2

Navy Personnel Research and Development Center

San Diego, CA 92152-6800 TR 89-6 February 1989

Development and Evaluation of Integrating Details: A Complex Spatial Problem Solving Test

Approved for public release, distribution is unlimited.

NPRDC TR 89-6

February 1989

Development and Evaluation of Integrating Details: A Complex Spatial Problem Solving Test

David L. Alderton, Ph.D.

Reviewed by
Robert F. Morrison, Ph.D.

Approved by
John J. Pass, Ph.D.

Released by
B. E. Bacon
Captain, U.S. Navy
Commanding Officer

and

J. S. McMichael, Ph.D.
Technical Director

Navy Personnel Research and Development Center
San Diego, California 92152-6800

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

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 is unlimited.
2b DECLASSIFICATION / DOWNGRADING SCHEDULE	
4 PERFORMING ORGANIZATION REPORT NUMBER(S) NPRDC TR 89-6	5 MONITORING ORGANIZATION REPORT NUMBER(S)
6a NAME OF PERFORMING ORGANIZATION Navy Personnel Research and Development Center 6b OFFICE SYMBOL *(If applicable)* Code 122	7a NAME OF MONITORING ORGANIZATION
6c ADDRESS *(City, State, and ZIP Code)* San Diego, California 92152-6800	7b ADDRESS *(City, State, and ZIP Code)*
8a NAME OF FUNDING / SPONSORING ORGANIZATION Chief of Naval Research Assistant Secretary of Defense 8b OFFICE SYMBOL *(If applicable)* Code 222 (EM&P/MN&PP)	9 PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c ADDRESS *(City, State and ZIP Code)* 800 N. Quincy Street, Arlington, VA 22217-500 The Pentagon, Washington, DC 20301	10 SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO	PROJECT NO	TASK NO	WORK UNIT ACCESSION NO
62233N	RM33M20		RM33M20.03

11 TITLE *(Include Security Classification)*

Development and Evaluation of Integrating Details: A Complex Spatial Problem Solving Test

12 PERSONAL AUTHOR(S)

David L. Alderton

13a TYPE OF REPORT Technical	13b TIME COVERED FROM Jan 86 TO Jun 88	14 DATE OF REPORT *(Year, Month, Day)* 1989 February	15 PAGE COUNT 46

16 SUPPLEMENTARY NOTATION

17 COSATI CODES		
FIELD	GROUP	SUB-GROUP
~~05~~	~~09~~	

18 SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*

Testing, spatial ability, mental imagery, componential analysis, ASVAB

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

This report summarizes the theory behind and the development, evaluation, and refinement of a complex spatial processing test, Integrating Details. In the course of revamping and making the Armed Services Vocational Aptitude Battery (ASVAB) adaptive, the opportunity exists for replacing or adding tests to the battery. Tests of spatial ability, a major dimension of human intelligence, are not represented in the ASVAB so they are reasonable candidates for inclusion. The vast psychometric literature on spatial ability suggests that only complex spatial tests are likely to be valid for both school and job performance. Theoretical and empirical work from visual cognition and mental imagery provided further guidelines for the development of a complex spatial test. Methodological techniques from componential analysis were used to instantiate the test and develop an information processing model for test item solution. The results suggest that Integrating Details is a complex spatial problem solving test, that it is relatively independent of verbal ability, that the test is reliable and has substantial construct

20 DISTRIBUTION / AVAILABILITY OF ABSTRACT ☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS	21 ABSTRACT SECURITY CLASSIFICATION UNCLASSIFIED	
22a NAME OF RESPONSIBLE INDIVIDUAL David L. Alderton	22b TELEPHONE *(Include Area Code)* (619) 553-7647	22c OFFICE SYMBOL Code 122

83 APR edition may be used until exhausted
All other editions are obsolete

SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED

validity. Furthermore, analyses demonstrate that the ability measured by Integrating Details is substantially unique from those measured by the ASVAB and thus there is ample opportunity for the test to augment the predictive validity of the ASVAB. A final 40-item version of the test is recommended for advanced development and validation with military enlisted personnel.

FOREWORD

This report describes an investigation of a complex spatial problem solving test conducted under the Personnel Performance Prediction (PPP) project (Work Unit No. 62233N RM33M20.03), which is sponsored by the Office of the Chief of Naval Research (Code 222) and the Office of the Assistant Secretary of Defense (Force Management and Planning/MN&PP). The objective of the PPP project was to develop and apply information processing methods and models to aptitude constructs and to evaluate their potential for computerized testing. Other PPP projects have investigated paragraph and mechanical comprehension.

The current research was initiated under a contract to Dr. Earl Hunt at the University of Washington (Contract No. N66001-85-C-0017), and part of it was subcontracted to Dr. James Pellegrino of the University of California, Santa Barbara. The purpose of the initial contract was to develop computer-based tests of spatial-visual ability to be used in further research by the Navy as possible classification tests. It is believed that spatial ability tests, which are not presently represented among military assignment tests, may improve the assignment of enlisted personnel to selected technical ratings. This report is an evaluation of one of the tests developed under the original contract: the Integrating Details test. The report summarizes the original research on this test and some of the subsequent research performed at the Navy Personnel Research and Development Center.

This report describes the development, evaluation, and suggested refinement of the Integrating Details test and will serve as the basis for further work. This report is required since there are plans to use the test as an experimental predictor of school and job performance in several projects sponsored by the Department of Defense (Office of the Assistant Secretary of Defense, Force Management and Personnel).

B. E. BACON
Captain, U.S. Navy
Commanding Officer

J. S. McMICHAEL
Technical Director

SUMMARY

Problem

In the context of making the Armed Services Vocational Aptitude Battery (ASVAB) adaptive and computer-administered, efforts have been made to expand the abilities measured by the current subtests. One major shortcoming of the ASVAB is that it contains no measure of spatial ability. Although a measure of spatial ability is an obvious candidate for including in the ASVAB, there are hundreds of tests of spatial aptitude from which to choose. A better understanding of the domain and range of spatial tests and a theory of spatial information processing is needed to guide the development and selection of a spatial test.

Purpose

This report summarizes the theory behind and development of a complex spatial problem solving test, *Integrating Details*. This report describes in detail the psychological and psychometric properties of the test and documents its evolution. In the end, a final version of the test is recommended for subsequent research on Navy enlisted personnel and in joint-service job performance measurement projects.

Approach

Literature reviews were used to establish guidelines for the selection and development of a spatial test that had a high probability of being valid in use with Navy enlisted personnel and one with a sound theoretical basis. Following the development of *Integrating Details*, combinations of psychometric and modeling techniques were used to evaluate and alter the test.

Results

The review of the paper-and-pencil spatial test literature revealed that complex spatial tests administered under relatively untimed conditions were the most valid predictors of both school and job success. Subsequent reviews of the experimental literature on visual cognition, imagery, and spatial information processing provided the outline of a theory for understanding performance on spatial problem solving tests. Combining the conclusions from the literature reviews gave pragmatic and theoretical direction for the development of a complex spatial test, *Integrating Details*. Following initial development, the test was used on a number of samples but performance data suggested many changes.

The psychometric properties of the test are summarized and found to be quite good. Detailed evaluations of the construct validity of the test are also conducted. These evaluations provide substantial support for the information processing model of item solution that guided item development and selection. Evaluations of the pattern of correlations between *Integrating Details* and other tests substantially supported the theoretical contention that the test assess complex spatial problem solving skills that are independent of verbal skills and also substantially unique from the ASVAB.

Recommendations

The changes incorporated into *Integrating Details* and documented herein have made substantial improvements in the psychometric properties of the test. The improved version of *Integrating Details* is recommended for advanced validation work.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Problem

In the context of making the Armed Services Vocational Aptitude Battery (ASVAB) adaptive and computer-administered, efforts have been made to expand the abilities measured by the current subtests. One obvious shortcoming of the ASVAB is that it contains no measure of spatial ability. Although a measure of spatial ability is an obvious candidate for inclusion in the ASVAB, there are hundreds of tests of spatial aptitude from which to choose. A better understanding of the domain and range of spatial tests and a theory of spatial information processing is needed to guide the development and selection of a spatial test.

Purpose

This report summarizes the theory behind and development of a complex spatial problem solving test, Integrating Details. The report describes in detail the psychological and psychometric properties of the test and documents its evolution. In the end, a final version of the test is recommended for subsequent research on Navy enlisted personnel and in joint-service job performance measurement projects.

Background

The following three sections, Spatial Test Domain, Spatial Information Processing, and Componential Analyses of Spatial Tests, outline the work and thought that influenced the selection and development of Integrating Details.

Spatial Test Domain

Except for the early work of Spearman (1904), which claimed that only a single general intelligence (g) existed, virtually all theories of intelligence based on the intercorrelations of test scores have hypothesized at least two distinct ability classes, verbal and spatial (Burt, 1949a, b; Cattel, 1971; Kelly, 1928; Thurstone, 1938; Vernon, 1950). Following the general acceptance of a distinct spatial ability, there began, especially among American psychologists, a search to discover the number of distinct facets or factors of spatial ability. Between 1928 and 1980, the number of reported spatial factors increased from 2 to over 30. Along with the growth of spatial factors came such an explosion in the number of tests of spatial ability that Eliot and Smith (1983) published An international directory of spatial tests containing reviews of 392 tests!

