here, but the rationale was similar to that given by Rose and Fernandez. However, Fairbank et al. also reported some total task score data (i.e., mean over all items). With reference to the previous discussion, such scores are useful to the extent that individual differences on scores derived from total scores do not exist.

One key result from the data published by Fairbank et al. (1984) data was that reliabilities tended to be high, indicating stable individual differences on most of these tasks. When considering average scores, reliabilities for all the elementary reaction time tasks exceeded .90. On the other hand, derived score reliabilities presented a mixed picture. Some of the derived scores, such as the NI-PI difference, had low reliability (less than .50), while others, such as the slope from memory scanning tasks (Sternberg, 1969), had high reliabilities (greater than .80). Pellegrino (1984) has also found high reliabilities among memory scanning tasks. Fairbank et al. also computed correlations between the information processing scores and standardized test score composites taken from the Armed Services Vocational Aptitude Battery (Department of Defense, 1984). The General (G), Administrative (A), Electronics (E), and Mechanical (M) composites are those used by the Air Force for personnel classification purposes. Thus, such correlations reflect the degree to which cognitive task data overlap the conventional test score data, that are used operationally in the Air Force.

The pattern of correlations with ASVAB composites showed that despite the fact that the cognitive task scores were fairly reliable, correlations with conventional measures tended to be fairly low: No correlation exceeded 0.50. This indicates that conventional tests do not tap, at least in a factorial pure sense, the processing skills tapped by the cognitive tasks. Thus, although cognitive tasks are not likely to replace conventional tests, there might very well be room to supplement them. Elsewhere, Kyllonen (1985b) reported data consistent with the Fairbank et al. finding on this matter. In a factor-analytic investigation of 17 cognitive tasks that tapped verbal, quantitative, and reasoning abilities, along with ASVAB tests, it was found that of eight cognitive factors identified by the conventional and cognitive tests, three (reasoning speed, verbal speed, and quantitative level) were not measured by the ASVAB tests.

There are a number of other interesting trends in the Fairbank et al. (1984) data. First, while most of the response time measures correlated highest with the A composite (which bears considerable resemblance to the Clerical or Perceptual Speed Factor in the differential literature), the percent correct measures computed on the more complex reasoning and memory tasks correlated higher with the G composite (the general ability measure). This suggests that error and latency data may reveal different aspects of performance, a finding consistent with the Kyllonen (1985b) results. It also was found that the intercept in the scanning tasks correlated highest with the A composite, while the slope showed no strong or consistent differential correlation pattern. The intercept is presumed to reflect time for encoding and response, and the slope is presumed to reflect the time it takes to perform a single memory comparison step. Also, Fairbank et al. found that with very few exceptions, scores from cognitive tasks correlated higher with either the A or G composite than they did with the M or E composite, presumably because these latter two scores reflect primarily the extent of specialized knowledge bases in either the mechanical or electronics domains.

Two other results are worthy of mention. One is that mean and standard deviation were highly correlated on almost all of the response time measures, reflecting increasing variability being associated with slower responding. Finally, on two linguistic transformation tasks, Sentence Verification and Three-Term Series, Fairbank et al. computed percent correct scores separately for the first and final blocks of items. In both cases, the correlation between percent correct and G was significantly higher on the first than on the final block, perhaps reflecting the fact

that general cognitive demands are greater early on in a test. This is consistent with Fleishman and Hempel's (1954) classic demonstration of changing cognitive demands on a psychomotor task as a function of practice.

At the time of this writing, Fairbank et al. are currently engaged in further analysis of their extensive data set to explore the degree to which parameters measured on one set of tasks correspond to parameters on various other sets. The purpose of the analysis is to explore the generality of the various processing parameters. Although similar kinds of analyses have been conducted by others (Carroll, 1980; Rose & Fernandez, 1977), as was mentioned previously, most of these have suffered from lack of power due to inadequate sample size. With cognitive task data collected on over 2500 Air Force trainees, it should prove possible to examine various hypotheses beyond the reach of others who have attempted to do so with smaller subject pools.

Analysis of Changes in Processing Efficiency

The final major issue to be addressed in conjunction with the general consideration of assessment methodology concerns the degree to which subjects improve with practice. Rose and Fernandez (1977) showed large improvements in the second day, and these findings are consistent with many published reports of performance on cognitive tasks that have not examined individual differences. A fairly well-established finding across many diverse cognitive tasks is that response time decreases in accordance with a power law (Lewis, 1980; Newell & Rosenbloom, 1981), as $RT(N) = aN^b$, where $RT(N)$ is response time for an item at trial $\underline{N}$, $\underline{a}$ is response time for the item on trial 1, and $\underline{b}$ is the parameter governing the rate of change. Recently, individual differences researchers have begun to examine the question of whether rate of improvement is an important source of individual difference variation that deserves special consideration.

In one noteworthy study, conducted as part of the Project LAMP, Pellegrino (1984) investigated individual differences in changes in information processing efficiency on fairly simple cognitive tasks of the type reviewed in previous sections. In his study, 60 young adults were administered three cognitive tasks in four to eight successive sessions. Tasks were designed to sample visual, verbal, and quantitative abilities. The visual processing task was a perceptual matching task in which subjects were required to compare matrices of varying size with one another for physical identity. The semantic processing task was a version of the matching task in which subjects were instructed to match by letter category, but letters could be the same (or different) physically or by name or category. The quantitative task presented elementary number facts involving either addition, subtraction, multiplication, or division, and the subject indicated whether the given answer was true. Analyses centered on (a) determining the adequacy of various mathematical models proposed which posited elementary operations such as encoding, comparison, decision, and response, (b) determining the relationships between information processing parameters and standard aptitude test scores, and (c) determining the relationships between standard test scores and slope and intercept parameters derived from a power law analysis of practice effects.