In an attempt to organize the vast literature of spatial ability, several investigators have reviewed much of the available factor analytic data (Lohman, 1979; McGee, 1979; Smith, 1964). The best summary of the correlational literature (Lohman, 1979) concludes that there is evidence for 11 different spatial factors (but see, Hunt, Pellegrino, Frick, Farr, & Alderton, 1988). However, a more meaningful understanding of the relationship between spatial tests (and factors) is gained by seriating tests along two related dimensions, speed and complexity (Lohman, 1979; Lohman, Pellegrino, Alderton, & Regian, 1987; Pellegrino, Alderton, & Shute, 1984; Pellegrino & Kail, 1982). Figure 1 contains sample items from six spatial tests, which exemplify this distinction. The first dimension reflects the classic distinction between speed and power tests (Cronbach, 1970). Speeded tests emphasize the number of items that can be solved per unit time while power tests attempt to determine how difficult an item an individual can solve. As one

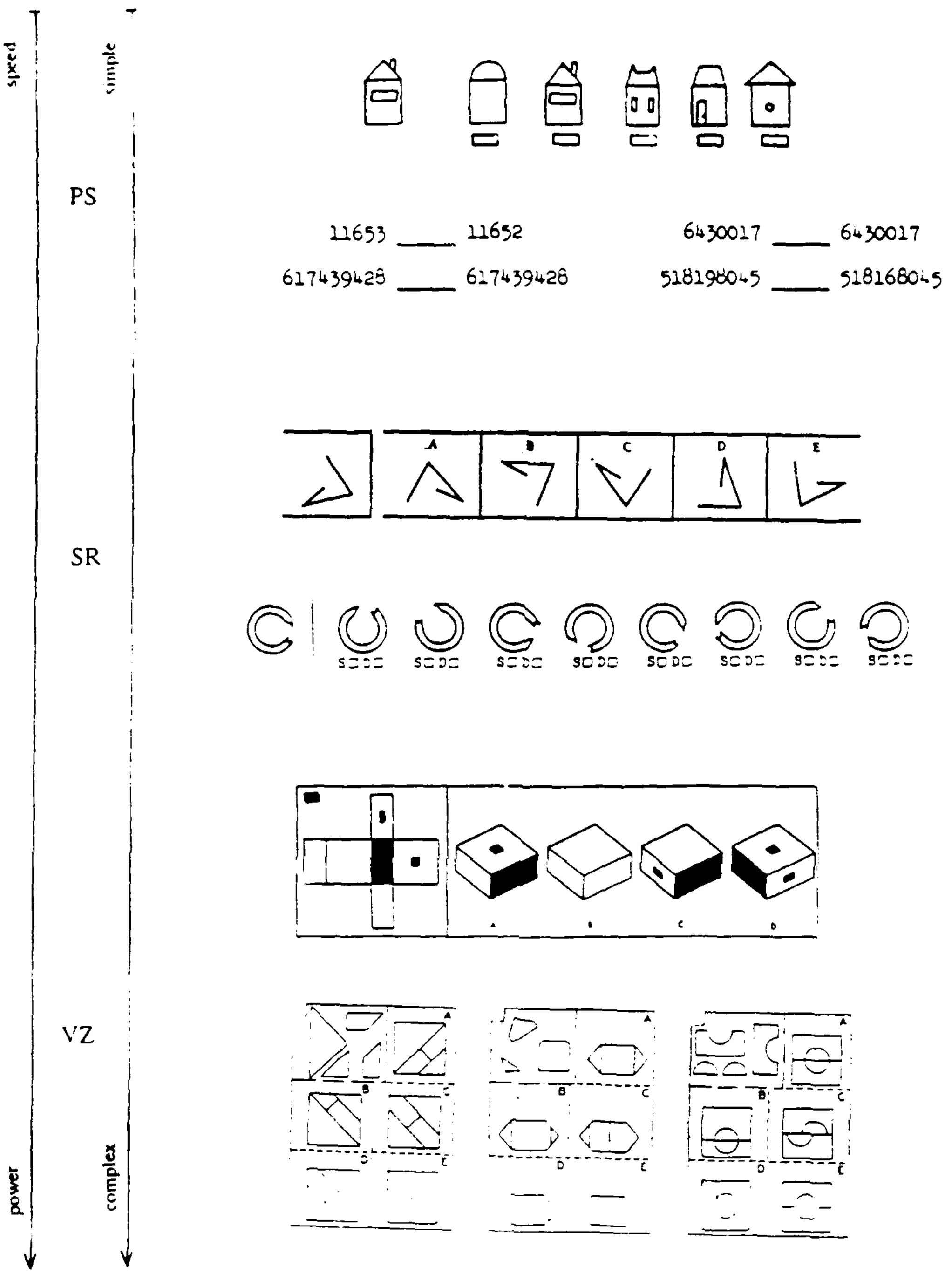


Figure 1. Spatial test sample items and the speededness of complexity continua.

moves from the top of the Figure toward the bottom, the test emphasis shifts from speed to power. A rough index of speededness is the number of items that must be solved per unit time. The first group of two tests, defining the Perceptual Speed (PS) factor, requires the solution of 32 items per minute. The next two tests, defining the Spatial Relations (SR) factor, requires that 17.7 items be solved per minute. The final two tests, defining the Visualization (VZ) factor, only require the examinee to solve 2.7 items per minute. The second dimension is item complexity. As one moves toward the bottom of the Figure, the items become increasingly more complex. In general, these are nearly parallel continua. Tests that emphasize speeded item solution typically employ very simple items that produce few errors but are administered under strict time limits; the principal source of individual differences is the speed of item solution. In contrast, tests that are relatively unspeeded tend to employ complex items, errors are frequent but time limits are generous; the principal source of individual differences is the accuracy of item solution.

Using the speed-power and item complexity dimensions as a heuristic for ordering test intercorrelations, produces a matrix that is arrayed as a simplex (Guttman, 1954). This has been done for the six tests from Figure 1 and is presented in Table 1. In a simplex, the largest column correlations are the immediate off-diagonal entries, and the magnitude of the remaining column entries steadily decrease the more removed it is from the main diagonal. Furthermore, moving from left to right, that is, from speed toward power tests and from simple toward complex items, there is a general increase in the magnitude of the correlations. This general increase in correlations suggests that there is a corresponding increase in the communality (shared variance) among the tests.

Table 1

Spatial Test Intercorrelations

		1	2	3	4	5	6
1.	Identical Pictures	--					
2.	CAB-Perceptual	.389	--				
3.	CAB-Spatial	.295	.381	--			
4.	PMA-Space	.251	.261	.671	--		
5.	DAT-Space	.247	.260	.410	.571	--	
6.	MPFB	.218	.254	.367	.450	.720	--
	Raven's	.19	--	.28	.32	.54	.62

In addition to this simplex pattern, once tests are arranged by speededness and item complexity, two other facts were revealed. First, tests using more complex items with an emphasis on solution accuracy are not only more correlated with other spatial tests, but they are also more correlated with other tests in general. The bottom row of Table 1 contains the correlation of each test with Raven's Advanced Progressive Matrices (APM) (Raven, 1962). Raven's is consensually considered to be a quintessential measure of general reasoning ability and can be used as a surrogate measure of IQ. As can be seen, as the spatial tests become less speeded and employ more complex items, the correlation with APM greatly increases.

The greater communality of the power tests and the greater correlation between power tests and general reasoning ability foreshadows the second fact, which is not apparent from the correlation matrix. As you move from left to right, that is, from speed toward power tests, the magnitude and breadth of the test's validity increases. For the simple (perceptual) speed tests (the first two tests in Figure 1 and Table 1) there is very little validity data outside of clerical occupations. For the spatial relations tests (the middle two tests in Figure 1 and Table 1) there is somewhat more validity evidence but the results are inconsistent. In contrast, the validity of power tests, those employing complex items, is vast and well documented. Table 2 contains a summary of the classroom validity data culled from the manuals of the Differential Aptitude Test battery's Space test (DAT) (Bennett, Seashore, & Wesman, 1974) and the Minnesota Paper Form Board (MPFB) (Likert & Quasha, 1970) (the last two tests in Figure 1 and Table 1). The first panel shows course titles common to both tests. The second panel shows additional classroom course criteria. The third panel shows course titles where the tests have demonstrated significant predictive validity. These titles represent a very broad spectrum of technical, academic, and professional classroom criteria.

Table 2

Course Work Validities of Spatial Tests

Common Course Grade Criteria

MPFB	Course	DAT
.30	Drafting	.32
.31	Geometry	.29
.30	Art	.32

Additional Course Criteria

Course	MPFB	DAT	Course
Electrical	.43	.34	Math
Engineering	.34	.37	Trigonometry
Topography	.53	.37	Cosmetology
Dentistry	.61	.47	Data Processing

Miscellaneous Grade Criteria

Architecture	Automotive Mechanics	Aviation Mechanics
Biology	Bricklaying	Carpentry
Chemistry	Engineering	Design Graphics
Metal Work	Physics	Printing
Text Editing	Watch Repair	Woodwork

Separate from the classroom criteria, complex (power oriented) spatial tests also have a proven history of validity against job success. Table 3 reports job titles and job categories where complex spatial tests have demonstrated significant predictive validity. The first panel lists job titles with significant spatial test validities; these were aggregated from military and industrial sources. The second panel shows the nine job categories used by Ghiselli (1973) in his summary of occupational validity. Spatial tests were found to be valid for training success and job proficiency in at least some jobs in each of these categories. The bottom panel of Table 3 contains the four occupational

Table 3

Jobs Requiring Spatial Ability

Jobs with Significant Spatial Test Validities	
Aircraft Inspectors	Aircraft Work Supervisors
Air Traffic Controllers	Automobile Mechanics
Bombardiers	Carpenters
Electrical Engineers	Electrical Scientists
Gun-Boat Operators	Motor-Launch Operators
Navigators	Saw Operators (Mills)
Sewing Machine Operators	Signalmen
Pharmaceutical Packers	Plumbers
Pilots	Dental Assistants

Job Categories with Significant Spatial Test Validities		
Managerial	Clerical	Sales
Protective	Services	Vehicle Operators
Trades and Crafts	Industrial	Metalworking

Jobs Categories Requiring Spatial Ability Above 90 Percent			
Engineers	Scientists	Draftsmen	Designers

categories that the United States Employment Service estimates as requiring spatial ability in the top 10 percent of the general population (adapted from Smith, 1964; McGee, 1979; original source, United States Employment Service, 1957). These are listed in order of the number of jobs within the category.

In summary, contrasting the relative speededness and item complexity of spatial tests is a useful heuristic for organizing and understanding the plethora of existing tests. Tests that are relatively speeded use simple items presented under stringent time constraints, errors are few; thus, individual differences are localized in the rate of item solution. In contrast, power tests use very complex items administered with generous or no time limits, errors are frequent; therefore, individual differences are most evident in how complex an item an individual can solve. Organizing test intercorrelations using the speed-complexity heuristic reveals a simplex in which complex power tests are more generally correlated with other spatial tests and with other tests in general. Complex spatial tests are the most predictively valid tests in the spatial domain. Indeed, complex spatial tests are significantly related to standard academic course grades, vocational course work, and job success across the entire occupational spectrum. Complex spatial tests are correlated with tests of general reasoning (e.g., Raven's APM), which accounts in part for the broad spectrum predictive validity of these tests. However, it is the spatial requirements of these tests that provides the predictive advantage of spatial tests against some course and job criteria.

The speed-power heuristic provides a categorization of spatial tests and focuses attention on the validity of power tests. However, what is lacking is a rigorous theory of spatial information processing and individual differences in spatial ability that can explain item difficulty, test intercorrelations, and test validity. It is this sort of information that is needed to provide guidance for the development of systematic items and tests that will be valid. The recent advances in spatial cognition, imagery theory, and componential analyses of spatial tests that provide this needed direction are discussed below.

Spatial Information Processing

Spatial information processing is central to normative theories of human cognition (e.g., Anderson, 1983) and individual difference theories of human intelligence (e.g., Vernon, 1950). Spatial information processing theories posit a continuum of structures and processes from visual sensation to object representation in a visual short-term buffer to ultimate long-term storage in a representational medium that preserves the configural relationships among elements in the perceived spatial array (Anderson, 1983; Chase, 1986; Kosslyn, 1981; Marr, 1982; Olson & Bialystok, 1983; Pinker, 1984). It is this specialized representational format and the associated processes required to operate on this format that gives rise to the distinction between verbal (propositional) and spatial (analog) ability classes long noted in individual difference theories of human intelligence. While the study of visual cognition has focused on the structures and processes employed to perceive and store spatial information, imagery research has (usually) emphasized the reverse; that is, once information has been stored, how it is reactivated and utilized internally in the absence of any perceptual stimulation to make judgements about spatial relationships. The two lines of research compliment one another and there is substantial experimental evidence that the same structures and processes underpin both activities (Farah, 1985; Finke, 1980, 1985; Finke & Shepard, 1986; Kosslyn, 1980; Pinker, 1984; Shepard, 1975, 1981, 1984) as well as neuropsychological data (Farah, 1984; Kosslyn, 1985; Kosslyn, Holtzman, Farah, & Gazzaniga, 1985; Sacks, 1985). That is, the same structures and processes that allow us to sense, perceive, represent, and remember the external spatial world support the ability to generate, inspect, and operate on self-initiated internal representations.[1]

[1] For the sake of completion, it is believed that through evolution the ecological constraints on the spatial environment now mirror the constraints on the visual information processing system, including the processes, structures, and representational format. Specifically, it can be argued that through evolution the most enduring constraints of the external world have become incorporated into our basic perceptual makeup. According to Shepard (1981):

> Especially basic are constraints conditioned by such facts as that space is locally Euclidean and three-dimensional and that significant objects, including our own bodies: (1) are bound by two-dimensional surfaces, (2) tend to conserve semirigid shape, (3) have exactly six degrees-of-freedom of overall position in space, and (4) tend, over time, to move between nearby positions according to a principle of least action. (p. 279)

The internal representation of the external spatial world consists of two structures and a set of processes to operate on them (Anderson, 1983; Kosslyn, 1980, 1981; Pinker, 1984; Finke, 1980, 1985).[2] The stable, long-term memorial representation that can be used for generating images consist of both depictive and descriptive data. Depictive data are literal encodings of objects that retain the information appropriate and necessary for generating an image. The descriptive or propositional data concerns how an object looks, including, the parts an object has, the location of its parts, the size category of a part or object, an abstract description of a part or object, the name of the object's superordinate category, and the list of depictive or literal encodings of the appearance of an object. This deep structure representation is accessed and activated in logical units or modules, thus making it possible to generate images only of parts of objects as well as whole objects. This deep or long term representation is used to generate a surface representation that has a particular surface structure, which is an active internal image. The surface representation or surface structure is a quasi-pictorial representation that occurs in a spatial medium that depicts an object or scene and fundamentally underlies the conscious experience of imagery. The surface image is projected or displayed in a medium referred to as the visual buffer. The properties of the surface image or surface structure representation is in large part a consequence of the properties of the visual buffer that supports the representation. The surface structure functions as an Euclidean coordinate space and it is viewer-centered. Further, the surface structure has a limited resolution or grain which is greatest at the center but deteriorates toward the periphery. Finally, surface structure representations are transient and begin decaying immediately upon activation. There are three classes of processes which act upon the deep (or long-term) and surface structures: generation, inspection, and transformation processes. Each process class consists of a number of processes which produce specific effects.

This sketch of the spatial information processing system thus far has been normative as are most of the data supporting the theoretical position. However, there have been two rigorous research efforts testing the applicability of the theory to individual differences and its relationship to traditional measures of spatial ability. Kosslyn, Brunn, Cave, and Wallach (1984) developed 10 tests based on imagery theory. The tests were designed to measure the quality of the surface representation and the efficiency of several specific processes operating on the surface representation.[3] Five points summarize this study. Performance measures derived from the spatial processing tasks showed large individual differences in processing efficiency and estimates of surface structure quality. The spatial processing measures were relatively uncorrelated (mean r = .15) with traditional measures of vocabulary, verbal fluency, memory span, and deductive reasoning. From a theoretical position, the spatial processing measures should be independent of one another; this was supported by a low mean intercorrelation (mean r = .28). However, based on the theoretical overlap of the processes employed in the 10 spatial tasks, the authors were able to substantially predict the rank ordering of the correlations between tasks (r = .82). Finally, Kosslyn, Brunn, Cave, and Wallach (1984) administered a traditional power measure of spatial ability (Form Board (Ekstrom, French, & Harman, 1976), similar to the sixth entry in Figure 1 and Table 1). Based on a post hoc analysis of the spatial processing requirements of the test, the authors were able to predict which of the theoretically derived tasks the test would be correlated with. Performance on this complex power test proved to be mostly related to the quality, capacity, and durability of the surface structure in the visual buffer.

[2] The majority of the following discussion is based on the extensive work in mental imagery of Kosslyn and his associates (e.g., Kosslyn, 1980, 1981; Kosslyn & Shwartz, 1977, 1981).

[3] The processes were sampled from the three classes of processes mentioned earlier: (1) generation, (2) inspection, and (3) translation processes. The details of the specific processes measured in the study are unimportant to the current discussion.

Poltrock and Brown (1984) investigated in more detail the relationship between individual differences in this theoretical account of spatial information processing and traditional measures of spatial ability. The authors developed six theory-based tests of spatial processing that produced nine measures; the measures reflected surface structure quality and the efficiency of several processes operating on the surface structure of the visual buffer. One result was similar to that reported by Kosslyn et al. (1984); all spatial processing measures showed large individual differences. Additionally, the spatial processing measures were well-defined and produced low average intercorrelations, supporting the theoretical contention that the spatial processing measures are relatively independent. The authors also administered eight traditional tests of spatial ability, which varied from being moderately speeded (e.g., mental rotation) to being nearly pure power measures (e.g., DAT; see the fifth test in Figure 1 and Table 1). Comparisons of the two test batteries (i.e., spatial processing measures and spatial ability tests) yield two important points. First, each of the spatial ability tests was significantly correlated with at least two of the spatial processing measures; surface structure quality and one or more of the processing efficiency measures. The other important point is that every single spatial test was significantly related to the quality of the surface structure representation in the visual buffer (mean r = .42) and, the magnitude of the correlation was a function of the tests' power orientation. This provides substantial support that the spatial information processing theory and traditional tests of spatial ability are related. The final analysis conducted by Poltrock and Brown (1984) tested the form of the relationship between the spatial processing theory and spatial ability. Specifically, the authors tested a model that predicted that surface structure quality and process execution efficiency was the source of a general spatial ability, which was expressed as scores on the various paper-and-pencil spatial tests. The model was resoundingly supported, producing a non-significant χ^2 goodness-of-fit statistic (χ^2 = 88.1, df = 83). Not surprisingly, the greatest contribution to general spatial ability was the quality of the surface structure representation and the greatest expression of general spatial ability was in performance on the complex power spatial tests.

This section can be summarized as follows. The combination of spatial information processing and imagery theories provides an internally consistent and parsimonious perspective from which to view spatial cognition. The general view is that the same structures and processes that allow us to sense, perceive, represent, and recall the external spatial world support the ability to generate, inspect, and operate on self-initiated internal representations. Much normative research has elaborated the nature and properties of the surface structure represented in the active visual buffer and the processes that operate on this surface structure. Efforts to extend this normative research to individual differences paradigms has further demonstrated the viability of this perspective as an individual differences construct. Large and reliable individual differences were found in the quality, precision, and durability of the active surface structure representation as well as in the efficiency of executing processes that generated, inspected, or transformed aspects of the surface structure representation. Importantly, this line of research has been extended to the traditional psychometric domain of tests of spatial aptitude. The research has shown that tests of spatial ability can be successfully interpreted in terms of the structures and processes posited by spatial information processing and imagery theories. Furthermore, individual differences in spatial and imaginal processing have been linked to individual differences in spatial test performance. In particular, the quality of the surface structure representation was the most powerful and general predictor of spatial test performance, especially for the power oriented spatial tests. This latter fact is significant since, in the last section, it was demonstrated that the power tests have the greatest communality in the domain of spatial tests and that the power tests are the most valid predictors of school and job performance.

These conclusions can further guide our efforts toward the development of a spatial test. The review of the spatial test literature suggested that efforts be focused on power tests, those that use complex items presented with generous time limits. The theoretical and research review above, provides an explanation of why the power tests have the greatest communality. It appears that this class of tests in particular, tax an individual's ability to create, retain, and operate on the active surface structure in the visual buffer. Since all spatial problem solving tests require representation in the visual buffer the greater the requirements on the buffer the greater will be the shared variance between two tests. In addition, the more complex the test the more processes will be required for item solution. As more processes are invested, there will be more process overlap between tests, which will further boost the proportion of shared variance between two tests.[4] Thus, the search for the type and nature of a spatial test that has the greatest opportunity for enhancing the predictive validity of the ASVAB has been substantially narrowed to a complex power test that will particularly tax an individual's ability to create, retain, and operate on the surface structure representation in the visual buffer. What is still lacking before development of such a test is feasible, is a methodology for test design that will permit a better understanding of test and test item difficulty. This methodology comes from componential analyses of spatial tests, the topic of the next section.

[4] The question remains as to why complex power spatial tests are also more correlated with other tests in general. There are two explanations for this phenomenon. The first is that many complex reasoning tests can be represented and solved using either propositional or spatial-analog representations and processes. This is true for many deductive (e.g., syllogisms) and inductive (e.g., analogies) reasoning tests and this is especially true for such "knowledge-free" tests as Raven's APM. Another explanation is that all conscious, attention demanding cognitive activity may draw on a single undifferentiated resource pool, a type of mental or cognitive energy. Accepting an energy view then, technically amounts to a view of cognition where there is a limited undifferentiated resource pool available for any and all cognitive activities. However, these views are not mutually exclusive and can be combined into a single framework. In this combined view, individuals would invest this general resource pool when performing tasks that principally require either analog or propositional representations and processes. This resource pool would thus set an upper bound on performance. However, if only the amount of general resources mattered in cognitive activity, then individuals high in general ability would be uniformly high on verbal and spatial tests. This is not the case, although there is a tendency for tests to be positively intercorrelated (which is the chief evidence for a general undifferentiated resource pool). If the existence of a general resource pool is accepted, how could this account for the non-uniformity of performance across tests (ignoring specific knowledge)? For example, individuals with poor quality surface structure representations yet a high resource level could have performance limited by the surface structure representation on tasks which tax this type of problem solving. Further, individuals with average general resources but with a very good quality surface structure representation could out perform individuals with more resources. Generally then, if it is assumed that there are at least two representational systems (analog and propositional) with dedicated structures and processes, each of which are large sources of individual differences, and a general undifferentiated resource pool that is vested in performance under both representational systems, then there is ample opportunity for individuals to compensate for deficits in any of the three sources of individual differences (general resources, analog, and propositional representational systems). This model of human intelligence logically explains both the phenomenon of positive manifold for mental tests and the non-uniformity of the level of performance across test classes. It is this view that guides the current work.