The initial analyses indicated that the models fit the data quite well, and parameters from the models were generally reliable. Also, Pellegrino found that individuals performed all the tasks faster as a function of practice, in accordance with the power law characterization of rate of change. The processing parameter-aptitude correlations were generally low, except in the case of the perceptual speed factor which was modestly related to the intercept parameter in some of the models (where the intercept reflected time for encoding and responding), consistent with the Fairbank et al. (1984) results. The analysis of starting point (intercept) and rate of change (slope) parameters showed that starting point in many cases was significantly related to perceptual speed, and also was related across tasks (median $\underline{r} = 0.48$). However, change rate was unrelated to any of the aptitude measures. Further, change rates across tasks were unrelated to

15

each other (median r = 0.08). This result suggests that there is not a general learning ability, at least insofar as this particular type of change is regarded as a form of learning. Rather, learning of this type seems to be task specific.

Recently, Ackerman and Schneider (1984) reviewed a number of studies that showed data consistent with Pellegrino's results. The authors observed that correlations between initial performance and final performance on a wide variety of psychomotor and cognitive tasks have typically been found to be quite low, which is to say that individual differences early in training do not map very neatly onto individual differences later in training. Considering that the purpose of ability tests when used as aptitude measures is to predict individual differences at the end of training, this is a lamentable finding. It also clashes with the view that intelligence is related to the ability to learn.

With these observations as an impetus, Ackerman and Schneider attempted to provide a comprehensive account of individual differences in changes in information processing efficiency by synthesizing current ideas on the structure of human abilities with considerations of automatic/controlled processing theory (Shiffrin & Schneider, 1977). The ability model they adopted is of the hierarchical variety, reviewed in a previous section (along the lines of Gustaffson's, 1984, HILI model). The processing model assumed that there are two distinct forms of processing: controlled and automatic. Automatic processing is fast, can be done simultaneously (i.e., in parallel) with other processing activities, and does not draw attentional resources. Controlled processing is slow, is performed in serial, and draws heavily on attention. It has been shown that processing can become more automatic with extensive practice, so long as processing requirements remain consistent from trial to trial. However, if requirements vary, processing will remain in a controlled state (Fisk & Schneider, 1983).

Ackerman and Schneider proposed a mapping between the concept of general attentional resources (Kahneman, 1973) and general intellectual ability, from which can be derived the prediction that general ability and broad-domain abilities should be related to success on tasks requiring substantial amounts of controlled processing (where controlled processing, by definition, requires a heavy investment of general or broad-domain attentional resources). Conversely, the authors also proposed a mapping of automatic processing and lower-order, highly specific abilities, which leads to the prediction that general ability will not be related to success (e.g., speed) during automatic processing, but some low-order factor might be. Ackerman and Schneider (see also Ackerman, 1984) reported data that lent some support to their proposal. They showed that on a task that prevented the development of process automaticity (by varying the processing requirements over trials), general ability and verbal ability (i.e., a broad-domain ability) were highly related to response time. The relationship between these broad abilities and performance on a task that enabled the development of process automaticity (by maintaining consistent processing requirements) was lower. It also was found that the relationship between perceptual/motor ability and task performance did not differ between the two tasks.

The Pellegrino and the Ackerman and Schneider studies represent important initial steps in the investigation of individual differences in a particular kind of learning, which Rumelhart and Norman (1978) have called "tuning." These researchers have gone beyond considering individual differences in initial performance in moving toward a characterization of differences in performance changes as a function of practice. Given that such changes are so extremely commonplace on cognitive tasks, and also that the purpose of using cognitive tasks for selection and classification is to predict the endpoint of often extensive training, it is likely that this topic and the general approaches employed by these researchers will become important cornerstones in future individual differences research. Applications may be particularly appropriate in endeavors that apply such methods to the prediction of success in high-speed decision-making activities such as air traffic controller and aircraft pilot. Although Ackerman and Schneider's call for a change in assessment procedures based on automatic/controlled processing theory may be

16

a bit premature (and impractical in that as much as 20 hours of testing may be required), a proposal to continue research along these lines is not premature.

V. ANALYSIS OF LEARNING

As Ackerman and Schneider (1984) rightfully point out, the purpose of ability testing is often to predict who will do best after extensive training. That being the case, it would seem appropriate to direct ability testing toward the analysis of learning ability. That is, it is important that ability assessment devices yield information about who learns best in a particular situation, and therefore, who will end up as the best performer after training. But just what is learning ability? Is it a unitary individual differences construct, or are there multiple varieties of learning ability?

Rumelhart and Norman (1978) have distinguished three types of learning, which they call accretion, restructuring, and tuning. Accretion refers to the accumulation of facts; restructuring, to the development of new cognitive procedures; and tuning, to the process of making existing cognitive procedures more efficient. Viewed along these lines, the previous section reviewed studies on individual differences in tuning processes. In the present section, the focus is on individual differences in the other forms of learning, particularly accretion.

Why is the examination of accretion processes important? In the conventional testing setting, the assessment procedures yield data that reflect the current skill level of an examinee. These data certainly must reflect to some degree the amount of prior exposure an individual has had to problems of the type administered. Yet it is possible that individuals reliably differ in their learning rate and thus relations between initial level and final level relations might be attenuated as a result. This attenuation would manifest itself in validity coefficients lower than might be obtained if learning rate were considered along with initial level. With this possibility as a motivation, a number of studies conducted in the last few years have been concerned with determining whether learning rate is likely to be an important variable for consideration in future assessment batteries. In the past, it has been difficult to conduct studies of this nature because of the difficulty of exercising sufficient control over stimulus presentation and response feedback, both of which may be important in investigating dynamic learning processes, but the current widespread availability of microcomputers alleviates this problem.