Componential Analysis of Spatial Tests

Componential analysis is a general set of techniques that have been successfully applied to a large number and variety of standard aptitude tests to reveal systematic sources of individual differences in test and test item difficulty (Carroll, 1976; Estes, 1974; Pellegrino & Glaser, 1979, 1980, 1982; Sternberg, 1977). Briefly, there are several steps in the componential analysis of an aptitude test. Initially, a standard psychometric test is selected for study and subjected to a detailed task analysis. In the first step of the task analysis test items are aggregated, sorted, and otherwise explored to determine the similarities and differences among the items with a keen eye toward determining the sources of item difficulty. Having successfully found the principal sources of item difficulty, these dimensions are subjected to a rational analysis to elucidate the probable structures and processes required for successful item solution. The information from the task analysis forms the bases for an information processing model of item solution. Following model specification, an experimental item set is developed that allows estimation of the model's parameters. Next, the experimental item set is administered to subjects and model parameters are estimated to determine the internal validity of the model. Internal validity refers to the model's ability to capture performance and is evaluated in terms of model fit and reliability. The final step is an evaluation of the model's external validity, which refers to how well individual differences in performance on the experimental task is related to individual differences in performance on the source aptitude test.

The power of this methodology over the traditional psychometric approach to item and test design is manifold. For example, traditional methods employ statistical (content and process free) criteria for item selection whereas a componential approach uses rational, rule governed criteria for selecting items. Additionally, the performance model that drives item creation and selection can be used to assess the adequacy of item performance and thus greatly reduces the uncertainty involved in creating new items and parallel forms. The process model for item solution provides the basis for predicting and explaining interitem and intertest correlations, which is not possible using the usual psychometric methodology. Therefore, theorv based componential tests have greater construct validity (Embretson, 1983). Finally, because the components for successful item solution are explicitly specified and assessed, componentially derived tests have substantially more diagnostic potential than traditional tests.

The componential approach has been successfully applied to a number of spatial tests which has substantially increased our understanding of the domain of spatial ability (see Lohman, et al., 1987, and Pellegrino, Alderton, and Shute, 1984, for reviews). Two examples will be used to illustrate the approach. The first example comes from Mumaw and Pellegrino (1984), which reports a componential analysis of the Minnesota Paper Form Board (MPFB) (Likert & Ouasha, 1970) (see entry 6 in Figure 1 and Table 1). The test requires individuals to determine which of the five completed puzzles could be made from the puzzle pieces in the upper left. The authors performed a task analysis of the MPFB and found that there were four sources of item difficulty: the number of pieces in the puzzles, locating corresponding puzzle elements, rotating the puzzle elements to determine if they match, and making the subtle discriminations between a correct and incorrect puzzle element. The authors used these dimensions of item difficulty to specify a performance model for item solution. The model specifies that solution duration would be a function of the number of puzzle elements encoded and compared, the number of elements searched for (i.e., that were not in corresponding locations relative to its location in the standard), the number of elements rotated, and the number of mismatching elements between the standard and test figures. Similar sources of difficulty were thought to affect the accuracy of item solution.[5] The authors next developed an item set that systematically and independently manipulated each of these sources of item difficulty. The experimental item set was then administered; latency and accuracy data were collected for each item. The process model was then fit to averaged group latency data, which produced an R^2 value of .90. Next, the model was fit to individual subject data, the average R^2 was .92. Thus, the model was capable of accounting for nearly all of the between item variability for both the group and individual subject data. Having successfully demonstrated the internal validity of the process model, the authors turned to the more general question of external validity. The latency derived process estimates and accuracy scores for each subject were used as predictors of the paper-and-pencil MPFB test score; this produced a multiple R of .78 (which approaches the reliability of the test itself, .81). Clearly then, the process model was tapping the same sources of individual differences as the paper-and-pencil test. Finally, more refined analyses of the pattern of errors for high and low ability subjects (based on incoming MPFB scores) suggested that the greatest source of item difficulty was making the fine discriminations between matching and non-matching puzzle elements. Interpreted in terms of the theory presented in the previous section, subjects low in spatial ability have a poorer quality surface representation in the visual buffer.

The Mumaw and Pellegrino (1984) study makes two important points. Standard psychometric tests can be interpreted in terms of the structures and processes posited by an information processing approach to aptitude. These structures and processes were capable of accounting for virtually all of the reliable variance in the actual paper-and-pencil test. The componential approach provides a much richer understanding of problem solving performance and sources of item difficulty. Based on the performance model and the model of item difficulty, it is possible to generate a nearly infinite set of items. More importantly, specifying the performance model allowed a detailed exploration of the construct validity of the task, both internally and externally, and provides the basis for predicting and interpreting the pattern of intercorrelations between spatial tests.

A more general variant of the componential analysis technique was used by Alderton (1986; see also, Alderton & Pellegrino, 1985). Instead of focusing exclusively on a particular test, the goal was to gain an understanding of a particular class of spatial tests referred to as surface development tests. Surface development tests, such as the Differential Aptitude Test battery's Space subtest (DAT) (Bennett, Seashore, & Wesman, 1974) (see the fifth entry in Figure 1 and Table 1), are generally intercorrelated in the low 70s with one another. Superficially, all surface development tests require the examinee to mentally fold a two-dimensional representation into a three-dimensional object and making judgements about the final object (usually against presented alternatives). The author began by taking items from several surface development tests and looked for similarities and differences across tests. Following this, an attempt was made to define the sources of item difficulty within each test and then across the tests. The results suggested three common sources of item difficulty: the number of folds that had to be made, the number of distinctly marked surfaces that had to be monitored during the folding process, and the type of surface marking (orientation free or orientation specific). These sources of item difficulty were then instantiated in an information processing performance model that predicted solution duration and accuracy would be a function of

[5] Note that this is simply a sketch of the more detailed information processing model specified by the authors. For a full account see the original source (Figure 2, p. 922).

the number folds, the number of marked surfaces, and the type of surface marking. To test this model, an experimental item set was developed that systematically and independently manipulated each variable. The items were administered to a large sample; latency and accuracy data for each item were collected. The model accounted for 96.4 percent of the group latency data; and the model significantly fit 90 percent of all individual subject data. Clearly, the model performed well. The latency derived parameters from individual subjects, along with accuracy scores, were then used to predict performance on the DAT; the multiple R was .60. Thus, individual differences in the parameters of the general surface development model were highly related to individual differences in performance on a specific paper-and-pencil test. Tc assess the generality of the model, a battery of spatial tests was factor analyzed revealing a general speed and general power factor; factor scores were computed for each subject and used as dependent measures with the model parameters as predictors. The regressions showed that the model produced a multiple R of .52 with the power factor and .50 with the speed factor. Further, the measures associated with the power factor score were those that reflected the quality and durability of the surface structure representation in the visual buffer. The important point is that the rationally developed model was capable of accounting for individual differences in problem solving on the experimental items, on a specific test of surface development problem solving, and, most importantly, on general measures of both the speed and power dimensions of spatial ability.

Several points summarize this section. Componential analysis provides a powerful methodology for developing internally consistent and externally valid measures of spatial ability. This is true for specific tests and classes of tests. This approach has many advantages over traditional psychometric approaches to test development. For one, the tests are grounded in cognitive theory and instantiated as a model of performance that can be explicitly tested and evaluated. Also, embedded in the test performance model are specific rules that govern item development and allow for easy generation of systematic test items.[6]

Summary

Combining the results from the previous three sections provides the direction needed to develop a systematic test of spatial ability with a strong potential for augmenting the predictive validity of the ASVAB. The review of the spatial test literature focused attention on power tests, those employing complex items administered under generous time limits. This is because power tests have the greatest communality with other spatial tests and, most importantly, because virtually all of the validity evidence for spatial tests is derived from these complex power tests. The previous section reviewed and organized the available theoretical literature on spatial information processing and imagery theory. It was concluded that there is substantial support that the same processes and structures

[6] No mention has been made of the diagnostic and remedial advantage of tests developed under componential approaches because it is not germane to the current discussion. However, briefly, componential tests have a diagnostic advantage over usual tests because performance is assessed at specific levels (e.g., rate of mental folding, sensitivity to various surface marking types) such that a pattern of strengths and weaknesses emerge from an individual's performance. This pattern or profile can serve as the bases not only for diagnosing a person's difficulties but this can also serve as the starting point for remediation.

used to sense, perceive, represent, and remember the external spatial world support the ability to generate, inspect, and operate on self-initiated internal representations. Further, the available evidence demonstrates that there are large differences among individuals in the quality of the surface structure representation and in the efficiency of the processes that generate, maintain, and operate on it. In particular, it is evident that the largest source of individual differences in imagery ability is in the quality of the surface structure representation. There was clear evidence that the two constructs substantially overlapped in the studies that related individual differences in mental imagery to individual differences in tested spatial ability. The proposed model (Poltrock & Brown, 1984) is one in which the processes and structures of mental imagery are the source of a general spatial ability (abstraction), which is expressed as scores on typical tests of spatial aptitude. Importantly, the model showed that the greatest contribution to general spatial ability was the quality of the surface structure representation and that the greatest expression of general spatial ability was in the complex power spatial tests. This provides a crucial link between the theory of spatial information processing and tested spatial aptitude. The clear implication is that complex power tests, the source of the greatest spatial test validity, particularly tax individuals' ability to create and operate on a high quality surface structure representation. Thus, the most promising direction for new test development is to develop a test that will place heavy demands on creating and maintaining a high quality mental image. Finally, the section on componential analysis revealed that there is a systematic methodology available for the development and evaluation of model-based, complex tests of spatial ability. Interestingly, the two componential studies reported also suggested that the quality of the surface structure representation was a particularly salient source of individual differences. These results guided the development of Integrating Details, reported in the following section.

INTEGRATING DETAILS

The previous section summarized the context, background research, and theory that provided direction in the development of Integrating Details. In the following sections, the developmental, history, and properties of Integrating Details will be fully discussed.

Description

Test Items

The idea for the actual Integrating Details test[7] items has roots in two previously investigated tests; a task developed by Poltrock and Brown (1984) called integrate (see pp. 111-113) and the MPFB (mostly its experimental analog (Mumaw & Pellegrino, 1984)).