Two recent studies by Allen and Morgan and colleagues were concerned with the relationship between initial level and learning rate and with the relationship between learning rate and conventional ability measures. In the first study (Allen, Secundo, Salas, & Morgan, 1983), 70 college students were administered three learning tasks (a fourth task was administered but data were not analyzed due to a computer malfunction). In a Coded Messages Task, subjects studied 12 word-symbol pairs, then made a series of same-different judgments on the equivalence of sentences and symbol strings. For example, an examinee might be questioned about whether the sentence "Enemy aircraft approaching from the North" was equivalent to the string "* I @ -." In an Emergency Procedures Task, subjects studied a set of procedures on how to handle emergencies and then were tested for their knowledge on the serial order of procedures. In a Security Checking Task, subjects studied a map of landmarks on a hypothetical Air Force base, where each landmark had an associated security level (high, medium, or low). In the test, subjects were asked questions such as "What is the security level of the second low security location after the tower?" Allen et al. computed three scores for each subject. An initial learning level score was the sum of correct responses during the first third (17 minutes) of the task, a final level score was the sum of correct responses during the final third of the task, and a rate score was the difference between the two level scores. The important findings were that first, initial learning level, by itself, did not predict final level accurately; the rate variable

17

significantly added to the prediction. Second, initial learning level was related across the different learning tasks, but rate and final level scores were not. Recall that Pellegrino (1984) found much the same result in his simpler classification and matching tasks.

In a follow-on study (Allen, Salas, Pitts, Terranova, & Morgan, 1984), these investigators readministered the same four learning tasks and three additional learning tasks (to separate groups of 63 and 60 students, respectively) along with a battery of conventional aptitude tests. The purpose of this study was to determine whether final levels of performance could be predicted solely by conventional test scores or whether scores from learning tasks would account for additional variance. They found that factor scores derived from an analysis of conventional ability tests were significantly related to final performance on four of the seven learning tasks (with r^2 ranging from 0.29 to 0.38) but that, on all but one task, goodness of prediction of final level was significantly enhanced by the inclusion of learning rate variables in the prediction equation (increment in r^2 ranging from 0.07 to 0.24). This result suggests that learning rate may be a reliable and somewhat generalizable individual difference variable that is not currently measured by conventional tests. A more compelling demonstration of the utility of learning rate measures, however, would show that such measures predict final performance levels in a long-term learning environment such as that found in standard 2-month technical training courses in industry or the military.

Thus, in a third study in the series, Allen, Pitts, Jones, and Morgan (1985) explored the utility of their learning rate measures in predicting final performance levels (course grade) in technical college courses (computer science, N = 90; bacteriology, N = 66; and engineering N = 48). The major hypothesis tested was that learning task parameters (slope and intercept) would add to the utility of the standard measures (high school grade point average and Scholastic Aptitude Test scores) in predicting final course grade. The intercept of the learning function in the Allen et al. study reflected the amount learned during the pre-performance instructional phase of the task; the slope reflected the average amount of performance improvement during each minute of the task. Thus, both parameters were in some way reflective of learning rate.

Analyses showed that the learning rate measures were significant predictors of final grade in all three courses when considered separately and that, in some cases, the rate measures accounted for additional variance in final grade beyond that accounted for by either high school GPA or SAT scores. However, the rate measures did not contribute to the predictive efficiency of the equation that included both GPA and SAT scores. Setting aside the issue of statistical power, this result could be interpreted as indicating that learning rate is already reflected to some degree in either or both the GPA and SAT scores.

Nevertheless, the series of studies taken together demonstrates that even a fairly rough approach to the analysis of learning may be profitable in providing assessment measures with practical utility. Yet, there is a need to consider at a somewhat more basic level, what it is that contributes to differences between people in variables such as learning rate. The theoretical framework outlined earlier in this chapter, for example, could be read to imply that differences in more fundamental source variables were responsible for differences in general acquisition proficiency. If so, then findings such as some of those emerging from the studies just reviewed need themselves to be explained in more detail. Recall the earlier argument that the reason such detailed analysis is sought, apart from general scientific pursuits, is to provide potentially greater adaptability and flexibility in a system of assessment.

In a recent study, Kyllonen, Tirre, and Christal (1984) addressed directly the question of how processing speed and efficiency, along with factual and procedural knowledge, can play a central role in learning. The logic employed ran as follows. In typical paired associates learning, the likelihood that a person would recall the response term, given the stimulus cue, depends on the density of the memory structure that was created at the time of study. That is,

if a person can create a highly elaborated, interactive structure that connects the stimulus and response terms during study, then it is more likely that the pair will be successfully retrieved (or recognized) at test. Alternatively, if the learner fails to create any structure that links items together at study time, then probability of retrieval is reduced. This is essentially the finding of the utility of mnemonic devices.

A richer long-term memory structure allows greater opportunity for retrieval because it allows more entry points to access that portion of the structure that is relevant to the memory task. The hypothesis was that subjects with high verbal knowledge (i.e., those who score well on standard vocabulary tests) come to the experimental session with an already well-developed declarative memory structure for words and associated concepts. If given plenty of time for study, these subjects will have a distinct advantage over their low verbal knowledge counterparts in their ability to integrate stimulus and response terms and thus successfully retrieve pairs at test. The reason for this is that the concepts activated during study should be more highly integrated and thus serve as good cues, or entry points, at test. The paucity of structure characterizing low verbal subjects leads to the activation of fewer and perhaps less distinctive concepts, thus leading to more retrieval difficulties at test.

If study time is short, on the other hand, the advantage for high verbal subjects might not be as great. Although by the logic above, there should still be some advantage, the most critical variable in establishing a connected structure to facilitate recall should be the rate at which relevant concepts can be retrieved. In sum, the prediction was that, with liberal study time, the high verbal knowledge individual should have a distinct advantage; with limited study time, the fast verbal processor should have the distinct advantage.

To test these hypotheses, a series of cognitive tasks was presented to Air Force enlistees; these tasks were designed to assess the enlistees' breadth of verbal knowledge and the speed with which they were able to access verbal concepts. A paired associates learning task was also administered to subjects. In the task, pairs were presented for study at one of five rates ranging from .5 to 8 seconds.

Two hypotheses were tested which related to these notions about learning ability. First, an individual's likelihood of correctly responding on items in which pairs were presented at the slow rate (8 s) would be more highly related to the breadth of an individual's verbal knowledge than would be the likelihood of correctly responding on items presented at the fast rate (.5 s). In both cases, some relationship was taken for granted but was expected to be greater at the 8 s condition. The second hypothesis was that the reverse relationship should hold when the variable of interest was verbal processing speed. That is, probability correct at the .5 s level should be more highly related to processing speed than probability correct at the 8 s rate. Indeed there was no compelling reason to believe that processing speed should have any effect on probability correct at the 8 s rate.

Although the analysis is fairly complex, it was found that under certain conditions the expected relationships did hold. The relationship between learning success and word knowledge was higher in the 8 s condition than in the .5 s condition. And the relationship between learning success and verbal processing speed was higher in the high speed (.5 s) study condition than in the leisurely 8 s condition. Because a number of measures of processing speed were included, each designed to tap a different configuration of psychological processes, it was possible to isolate the source of processing speed differences operating in the paired associates task. The analysis showed that simple motor speed differences, or even simple comparison speed differences, could be ruled out. The critical speed seemed to have been how quickly an individual was able to search memory to retrieve a relevant concept.

19

One of the implications to be drawn from this study is related to the use to which verbal (or more generally, semantic) processing speed tasks might be put in application efforts. Much attention has been given in recent years to the letter classification task as a useful index of verbal ability (Hunt, 1978). This results from the modest but apparently reliable relationship found between response time measures on the task and composite verbal aptitude measures. The original interest in the finding was related entirely to the theoretical issue of understanding a cognitive component that might have a causal linkage to the development of verbal knowledge. But there might also have been the implicit belief that the task, along with other similar tasks, might somehow serve as replacement for knowledge-dependent verbal aptitude tests. The Kyllonen et al. study suggests a more appropriate use for verbal processing speed tests. That is, under conditions of high information flow, such as experienced in the cockpit or a variety of similar situations, such a measure of processing efficiency might predict who will remember information most effectively as it flows through in real time. (See Christal, Tirre, and Kyllonen (1984) for further discussion along these lines.)

Although the studies reviewed in this section present provocative findings, the limitations of this research should be emphasized. In both the Allen et al. and the Kyllonen et al. studies, the criterion tasks were fairly simple fact acquisition tasks. Understanding the relationships between parameters on such tasks, and understanding the cognitive determinants of performance on such tasks, is only a useful first step to understanding what it is that causes some to learn faster than others. What is needed to make true progress in understanding learning ability is the analysis of how skills develop in the context of more realistic and long-term learning environments. The typical validity studies that identify correlates of final GPA after weeks of technical training are not the solution in that they present a whole class of new problems, mostly related to the lack of control over the conditions of learning, and the failure in yielding dynamic learning progress indices.

A new development that seems promising at the present is the analysis of learning in the computerized intelligent tutoring system environment where a great deal of information about what progress a student is making could be computed (in principle), and full control over the conditions of learning is possible. One study conducted along these lines compared a variety of aptitude and motivation test scores to a whole host of dynamic learning variables (Snow, Wescourt, & Collins, 1983). The study, although only a pilot (in that only a small number of students were available for analysis), showed the great potential for integrating cognitive assessment methods with computerized instruction in an effort to discover the underlying sources of the ability to learn. It is likely that future efforts of this type will be forthcoming.

VI. ANALYSIS OF COGNITIVE SKILL

One of the more exciting recent developments in cognitive assessment comes from the analysis of complex cognitive skills such as reading, troubleshooting (of electronic systems), typing, mathematics problem solving, and computer programming. Much of this research employs many of the same techniques and methods employed by researchers investigating learning and more elementary cognitive skills, but the goals of the research tend to center around the issue of the nature of basic skills for potential training applications rather than selection and classification applications. The general idea is that if the underlying cognitive constituents of complex behavior are understood, more effective diagnoses of learning and performance disabilities might be possible and thereby result in prescribing more effective remedial training.

One particularly nice application of this strategy can be found in a systematic program of research studies conducted by Frederickson and colleagues in the area of general reading skills (Fredericksen, Weaver, Warren, Gillotte, Rosebery, Freeman, & Goodman, 1983). Initial work on this project concerned the identification of the components of reading (Fredericksen, 1981; 1982)

through the use of differential assessment techniques of the type discussed throughout this chapter. The research strategy was to administer cognitive tasks such as letter matching, word recognition, and anagram encoding, and to test various models that accounted for the pattern of relationships among measures derived from the tasks. This analysis then resulted in an identification of component skills, individual estimates of which in turn were correlated with scores on standard reading tests to determine the skills that differentiated good from poor readers. Fredericksen et al. (1983) then selected three of the components as particularly critical for reading, and developed specific computerized remedial training of the components. The training proved successful on a series of reading tasks, but even more interesting was the fact that Fredericksen et al. administered a variety of cognitive criterion measures to identify which training had an effect on which particular set of skills.