[7] For completeness, historical accuracy, and posterity, the development of the original Integrating Details test was the result of a collaborative effort between James Pellegrino, Thomas McDonald, Ronald Abate, and David Alderton at the University of California, Santa Barbara between January and March 1985 under contract from the Navy Personnel Research and Development Center (NAVPERSRANDCEN) (see NPRDC TR 87-31). The test was also reviewed by Earl Hunt (University of Washington, comiprincipal investigator with James Pellegrino) and inspected by John Wolfe (NAVPERSRANDCEN contract monitor).

Sample items from both tests are reproduced in Figure 2. The Poltrock and Brown (1984) integrate items were sequentially presented on five slides. The first four slides contained a single irregular figure. For the first slide, the subject had to memorize the figure and remember where the next three figures were to be integrated (indicated by colored edges). The next three slides each contained an irregular figure with a single colored edge indicating where the figure was to be mentally attached to the first figure. Thus, the task required the sequential integration of irregular figures, which placed very high demands on both the quality and durability of the surface structure representation. The final (fifth) slide contained four alternatives from which an answer had to be selected. There were several problems with this test, which made it undesirable for straight-forward adoption. The test was exceedingly difficult producing a very low average accuracy. Also, the amount of time spent on each succeeding slide decreased, which indicates that individuals may have been rushing through the slides to minimize the effects of a rapidly deteriorating internal representation. (This was not the explanation given by the original authors.) The use of very irregular figures made it difficult to control extraneous sources of item variance. Finally, creation of the alternative set was unsystematic, making it difficult to distinguish the source of accurate and inaccurate performance (e.g., a correct decision could result because the correct alternative may have been easily distinguishable, or a wrong decision could result because the alternatives may have been too similar). The chief merit and interest in the test was the heavy demands placed on creating and maintaining a stable and detailed internal representation.

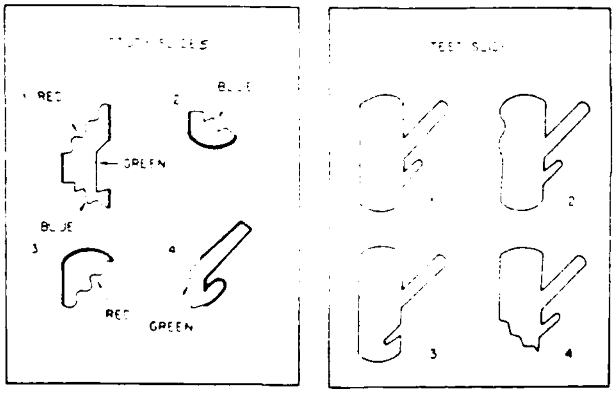


FIG 5 Sample integrate problem (Words and arrows identify regions that were colored in the original problem.)

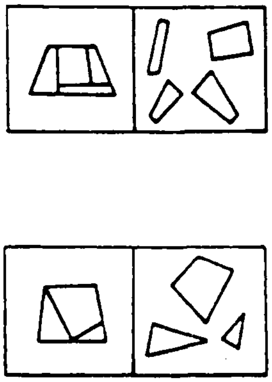
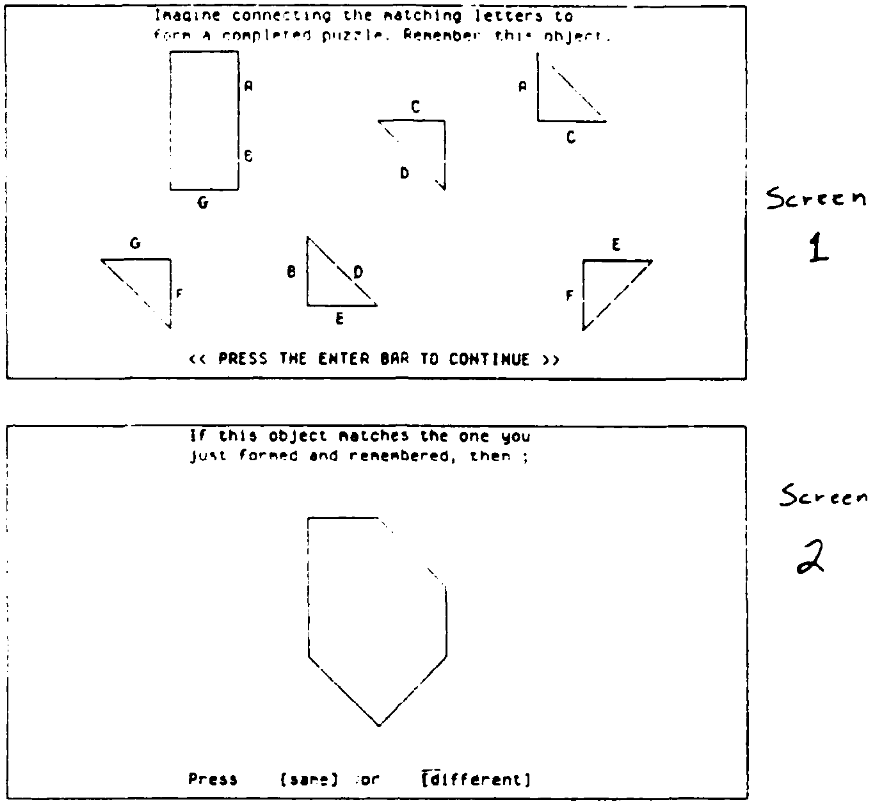


Figure 2. Sample integrate items and sample MPFB analog items.

The second panel of Figure 2 contains two items from the experimental analog of the MPFB, which also supplied ideas for the development of Integrating Details. This test presented individuals with puzzles, composed of small semi-regular geometric forms that ostensibly were put together and then compared. In fact, the performance model demonstrates that individuals do not actually put the puzzle pieces together. Rather, since the complete and disheveled puzzles were simultaneously presented, individuals encoded and compared each of the puzzle elements in turn, without ever forming a complete puzzle. Both latency and errors were found to be linearly increasing functions of the number of puzzle elements. The chief drawback of this test was that all of the visual information necessary for making decisions was simultaneously present, which greatly reduced the representational demands of the test. The merit of the test was that it used a relatively bounded set of semi-regular geometric forms making item creation easy, and the systematic manipulation of the number of puzzle elements (from 3 to 6 pieces) provided a sound index of item difficulty. Furthermore, these manipulations provided the basis for a performance model that was employed to ascertain the degree to which subjects were performing the task as expected.

The combined merits of these two tasks provided the essential design characteristics for Integrating Details. A sample Integrating Details item is presented in Figure 3. It was decided that puzzle problems would be created from a limited set of regular geometric forms. A total of 11 regular geometric shapes were decided upon. Several of these shapes are presented in the upper panel of Figure 3; note that a shape can be used more than once. Using only this 11 shape vocabulary sets bounds on the creation of items while controlling obtuse sources of error variance. When the item is presented, the subject is required to connect the puzzle elements mentally by connecting the matching letters (e.g., A to A, B to B, etc.). (Letters were used instead of colored edges to allow the use of monochrome computer monitors.) The number of puzzle pieces were systematically varied (originally from 3 to 6) providing the beginning of a performance model (to be described in the next section). Next, it was decided that the puzzle elements should be presented separately from the decision options. Thus, the top panel in Figure 3 is actually the first screen[8] of an item. This was to ensure that there would be substantial demands placed on the surface structure representation. To further tax this representation, and ensure that individuals mentally integrated the puzzle elements, it was decided that the alternatives would not contain the internal connecting edges of the elements, only the outline of the completed puzzle would be presented. An alternative is presented in the lower panel of Figure 3 and is the second screen of an Integrating Details item. This is in contrast to the MPFB items shown in Figure 2 but similar to the integrate alternatives used by Poltrock and Brown (1984). Further, because of the problem in scaling the difficulty of response alternatives, only a single alternative is presented. In Figure 3, the presented alternative could be made by correctly connecting the matching letters in the first screen; therefore, this item requires a same (or match) response. If the figure presented on the second screen could not be made by correctly connecting the matching letters, then a different (or mismatch) response would be required. Certain rules were used to guide the creation of the different trial alternatives. Specifically, the number of outline edges for the incorrect alternative would be the same as the number of edges in the correct alternative. Simple rotations of the correct alternative were also disallowed. Finally, attempts were made to roughly keep the area of the incorrect alternative similar to the correct alternative.

[8] "Screen" is used since Integrating Details has only been computer-administered.


Imagine connecting the matching letters to form a completed puzzle. Remember this object.

<< PRESS THE ENTER BAR TO CONTINUE >>

If this object matches the one you just formed and remembered, then ;

Press [same] or [different]

Figure 3. Sample integrating details item.

Each Integrating Details item is presented in the following sequence. First, the puzzle elements are presented alone. The subject is allowed as much time as needed to correctly connect the matching letters to form a complete object and store the object in memory. When she/he has formed and remembered the object, a key is pressed and the puzzle pieces are replaced by an alternative. The examinee then has as long as necessary to decide if the object in memory matches the one presented, and then selects either a same or a different response. Each item produces three dependent measures: time spent putting together the puzzle elements (Integrate Time), time spent deciding if the presented alternative is correct (Decision Time), and response accuracy (Accuracy).

Performance Model

By design the items lend themselves to the development of a reasonably well specified performance model. Referring to Figure 3, upon presentation of the puzzle elements, the individual must encode one of the elements in the visual buffer (e.g., the rectangle) and decide which of the lettered edges to match (assume it is the "A"). Next, he/she must search among the remaining elements to locate the corresponding letter (the "A" on the isosceles right-triangle in the upper right of screen 1). Having found the letter, the shape must be encoded then synthesized at the correct location on the first figure, and this new product must be stored in the visual buffer. The new product in memory must then be scanned to determine the next letter to search for (assume "C"). This next letter ("C") must be found among the puzzle pieces, the shape encoded and synthesized at the correct location and the new product must be stored.

This search-encode-synthesize-store cycle must be iteratively executed until all of the objects have been synthesized. Table 4 contains a schematic representation of this (integration) phase of the performance model. The first column indicates the cycle number, the next column contains the cumulative step number. The column labeled Processing Activity describes what is being done in each step. The final column indicates whether the stimulus being operated on (or searched for) is physically present (external) or not (internal). There are several important details to note about the model. Most importantly, following the first cycle (Cycle 1 in Table 4 or connecting "A" to "A" in the example above), the subject no longer has all of the necessary information physically available. That is, for the first cycle, the two pieces that are synthesized are both present on the screen, but beginning with Cycle 2 (step 8), the subject is working on the synthesized and stored product from the prior cycle. This is easily seen in the last column of Table 4. Thus, beginning with the second cycle, the task places increasing demands on the quality of the internal surface structure representation. Recall from the earlier discussion, that the surface structure representation begins decaying immediately. Therefore, beginning with Cycle 2, the subject must execute additional processes that refresh the constantly decaying surface structure representation, making the task more difficult and effortful.

Table 4

Integration Performance Model

Step	Processing Activity	Stimulus
Cycle 1		
1	Encode Shape(i)	External
2	Scan Lettered Edges of Shape(i)	External
3	Select Lettered Edge(j) of Shape(i)	External
4	Search for Letter(j)	External
5	Encode Shape(i+1)	External
6	Synthesize Shape(i+1) at Shape Edge(j)	External
7	Store Product(k)	Internal
Cycle 2		
8	Scan Lettered Edges of Product(k)	Internal
9	Select Lettered Edge(j+1) of Product(k)	Internal
10	Search for Lettered Edge(j+1)	External
11	Encode Shape(i+1)	External
12	Synthesize Shape(i+1) at Product Edge(j+1)	Internal
13	Store Product(k+1)	Internal
14	Scan Product(k+1) for additional Lettered Edges	Internal
15	if Yes recycle to Scan(8)	
16	if No EXIT	

Following the integration of all the puzzle elements on the first screen, the individual indicates that she/he is ready for the alternative. In terms of Table 4, this is represented by exiting the integration phase of the model at step 16. At this point, the individual has created the completed puzzle and is actively maintaining (refreshing) the product. The alternative on the second screen is encoded and compared with the mental composite, a decision is made and a response emitted (*same* if it matches or *different* if it does not match). Unfortunately, this description does not capture the complexity of the matching process, which is only poorly understood. Logically, there are two distinct stages in this portion of item solution that affects both latency and accuracy; encoding and decision. In general, encoding difficulty (time and accuracy) is influenced by the complexity of the presented figure. "Complexity" is a function of several aspects of the figure, including the size (or area), number of edges, number of oblique lines, and degree of Gestalt "goodness-of-form." The decision or matching phase is influenced by the degree of overlap between the presented and true alternatives, including the area and the number of overlapping edges. Both of these phases are affected by the quality of the representation created during item integration and the durability of the created representation. Additional research is required before a comprehensive performance model for this processing stage can be developed (see the section on *Future Research*).

Collectively, this characterization makes several predictions about group and individual performance on the items. Most obvious, is that Integration Time should increase as a function of the number of puzzle elements synthesized. This is the result of having to execute additional encode-search-synthesize-store cycles. Also, accuracy should decrease as more of these cycles are executed because it is assumed that the probability of correctly executing any process (steps 1-16 in Table 4) is less than one. Therefore, an error can occur at any point. Importantly, since the output of each process is the input to the next process, errors will be perpetuated throughout the remaining problem solving steps. This sequential dependency predicts that as more processes are executed, overall, the probability of successful problem solution decreases. A second source inaccuracy related to the number of elements is the constantly decaying surface structure representation. Minimizing surface structure decay necessitates the execution of regeneration or refreshing processes which requires effort borrowed from the execution of the other problem solving processes. If the representation decays beyond the point where it can be fully regenerated, the regenerated representation may contain errors[9] that were not previously part of the representation. For the decision portion of the item, solution duration should be a minor function of the number of puzzle pieces in the first half of the problem since, in general, there will be more features to compare. For this reason, the number of pieces should also affect the accuracy of decision processes. However, the irregularity or complexity of the final shape should have an even greater impact on decision time and especially accuracy. Irregularity or complexity is a function of the number of vertices, the number of oblique angles, the number of indentations, and the presence or absence of symmetry in the final object. Therefore, the more irregular or complex the final shape, the more difficult the object will be to maintain in the visual buffer and there will be more features to compare. Thus, complexity should affect both decision time and accuracy. Since during the decision stage the object in memory is compared to the presented alternative, the similarity between the object in memory and the presented alternative should have an impact on decision accuracy and to a lessor extent on decision latency (but no effect on integrate time is expected). These predictions will be evaluated in a later section.

[9] Based on the large form perception literature, the most likely class of regeneration errors should be those which "regularize" the original image. That is, there is a tendency to smooth and make the product more systematic and regular than it should be; such as making a slightly flat square into a true square.

Methods

Samples Tested

Between June 1986 and December 1987 well over 2,000 individuals have been administered Integrating Details. This report will focus on performance from 1,833 of these individuals.[10] Six separate samples have been tested ranging in size from 84 to 542 subjects. The six samples are summarized in Table 5, roughly in the temporal order in which they were tested (albeit, there was overlap). Because the test has been under continual development, each of the samples have received somewhat different versions of the test. The type of changes from one sample to the next included changes in the number of items, instructions, range of item difficulty, minimum response time limits, etc. These changes will be documented below.

Table 5

Sample Constituency and Sizes

Label	Subjects	N	Label	Subjects	N
A.	College Students	170	B.	Navy Machinist Mates	84
C.	Navy Recruits	460	C'.	Retest Subset of C	127
D.	Navy Recruits	427	E.	High School Students	542
E'.	Retest Subset of E	445	F.	Navy Recruits	150

Integrating Details was originally designed as a 48 item test and administered to 170 college students (Sample A; see Hunt, Pellegrino, Abate, Alderton, Farr, Frick, & McDonald, 1987, for details). Half of the items required a same response and half a different response. The items required the synthesis of either 3, 4, 5, or 6 pieces. No minimum or maximum times were set. The 48 items were grouped into six blocks of 8 items each. Each block contained two each of the 3, 4, 5, and 6 piece items; one requiring a same response and one a different response. The six blocks were presented in one of six different orders to each subject; presentation order was determined randomly across subjects. The items were presented on Apple II class micro-computers. In January 1986, the 32 item test was shortened for presentation to Navy Machinist Mates (Sample B) on-board ship. The test was shorted to 32 items by simply eliminating two of the original six blocks of 8 items. No formal or empirical basis was used to decide which of the six blocks to discard (the last two were arbitrarily chosen). In the spring of 1986, the test was converted to run on Apple III micro-computers for presentation to Navy Recruits at the San Diego Recruit Training Center (RTC). The conversion was required to take advantage of available testing equipment at RTC. In the process of converting the program for Apple III use, several minor changes were made to the instruction set. No other changes were incorporated.

[10] The data from the remaining subjects are in various stages of analysis and not available at this time (1 June 1988).

Following detailed data analyses of the performance of Samples B and C, several problems became evident that suggested important changes in the test. First, individuals who found the test particularly difficult tended to simply press response buttons to finish the test. However, some of these people may have been trying to solve the items by using the perceptual after-image of the puzzle pieces on the first screen to decompose the final object. To discourage the "button-pushers" and alter the after-image strategy, minimum response times were set. Specifically, if an individual responds in less than 1.25 seconds when the puzzle pieces are presented, he/she is warned that the only successful way to solve the problems is to spend more time putting the pieces together. Following the warning, the pieces are again presented and the total time accumulated. For the Decision portion of the item, if an individual responds in less than .5 seconds, then she/he is warned to spend more time deciding on their response; the alternative is then presented again, and total time accumulated. Another change was indicated by item analysis. The test contained several faulty items; faulty in that they did not discriminate and/or mean performance was well below chance. Four (of 32) items were corrected. An additional change also seemed warranted. The majority of the items proved quite difficult, which appeared to frustrate examinees and reduce motivation. Therefore, 2-piece puzzle problems were created and incorporated to engender some sense of accomplishment and maintain motivation.[11] The addition of 2-piece problems altered the final block design such that each block contained 10 items, two each of the 2, 3, 4, 5, and 6 piece puzzle problems, one of each piece value in a block required a <u>same</u> response, the other required a <u>different</u> response. The instructions were also altered to reflect these changes. (A 2-piece problem was added to the practice set and the practice problems were reordered so that each subject is presented with a 2-, 3-, 4-, 5-, and 6-piece puzzle problem, in order.) Because of the many changes, and the identification of several faulty items, it was decided that an enlarged item set should be field tested on a sample of Navy recruits. The enlarged set included the 48 original items (less the 4 items that were found to be bad), the 4 corrected items and 12 new 2-piece problems. This resulted in a total of 60 items, organized into six blocks of 10 items.

At nearly the same time as the changes were being made, it was decided to switch over from Apple III computers to Hewlett-Packard Integral Personal computers. This was to ensure future compatibility with the Accelerated Computer Adaptive Testing program, which would be the ultimate delivery vehicle for the test, if it was adopted for administration to applicants. Therefore, the entire test was reprogrammed on Hewlett-Packards under UNIX® and in the language "C." The 60 item version was then administered to Navy recruits on Hewlett-Packard micro-computers at the San Diego RTC; this is Sample D in Table 5. Coincidental with these changes, a contract was awarded to collect retest reliability data on a monetarily motivated applicant-like sample. It was decided that the 60 item version was too long and that not enough was known about the 2-piece puzzle items to risk administering them under the retest contract. Therefore, the corrected set of 32 items running on the Hewlett-Packard was used in the retest study on high school students; this is Sample E in Table 5. Finally,

[11] Coincidentally, by chance over 3/4s of all subjects were receiving the most difficult item type, a 6-piece puzzle, as the first item in the test which undoubtably reduced subsequent effort. Therefore, with the addition of the 2-piece puzzle items it was decided that subsequent block construction would be constrained to have the first item of each block be a 2-piece problem. Therefore, every person taking the test would begin with an easy item.

because it seemed possible that changing from the Apple to the Hewlett-Packard might have produced some unknown changes in the test, a sample of Navy recruits was administered the 32 item Hewlett-Packard version and the original (without the 4 corrected items) Apple II version of the test; this is Sample F in Table 5.

The relationship between the changes in the test and the various samples tested is summarized in Table 6. The first column indicates the sample and can be related back to Table 5. The next column gives the type of computer used (note that Sample F received two different computers). "No. Items" reports the number of items in the version for the sample. "Piece Value" gives the range of item difficulty used in the study (a constant 3-6 for all but Sample D). The "Min. Time" column refers to whether or not the 1.25 and .5 second minimum response times were incorporated into the version of the program. "Inst. Change" means "Instruction Change" and refers to whether or not the instructions had been changed relative to the original (Sample A) version of the test. The last column answers whether or not the four faulty items had been replaced or not.

Table 6

Relationship Between Changes in Integrating Details and Samples Tested

Sample	Computer	No. Items	Piece Value	Min. Time	Inst. Change	Item Change
A	Apple II	48	3-6	No	N/A	N/A
B	Apple II	32	3-6	No	Minor	No
C	Apple III	32	3-6	No	Minor	No
D	Hewlett-Packard	60	2-6	Yes	Yes	Yes
E	Hewlett-Packard	32	3-6	Yes	Yes	Yes
F	Hewlett-Packard	32	3-6	Yes	Yes	Yes
	and Apple II	32	3-6	No	No	No

Results

Initial Descriptive Statistics

Table 7 contains mean performance for the first five samples[12] for each dependent measure: Integrate Time, Decision Time, and Accuracy. Table 8 contains the corresponding standard deviations.