The Fredericksen study is only one of a number of similar (if not as wide-scoped and systematic) studies that have compared experts in a particular subject-matter domain with novices in an attempt to define the underlying component skills of expertise. Thus far, research has tended toward the analysis of more academic expertise, such as physics (Chi, Glaser, & Rees, 1981; Larkin, McDermott, Simon, & Simon, 1980), but it is not out of line to imagine the analysis of more prosaic but nevertheless critical areas of technical training.

Recently, the Air Force has become convinced that new cognitive methods of analysis hold the key to a redefinition of what it is that constitutes a "basic skill." Gott and Davis (1983) have related this reconceptualization to a switch from a focus on what they call a power-based strategy for assessing general facility to a knowledge-based approach that recognizes the narrow, domain-specific nature of cognitive skill. Traditionally, both within and outside the military system, basic skills have been defined as the three Rs of reading, writing, and arithmetic. However, there is a growing realization that skills defined at this level of generality do not lead easily to prescriptions for how remediation for skills deficiencies can be accomplished. There is the hope that a more fundamental domain-specific characterization of skill might more naturally suggest techniques for overcoming particular deficiencies.

In the first large scale effort of this kind, two occupational specialties in the Air Force (Jet Engine Mechanics and Avionics Troubleshooting) have been examined (Bond, Eastman, Gitomer, Glaser, & Lesgold, 1983). In an extensive dissertation, Gitomer (1984) has documented a number of studies concerned with identifying the basic skills involved in troubleshooting electronic aircraft equipment. The study was motivated by the observation that there are tremendous differences between first-term airmen in their ability to perform troubleshooting effectively, despite the fact that the airmen considered had all completed extensive technical training. Based on ratings by supervisors, Gitomer divided 16 airmen into two groups (of $N = 7$ high-skilled and $N = 9$ low-skilled) and proceeded to administer a series of tasks groups to isolate the source of the skill difference. Tasks ranged in complexity from a complex troubleshooting simulation to simple picture-name classification. Some were variants on standard cognitive tasks discussed elsewhere in this paper, such as one that required examinees to identify the name of a pictured component, which bears a resemblance to the name identity task. Others, such as a series of component clustering tasks, and some open interview tasks (e.g., "tell me all you know about azimuth hydraulic actuators") have been used in connection with studies on physics expertise (Chi, Feltovich, & Glaser, 1981). Still others were fairly domain-specific cognitive tasks that resulted from a careful task analysis of troubleshooting activities. An example is the Logic Gate Computation task in which an examinee is required to fill in a blank given a partially complete logic gate truth table (e.g., given the relationship, "NAND," and the input values, high, and low, what is the output value?).

From the pattern of differences found over the 13 tasks administered, Gitomer was able to paint a picture of the constitution of troubleshooting skill. He found that the more skilled performers differed from the less skilled in that they were driven by better specified goals more

21

consistent with task demands, they had more methods available to them for attacking a problem, they were able to execute such methods more efficiently, and they were better able to select appropriate problem solving methods across different situations. Some of the areas in which differences were not found may be as revealing as areas of difference. Gitomer found no effect for time in training, and he found (surprisingly) no differences between the two groups on their Electronics Aptitude score. He also found that both groups had poor knowledge of general electronics principles, presumably because after the first few weeks of formal training, such general principles played no part in job task activities.

One of the important issues related to studies along the lines of this one is the question of how general are basic skills. Although the generality issue cannot be addressed systematically on the basis of a single study, there is some evidence from Gitomer's work that not all the differences were narrow domain-specific differences. The logic gate task, although directly a part of avionics troubleshooting activities, is actually quite general in that it plays a role in a broad variety of complex cognitive activities such as logical analysis and computer programming. And the task revealed substantial differences between high and low skilled performers. Although Gitomer did not completely spell out the reasons why the difference might have shown up so clearly, it is possible that the cause may have been related to differences in general working memory capacity, one of the sources identified in the interactive common sources framework (see Figure 1).

The important point to draw from this work, which is really only in its preliminary stages, is that through the administration of carefully constructed cognitive tasks it should prove possible to isolate the sources of differences among people who perform complex activities with differing degrees of proficiency. The results of such analyses should be prescriptive statements about how the less skilled individuals might be tutored to overcome specific deficits. In this regard, Gitomer has shown how the lack of differences in many of the tasks he administered is actually a quite encouraging sign in that it demonstrates that the poor troubleshooters are not simply worse at all cognitive skills, but rather they suffer particular and isolable deficiencies. Powerful prescriptions for training are much more likely to result from considerations for these specific deficiencies than from global recommendations to train people to "read better."

VII. SUMMARY

The purpose of cognitive assessment is to apply current understanding of how people think, learn, and remember toward the measurement of an individual's proficiency level in these activities. The most obvious way in which a new technology for cognitive assessment might have an impact is in improving present selection and classification systems. In this paper, some of the most recent attempts to explore issues related to the feasibility of new measurement methods were reviewed. It appears that individual differences on elementary cognitive tasks are generally substantial, and there is evidence that such differences are not being captured by conventional ability tests. This suggests a role for cognitive tasks as supplements to conventional tests, and it may be that they will be particularly valid performance predictors for specialized occupations that require particular kinds of psychological processing. A second possible way in which current selection and classification systems might be supplemented is with measures that directly assess changes in processing efficiency as a function of practice. Because the form of such changes determines an examinee's expected performance level at the end of training or practice, it is reasonable to expect that current test batteries, which assess an examinee's current state of knowledge, can be profitably augmented by including measures of changes in processing efficiency.