[12] Sample F is not included because of the machine dependent presentation order effects. That is, the design of the study required that half of the subjects take the Hewlett-Packard version first followed by the Apple II version and the reverse was true for the other half of the sample. Therefore, means (as well as other descriptive statistics) are confounded by machine and presentation order. These data will be discussed separately in a later section.

Table 7

Mean Performance on Integrating Details

Measure	Sample A	Sample B	Sample C	Sample D	Sample E
Integrate Time	19.80 s	18.04 s	16.08 s	18.81 s	17.33 s
Decision Time	3.83 s	3.45 s	3.37 s	3.87 s	3.51 s
Accuracy	.71	.63	.65	.71	.70

Table 8

Standard Deviations of Performance on Integrating Details

Measure	Sample A	Sample B	Sample C	Sample D	Sample E
Integrate Time	8.40 s	8.46 s	8.60 s	8.42 s	8.66 s
Decision Time	1.81 s	1.61 s	1.64 s	1.69 s	1.77 s
Accuracy	.13	.14	.12	.12	.14

Looking across the five samples reveals a rather remarkable consistency, both for the means and standard deviations. This is especially true given the number of changes in the test and test items as well as the differences in the samples, sample sizes, testing situations, computers used, and versions of the test. The only real exception is the low mean Integrate Time of Sample C. There were a number of difficulties with this sample that will be noted throughout this report.

Scoring Changes and Descriptive Statistics

The measures in the previous section were derived using the scoring rules from the original version of Integrating Details. These rules specified that latency measures would be based on correct trials only and that raw latencies (i.e., untransformed time in seconds) should be used. However, these rules were not empirically based. Various alternative scoring systems were explored using the data from Samples C, D, and E. For each of the three samples, and for both latency measures (Integrate Time and Decision Time), comparisons were made for several possible transformations of the data, including, mean time, log mean time, mean log time, square root mean time, and mean square root time. By far, the best transformation for Integrate and Decision Time, in each sample was mean log time (i.e., perform a (common) log transformation on each item latency and then compute the arithmetic mean). Since the same results occurred for all three samples the data were combined (N = 1,388). Table 9 compares the means, standard deviations, skewness, and kurtosis values for mean time and mean log time (for correct trials only) for Integrate and Decision times. There is a notable improvement in the symmetry and peakedness of both distributions following the log transformation of each item latency.

Table 9

Distributional Comparisons of Mean and Mean Log Times (N = 1,388)

Parameter	Integrate	Log Integrate	Decision	Log Decision
Mean	17.019	1.065	3.457	.433
SD	8.513	.207	1.636	.136
Skewness	1.775	-.429	3.561	1.085
Kurtosis	7.517	.739	23.242	2.874

The next set of analyses focused on the necessity of excluding incorrect latency trials. This idea was initially considered because logically, Integrate Time is neither correct nor incorrect. That is, following the integration of the puzzle elements a subject simply responds that she/he is ready for the alternative and there is no objective measure as to if the puzzle pieces were correctly or incorrectly synthesized. Using the data from Samples C, D, and E the relationship between mean log correct times and mean log all times (all trials) was explored. The correlation between errorfree mean log Integrate Time and errorful (all trials) mean log Integrate Time was a striking .984 (N = 1,388). The scatter plot for this correlation is presented in Figure 4. Based on these results, the relationship between Decision Times for correct and all trials was also explored. The correlation between errorfree mean log Decision Time and errorful (all trials) mean log Decision Time was found to be .975. The scatter plot for this correlation is presented in Figure 5. Comparisons of the distributional properties of correct and all mean log times is presented in Table 10. Note that mean all log times are somewhat higher and have larger standard deviations relative to the measures based on correct trials only; this is true for Integrate and Decision Times. However, note that the symmetry and peakedness gains obtained from using log times have not been lost by including error trials in the latency scores.

Although using the mean log time of all trials provided measures with the best distributional properties and retained all of the information in the original (correct trials only) measures, they may not be the most reliable measures. Therefore, one final set of analyses was performed. For each sample, the item level data was split into odd and even halves, then mean correct times, mean log correct times, mean all times, and mean log all times were calculated for each half and separately for each sample. Following this, the measures derived from odd halves was correlated with the corresponding measure from the even half; the Spearman-Brown half-length correction was then applied. These reliabilities are reported in Table 11. (Because of the very different number of items for sample D, the results are reported separately for each sample.) This analysis makes it quite clear that the reliability of the mean log times is equal or superior to mean raw times, for both Integrate and Decision times. Furthermore, the advantage of mean log times over raw times is enhanced when all trials are used instead of using only correct trials.

Collectively, the preceding analyses suggest that all trials should be used in calculating latency measures and that these measures should be based on log transformed values. Besides ameliorating distributional abnormalities and increasing group reliability, using mean log all times has several additional benefits. For one, the improved precision

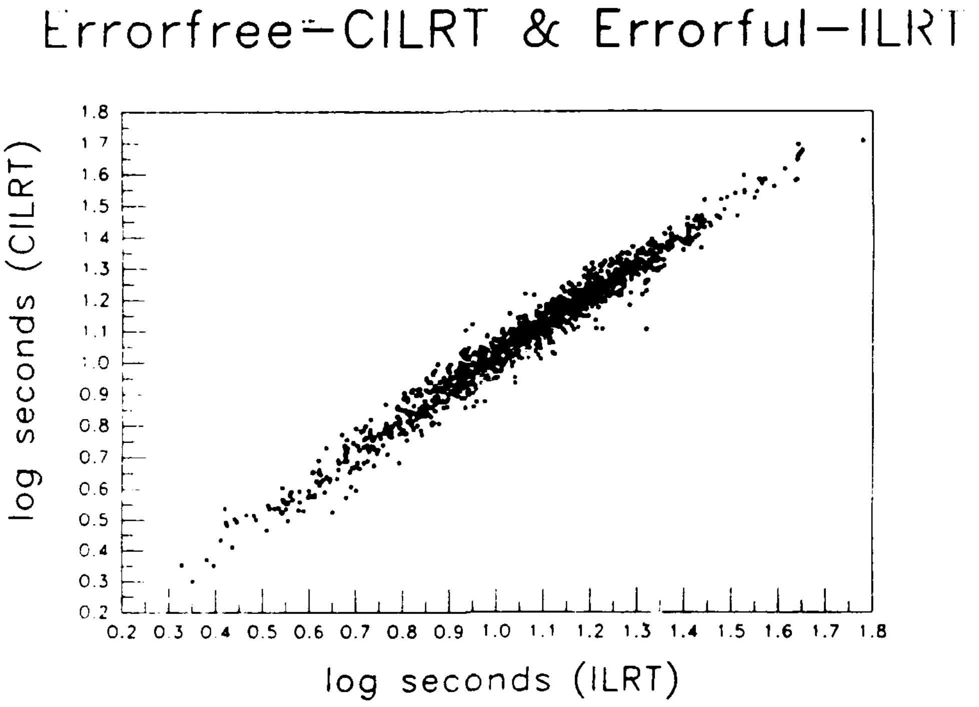


Figure 4. Scatter plot of correct log with all log integrate times.

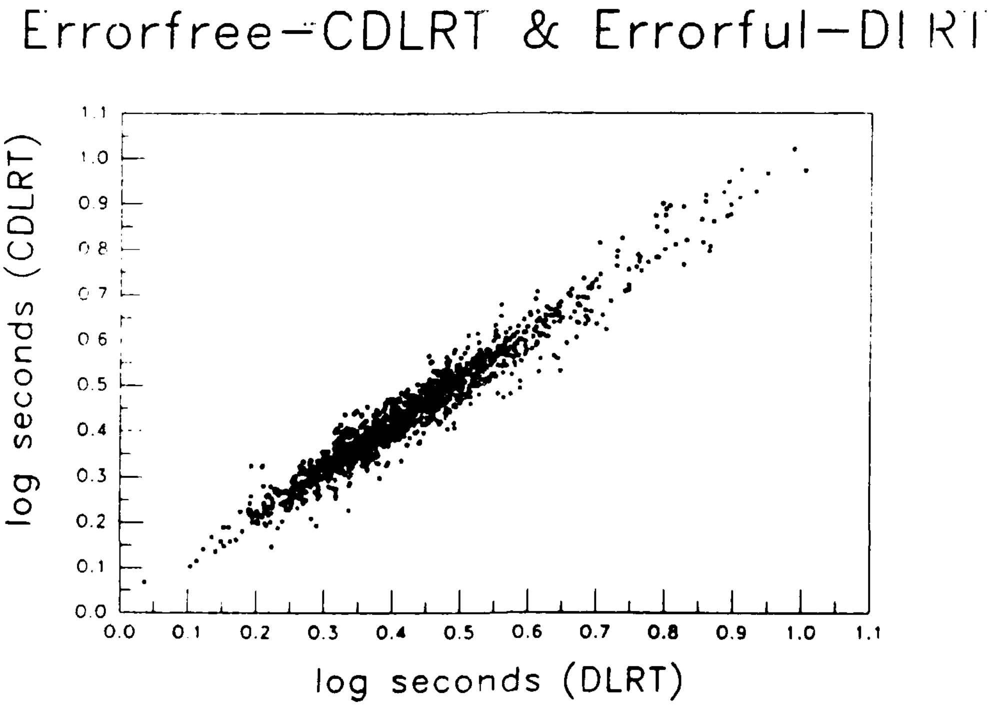


Figure 5. Scatter plot of correct log with all log decision times.

Table 10

Mean Log Correct Times and Mean Log All Times Distributional Comparisons (N = 1,388)

	Log Integrate Time		Log Decision Time	
Parameter	Correct	All	Correct	All
Mean	1.065	1.080	.433	.445
SD	.207	.208	.136	.138
Skewness	-.429	-.532	1.085	1.069
Kurtosis	.739	.811	2.874	2.767

Table 11

Reliabilities of Mean All and Mean Correct Trials for Log and Raw Latency Scores

Measure		Sample C		Sample D		Sample E	
		Log	Raw	Log	Raw	Log	Raw
Integrate	All	.98	.96	.97	.94	.94	.93
	Correct	.95	.93	.95	.92	.93	.91
Decision	All	.93	.92	.95	.93	.90	.87
	Correct	.88	.88	.91	.86	.86	.80

of estimating latency scores will be particularly apparent for low ability individuals who typically miss more than 35 percent of the items (mean accuracy is roughly 70%). Further, since lower ability subjects tend to miss the more difficult items, computing a mean latency score on correct items only amounts to computing a mean score on the easiest subset of items, making comparisons with high ability subjects questionable. Indeed, excluding incorrect trials makes any between subject comparisons of latency scores dubious since the denominator (number correct) will vary across subjects and the actual items entering into the mean estimate will also vary.