22

There has been much discussion recently about what possibilities microcomputers hold for changing the way assessment is accomplished. Much of this discussion centers on adaptive testing technology and computerized versions of existing aptitude tests (Moreno, Wetzel, McBride, & Weiss, 1983; Weiss, 1983), but more recently, attention has increasingly turned toward the issue of whether new abilities can now be measured (Belmont, 1983; Hunt, 1982). In a thoughtful review, Hunt and Pellegrino (1984) have discussed changes that computerized testing can bring in assessment of the traditional spatial, verbal, and reasoning abilities, as well as in the previously unmeasured abilities related to learning, attention, and psychomotor skills, and thus their report might be read as a supplement to the views expressed in this paper. Hunt (1982) concluded, in his earlier paper, that although it is not yet practically feasible, a concerted 5-year research program that aimed to consolidate cognitive measurement techniques and explore their utility as intelligence tests might have important long-term benefits in leading to increased predicitive validity of aptitude batteries. This seems to sum up the current state of the art in ability measurement: The new cognitive assessment techniques show promise, but considerable extra research effort will be required before such tests will be feasible for personnel decision-making purposes in operational settings. And further, true significant strides in ability measurement applications await further developments in establishing adequate criterion measures. One particularly promising area in this regard is the use of computerized instructional environments to serve as test beds in validation research.

An area that potentially may more quickly benefit from new forms of cognitive assessment has to do with the identification of basic skills. Two studies were reviewed here that demonstrated how a careful task analysis followed by a comparison of performers at various skill levels can lead to the identification of component skills. Skills identified in such a fashion tend to be less general than the traditional skills of reading, writing, and arithmetic, and by virtue of the specificity of such skills, cognitive diagnosis is more easily accomplished and prescriptions for training specific deficits more naturally result. Here again, the general approach is only beginning to be explored, but given the tremendous cost of training, the benefits of such an approach are likely to be realized in the near future, and diverse application efforts are likely to be seen.

In sum, cognitive assessment of the type that has been the main focus of this paper is a promising technology, but one that is not yet ready to be applied in the workplace. As the cost of already fairly inexpensive microcomputers comes down even further, while at the same time applied research programs provide more and more demonstrations of the utility of new and diverse forms of cognitive assessment, some of the best ideas in this field are likely to be transferred to practical applications. Such a move will significantly expand the ideas on where, how, and for what purpose people's cognitive capabilities should be measured.

23

REFERENCES

Ackerman, P.L. (1984). A theoretical and empirical investigation of individual differences in learning: A synthesis of cognitive ability and information processing perspectives (Doctoral Dissertation, University of Illinois, 1983). Dissertation Abstracts International, 45, 1, 381-B.

Ackerman, P.L., & Schneider, W. (1984, August). Individual differences in automatic and controlled processes (Tech. Rep. No. HARL-ONR-8401). Champagne: University of Illinois, Department of Psychology.

Allen, G.A., Pitts, E.W., Jones, R.J., & Morgan, B.B., Jr. (1985). Measures of learning rate, level of learning, and scholastic aptitude as predictors of performance in training-oriented academic courses (Tech. Rep. No. ITR-85-32). Norfolk, VA: Old Dominion University, Center for Applied Psychological Studies.

Allen, G.A., Salas, E., Pitts, E.W., Terranova, M., & Morgan, B.B., Jr. (1984). A comparison of cognitive ability factors, learning level measures, and learning rate measures as predictors of learning performance (Tech. Rep. No. ITR-84-31). Norfolk, VA: Old Dominion University, Center for Applied Psychological Studies.

Allen, G.A., Secundo, M.D., Salas, E., & Morgan, B.B., Jr. (1983). Evaluation of rate parameters as predictors of the acquisition, decay, and reacquisition of complex cognitive skills (Tech. Rep. No. ITR-82-27). Norfolk, VA: Old Dominion University, Center for Applied Psychological Studies.

Belmont, J.M. (1983). Concerning Hunt's new ways of assessing intelligence. Intelligence, 7, 1-7.

Bond, L., Eastman, R., Gitomer, D., Glaser, R., & Lesgold, A. (1983, December). Cognitive task analysis of technical skills: Rationale and approach. Unpublished manuscript. University of Pittsburgh, Learning Research and Development Center.

Burt, C. (1940). The factors of the mind. London: University of London Press.

Carroll, J.B. (1980). Individual difference relations in psychometric and experimental cognitive tasks (Tech. Rep. No. 163). Chapel Hill: University of North Carolina, The L.L. Thurstone Psychometric Laboratory.

Carter, R.C., & Krause, M. (1983). Unreliability of slope scores. Unpublished manuscript. New Orleans: Naval Biodynamics Laboratory.

Cattell, R.B. (1971). Abilities: Their structure growth and action. Boston: Houghton Mifflin.

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

Chi, M.C., Glaser, R., & Rees, E. (1981). Expertise in problem solving. In R.J. Sternberg (Ed.), Advances in the psychology of human intelligence (Vol 1, pp. 7-76). Hillsdale, NJ: Lawrence Erlbaum Associates.

Christal, R.E. (1976). What is the value of aptitude tests? Paper presented at the 18th Annual Conference of the Military Testing Association, Gulf Shores, AL.

Christal, R.E. (1981). The need for laboratory research to improve the state-of-the-art in ability testing. Paper presented at the National Security Industrial Association-DoD Conference on Personnel and Training Factors in Systems Effectiveness, San Diego, CA.

Christal, R.E., Tirre, W.C., & Kyllonen, P.C. (1984, April). Two for the money: Speed and level scores from a computerized vocabulary test. Paper presented at the Ninth Psychology in the DoD Symposium, USAF Academy, Colorado Springs, CO.

Cronbach, L.J. (1957). The two disciplines of scientific psychology. American Psychologist, 12, 671-684.

Cronbach, L.J., & Snow, R.E. (1977). Aptitudes and instructional methods: A handbook for research on interactions. New York: Irvington.

Dailey, J. (1948, November). Development of the Airmen Classification Test Battery (Research Bulletin 48-4). Lackland AFB, TX: Air Training Command.

Department of Defense. (1984). Test manual for the Armed Services Vocational Aptitude Battery (DoD 1340.12AA). North Chicago, IL: US Military Entrance Processing Command.

Eaton, N.K., & Shields, J. (1985). Validating selection tests for job performance. In J. Zeidner (Ed.), Human productivity enhancement: Vol. 2. Acquisition and development of personnel. New York: Praeger.

Estes, W.K. (1974). Learning theory and intelligence. American Psychologist, 29, 740-749.

Fairbank, B., Tirre, W., & Anderson, N.A. (1984, December). The selection, administration, and analysis of 30 cognitive tests of individual differences. Paper presented at the NATO Conference on Learning and Cognition, Athens, Greece.

Fisk, A.D., & Schneider, W. (1983). Category and word search: Generalizing search principles to complex processing. Journal of Experimental Psychology: Learning, Memory, and Cognition, 9, 177-195.

Fleishman, E.A., & Hempel, W.E. (1954). Changes in factor structure of a complex psychomotor task as a function of practice. Psychometrika, 19, 232-252.

Fredericksen, J.R. (1981). Sources of process interaction in reading. In A.M Lesgold & C.A. Perfetti (Eds.), Interactive processes in reading. Hillsdale, NJ: Lawrence Erlbaum Associates.

Fredericksen, J.R. (1982). A componential theory of reading skills and their interactions. In R.J. Sternberg (Ed.), Advances in the psychology of human intelligence (Vol. 1, pp. 125-180). Hillsdale, NJ: Lawrence Erlbaum Associates.

Fredericksen, J.R., Weaver, P.A., Warren, B.M., Gillotte, H.P., Rosebery, A.S., Freeman, B., & Goodman, L. (1983, March). A componential approach to training reading skills: Final report (Tech. Rep. No. 5295). Cambridge, MA: Bolt, Beranek, and Newman, Inc.

Gilbert, T.F. (1976). The high cost of knowledge. Personnel, 53(2), 11-23.

Gitomer, D. (1984). A cognitive analysis of a complex troubleshooting task. Unpublished doctoral dissertation, University of Pittsburgh, Pittsburgh, PA.

Gott, S., & Davis, T. (1983, November). <u>Introducing specific knowledge domains into basic skills instruction: From generalized powers to specified knowledge</u>. Paper presented at the National Adult Education Conference, Philadelphia, PA.

Gould, R.B., & Hedges, J. (1983, August). <u>Air Force job performance criterion development</u>. Paper presented at the American Psychological Association Annual Convention, Anaheim, CA.

Gould, S.J. (1982). <u>The mismeasure of man</u>. New York: Norton.

Gragg, D.B., & Gordon, M.A. (1950). <u>Validity of the Airman Classification Battery AC-1</u> (Research Bulletin 50-3, ATI 093 440). Lackland AFB, TX: Air Training Command, Human Resources Research Center.

Guilford, J.P. (1982). Cognitive psychology's ambiguities: Some suggested remedies. <u>Psychological Review, 89</u>, 48-59.

Guilford, J.P., & Lacey, J.I. (Eds.). (1946). <u>Printed Classification Tests: Army Air Forces Aviation Psychology Research Program Reports No. 5</u>. Washington, DC: Government Printing Office.

Gustaffson, J.E. (1984). A unifying model for the structure of intellectual abilities. <u>Intelligence, 8</u>, 179-203.

Holtz, B., & Schmitz, E. (1985). Technology for person-job matching. In J. Zeidner (Ed.), <u>Human productivity enhancement: Vol. 2. Acquisition and development of personnel</u>. New York: Praeger.

Hopkin, V.D. (1982). <u>Human Factors in Air Traffic Control</u> (North Atlantic Treaty Organization, Advisory Group for Aerospace Research and Development, AGARD-AG-275). Farnborough, Hampshire, UK: Royal Air Force Institute of Aviation Medicine. (DTIC AD-A116 394)

Horn, J.L. (1968). Organization of abilities and the development of intelligence. <u>Psychological Review, 75</u>, 242-259.

Hunt, E.B. (1978). Mechanics of verbal ability. <u>Psychological Review, 85</u>, 109-130.

Hunt, E.B. (1982). Toward new ways of assessing intelligence. <u>Intelligence, 6</u>, 231-240.

Hunt, E.B., Frost, N., & Lunneborg, C.E. (1973). Individual differences in cognition: A new approach to intelligence. In G. Bower (Ed.), <u>The psychology of learning and motivation: Advances in research and theory</u> (Vol. 7). New York: Academic Press.

Hunt, E.B., & Pellegrino, J.W. (1984). <u>Using interactive computing to expand intelligence testing: A critique and prospectus</u> (Tech. Rep. No. 84-2). Seattle, WA: University of Washington, Department of Psychology.

Kahneman, D. (1973). <u>Attention and effort</u>. Englewood Cliffs, NJ: Prentice-Hall.

Kennedy, R.S., Bittner, A.C., Carter, R.C., Krause, M., Harbeson, N.M., McCafferty, D.B., Pepper, R.L., & Wiker, S.F. (1981). <u>Performance evaluation tests for environmental research (PETER): Collected papers</u> (Tech. Rep. No. NBDL-80R008). New Orleans, LA: Naval Biodynamics Laboratory.

Kennedy, R.S., Bittner, A.C., Harbeson, N.M., & Jones, M.I. (1981). <u>Perspectives in performance evaluation tests for environmental research (PETER): Collected papers</u>. (Tech. Rep. No. NBDL-80R004). New Orleans, LA: Naval Biodynamics Laboratory.

Kyllonen, P.C. (1985a). Individual differences in learning and cognition: The four-source proposal. Unpublished manuscript. Brooks AFB, TX: Air Force Human Resources Laboratory.

Kyllonen, P.C. (1985b). Dimensions of information processing speed (AFHRL-TP-84-56, AD-A154 778). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Kyllonen, P.C., Tirre, W., & Christal, R.E. (1984). Processing determinants of associative learning. Paper presented at the Psychonomic Society, San Antonio, TX.

Larkin, J.H., McDermott, J., Simon, D.P., & Simon, H.A. (1980). Models of competence in solving physics problems. Cognitive Science, 4, 317-345.

Lewis, C.H. (1980). Speed and Practice. Unpublished manuscript. Pittsburgh, PA: Carnegie Mellon University, Department of Psychology.

Moreno, K.E., Wetzel, C.D., McBride, J.R., & Weiss, D.J. (1983, August). Relationship between corresponding Armed Services Vocational Aptitude Battery (ASVAB) and Computerized Adaptive Testing (CAT) subtests. (Tech. Rep. No. NPRDC-TR-83-27). San Diego, CA: Navy Personnel Research and Development Center.

Nairn, A., & Associates. (1980). The reign of ETS: The corporation that makes up minds. The Ralph Nader Report on the Educational Testing Service. Washington, DC: Ralph Nader.

Neisser, U. (1967). Cognitive psychology. New York: Appleton.

Newell, A., & Rosenbloom, P. (1981). Mechanisms of skill acquisition and the law of practice. In J.R. Anderson (Ed.), Cognitive skills and their acquisition (pp 1-55). Hillsdale, NJ: Lawrence Erlbaum Associates.

Payne, D., & Tirre, W. (1984). Individual differences in learning rate. Paper presented at the Ninth Psychology in the DoD Symposium, USAF Academy, Colorado Springs, CO.

Pellegrino, J.W. (1984). Individual differences in information processing efficiency. Paper presented at the American Psychological Association Annual Convention, Toronto, Canada.

Pellegrino, J.W., & Glaser, R. (1979). Cognitive correlates and components in the analysis of individual differences. Intelligence, 3, 187-214.

Posner, M.I., & Mitchell, R.E. (1967). Chronometric analysis of classification. Psychological Review, 74, 392-409.

Rimland, B., & Larson, G.E. (in press). Individual difference: An undeveloped opportunity for military psychology. Military Psychology.

Rogosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. Psychological Bulletin, 90, 726-748.

Rogosa, D., & Willett, J. (1983). Demonstrating the reliability of the difference score in the measurement of change. Journal of Educational Measurement, 20, 335-343.

Rose, A.M., & Fernandez, K. (1977, November). An information processing approach to performance assessment: I. Experimental investigation of an information processing performance battery (Tech. Rep. No. AIR-58500-TR). Washington, DC: American Institutes for Research.

Rumelhart, D., & Norman, D. (1978). Accretion, tuning, and restructuring: Three modes of learning. In J.W. Cotton & R. Klatzky (Eds.), Semantic factors in cognition. Hillsdale, NJ: Lawrence Erlbaum Associates.

Shiffrin, R.M, & Schneider, W. (1977). Controlled and automatic human informaton processing. II. Perceptual learning, automatic attending, and a general theory. Psychological Review, 84, 127-190.

Snow, R.E. (1978). Theory and method for research on aptitude processes. Intelligence, 2, 225-278.

Snow, R.E. (1980). Aptitude processes. In R.E. Snow, P.A. Federico, & W.E. Montague (Eds.), Aptitude, learning, and instruction: Vol. 1. Cognitive process analyses of aptitude. Hillsdale, NJ: Lawrence Erlbaum Associates.

Snow, R.E., & Lohman, D.F. (1981). Cognitive and learning in young adults (Tech. Rep. No. 13). Stanford, CA: Stanford University, Aptitude Research Project, School of Education.

Snow, R.E., Wescourt, K., & Collins, J.C. (1983). Individual differences in aptitude and learning from interactive computer-based instruction (Tech. Rep. No. 10). Stanford, CA: Stanford University, Aptitude Research Project, School of Education.

Spearman, C. (1905). General intelligence objectively determined and measured. American Journal of Psychology, 15, 201-293.

Sternberg, R.J. (1977). Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities. Hillsdale, NJ: Lawrence Erlbaum Associates.

Sternberg, R.J. (1980). Nothing fails like success: The search for an intelligent paradigm for studying intelligence. Journal of Educational Psychology, 73, 142-155.

Sternberg, S. (1969). Memory-scanning: Mental processes revealed by reaction-time experiments. American Scientist, 57, 421-457.

Thurstone, L.L. (1938). Primary mental abilities. Chicago: University of Chicago Press.

Vernon, P.E. (1961). The structure of human abilities (2nd ed.). London: Methuen.

Weeks, J.L., Mullins, C.J., & Vitola, B.M. (1975, December). Airman classification batteries from 1948 to 1975: A review and evaluation (AFHRL-TR-75-78, AD-A026 470). Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory.

Weiss, D.J. (Ed.). (1983). New horizons in testing: Latent trait test theory and computerized adaptive testing. New York: Academic Press.

Wilson, K.M. (1982, October). A study of the validity of the restructured GRE aptitude test for predicting first-year performance in graduate study (GRE Board Research Report GREB No. 78-6R, ETS Research Report 82-34). Princeton, NJ: Educational Testing Service.

END

FILMED

3 - 86

DTIC