The foregoing section was a straight-forward evaluation of various ways to score each of the dependent measures separately but this leaves open the question of what summary measure should be used for the test. The fundamental problem is that a latency score from a test like Integrating Details cannot be directly used for selection and/or classification since individuals would begin responding as rapidly as possible. One simple solution would be to drop the latency measures from consideration in any future applied setting. However, this is unsatisfactory since the two latency measures have a clear theoretical interpretation and tap reliable (see Table 11) sources of unique variance--unique from the accuracy score and from one another as well. Table 12 contains the

Table 12

Intercorrelations of Dependent Measures
from Integrating Details
(N = 1,388)

Measure	IT	DT	Acc
Integrate Time (IT)	1.000		
Decision Time (DT)	.128	1.000	
Accuracy (Acc)	.448	-.024	1.000

intercorrelations among the three measures derived from Integrating Details. Note that Integrate Time and Accuracy are moderately and positively correlated but both of these are nearly independent of Decision Time.

The essential problem is to develop a summary score based on the three dependent measures that will retain as much information as possible from each of the measures while eliminating or minimizing easy and obvious response strategies. To best compare the relative merits and shortcomings of the summary scores to be presented, a brief digression is warranted to reemphasize the best strategy for accurate performance on this test. Theoretically, the best strategy for this test is to spend a great deal of careful time integrating the puzzle elements to form a clear, stable representation of the completed puzzle and then to rapidly encode, compare, and respond to the presented alternative (a rapid comparison is necessary to minimize surface structure decay and thus errors). Initial thought on the problem suggested three possible summary scores for evaluation. One summary score is a standard rate score, the number of seconds per unit of obtained Accuracy. This was computed by dividing Accuracy by the sum of the two latency scores (retransformed to raw seconds). A second rate score was calculated by dividing the sum of the two log latency measures by Accuracy, producing a score estimating the average number of log seconds needed for perfect Accuracy. A final summary measure was calculated by normalizing each of the dependent measures separately, then summing them. Note, in computing this, Decision Time was reflected since "fast" decisions are considered better.

Table 13 summarizes the results from the three summary measures. The first row reports the correlation between the Accuracy per Second rate score and Integrate Time, Decision Time, and Accuracy. This rate score has retained much of the information from the original latency measures but contains no information about Accuracy, which is unsatisfactory because this sco.'e could be maximized by fast performance alone. The second rate score (log seconds for perfect Accuracy) shows a very different pattern; a high correlation with Accuracy and a high negative relationship with Decision Time. This summary score is a substantial improvement since maximum performance is obtained by being accurate and deciding quickly. The sum of z-scores is contained in the last row. Note that this score retains the most information from Accuracy, relative to the other summary scores and relative to the two latency measures in this score, and large portions of the information from the two latency measures. This alone would make it the preferred summary score. However, notice also that the pattern of correlations with the original variables gives this summary score a further advantage. Specifically, this score is maximized by being accurate, spending a long time integrating the puzzle pieces and relatively little time in the decision portion of the problem. As indicated above, this is exactly the pattern of performance that will maximize success on the test.

Table 13

Correlations Between Summary Scores and Integrate Time, Decision Time, and Accuracy

Measure	Integrate Time	Decision Time	Accuracy
Accuracy per Second	-.781	-.409	.090
Seconds for Perfect Accuracy	-.282	-.527	.619
Sum of Z-Scores	.688	-.467	.766

This section has summarized a number of important enhancements to and suggestions for scoring Integrating Details. The data argue that for the latency scores, mean log latency measures based on all trials should be used. Using (common) logarithms corrects distributional abnormalities whereas using all trials (vs. correct only) increases interpretability and reliability. The final section discussed the need for devising a summary score that combines the information from all three variables. Several measures were proposed but the best measure appears to be the sum of z-scores from all three dependent measures. This score was best in that it retained the most information from the original dependent measures and is only maximized when an individual performs the test in the best possible way. However, further research is needed to determine an optimal summary score. The sum of z-scores is good but may encourage individuals to spend somewhat more time integrating the puzzle elements than is necessary, to boost the summary score. Also, this score may penalize individuals that are good at the task yet exceptionally fast integrating the puzzle pieces.

Reliability

The split-half reliabilities from all five samples are presented in Table 14. Note that the reliability estimates for the latency measures from Samples C, D, and E are based on mean log all times whereas the reliabilities for Samples A and B are for mean correct times. (Item latency data are not available for Samples A and B.) By and large, even with the differences between the samples, the magnitude of the reliabilities are quite similar with the exception of the Accuracy measure for Sample C (.58). It was the poor performance of this sample that suggested needed changes in the test and in the testing situation. The changes in the test have already been discussed. Although not made explicit, Integrating Details was not administered alone but always in the context of other experimental tests. The majority of the individuals in Sample C received Integrating Details as the last test in a long battery of other experimental tests (e.g., simple and choice reaction time, sentence verification, complex arithmetic, vocabulary, mental rotation, and others), which reduced their motivation. It should also be emphasized that the Navy recruits in this sample were in "boot camp," and thus quite tired, and were informed that the tests had no impact on their careers; therefore, they were generally unmotivated to perform. The changes in the testing situation included shortening the total session testing time (from 3 hours to approximately 1 hour) and testing in smaller groups allowing closer monitoring of examinees. The changes in the test and testing conditions where instituted with Sample D; note the improvement.

Table 14

Split-half Reliabilities of Integrating Details

Measure	Samp'e A	Sample B	Sample C	Sample D	Sample E
Integrate Time	.92	.98	.98	.97	.94
Decision Time	.94	.95	.93	.95	.90
Accuracy	.75	.73	.58	.79	.73
Sum of Z Scores	---	---	.87	.93	.89

The bottom row of Table 14 contains the split-half reliabilities for the sum of the z-scores for the three dependent measures, for Samples C, D, and E. These values are quite high for all three samples, even for Sample C, and thus provides additional support for developing such a summary score.

Table 15 contains the alternate form/alternate computer reliabilities for Integrating Details for Sample F. In this study, 150 Navy recruits were tested on the earliest 32 item Apple II version of the test, and a substantially updated 32 item Hewlett-Packard version. As reported in Table 6, the two versions of the test differed in terms of the minimum response times, instructions, items, and presentation vehicle (or type of computer). Each recruit took both versions of the test on the same day roughly 1.5 hours apart :. ith half of the subjects taking the Hewlett-Packard version first.[13]

Table 15

Alternate Form/Alternate Computer Reliability of Integrating Details

Measure	Sarnple F
Integrate Time	.68
Decision Time	.65
Accuracy	.70

[13] This study compared Apple II and Hewlett-Packard versions of three tests: Integrating Details, Mental Rotation, and Intercept. The second two tests are described in Hunt et al. (1987). Therefore, each recruit was actuallv tested for approximatelv 3 hours, which is longer than generally desirable, as noted earlier, and this may have lowered the reliabilities.

Test-retest reliabilities are reported in Table 16 for the subset of subjects from Samples C and E that returned for the second testing. The delay period between the two testings was a minimum of 4 weeks and was just over 5 weeks for a small minority of individuals. The two latency measures produced roughly equivalent values across the two samples. Note that the latency retest estimates are much lower than the corresponding split-half measures reported earlier (Table 14). There is very little literature on the long term stability of latency measures. However, there is an extensive literature on psychological and physiological influences on latency measures, such as practice, motivation, fatigue, and blood sugar level, which suggest that there is a great deal of situational specificity in latency measures. This appears to be borne out by the large difference between the split-half and retest reliability estimates. Indeed, the average difference between the within-session and across-session estimates is .31, suggesting that 31 percent of the reliable variance in each latency measure is situationally specific.

Table 16

Retest Reliability of Integrating Details

Measure	Sample C'	Sample E'
Integrate Time	.63	.66
Decision Time	.63	.67
Accuracy	.57	.74
Sum of Z-Scores	.68	.80

The retest reliability of the accuracy measure is quite reasonable for Sample E (.74) but disappointingly low for Sample C (.57). The probable reasons for the poor performance of Sample C were discussed earlier. Focusing attention on the retest reliability of Sample E (.74), it should be noted that the split-half reliability of the accuracy measure for the first session was .73 and .75 for the second testing. This shows that, in contrast to the latency measures, accuracy on Integrating Details is as stable within a session as it is across time. The last row of Table 16 contains the retest reliability for the z-score summary measure. The values are quite good, showing a large degree of stability, even for Sample C.

Subgroup Differences

To date, only gender differences have been assessed and then only for Samples A and E. Historically, claims have been made that spatial tests favor males, although the data have been inconsistent. More recently, a growing body of evidence suggests that gender differences in spatial ability are actually related to the test complexity-speededness dimensions. Specifically, females tend to out perform males on Perceptual Speed tests, do less well than males on tests of Spatial Relations, and perform as well as males on complex Visualization tests (e.g., Alderton & Pellegrino, 1985; Linn & Peterson, 1986). If this is so, then there should only be a small correlation between gender and performance on Integrating Details. Table 17 presents the point biserial correlations between each of the Integrating Details measures and gender, with males coded as 1 and females as 0. There is essentially no relationship between gender and the latency measures. Accuracy shows a

Table 17

Gender Differences on Integrating Details
Point Biserial Correlations

Measure	Sample A	Sample E
Integrate Time	.01	-.09
Decision Time	.10	-.09
Accuracy	.17	.15
Sum of Z-Scores	---	.07

small tendency for males to out perform females. However, the .17 and .15 values are much smaller than the point biserial correlations between two tests of mental rotation and gender from Sample A (.26 and .24). Finally, the sum of z-scores shows a very minor relationship with gender.

Practice Effects

The retest data from Samples C and E can be used to ascertain the effects of practice; these are summarized in Table 18. The M-Diff column contains the difference between the two means (first session mean minus second session mean). This provides a measure of the absolute change with repeated test administrations. The M-Gain was calculated as the difference between performance on the first and second sessions (Mean-1 minus Mean-2), divided by the standard deviation of the first testing ((M1-M2)/SD1), and thus shows a relative performance change. All measures show some improvement but only the two latency measures from Sample E were statistically significant. The SD1/SD2 quantity is the ratio of the standard deviations (first divided by second). The fact that the standard deviation ratios are all near one indicates that although there were mean gains on the second testing, practice did not reduce individual differences.

Table 18

Summary of Practice Effects for Retest Samples on Integrating Details

	Sample C'			Sample E'		
Measure	M-Diff	M-Gain	SD1/SD2	M-Diff	M-Gain	SD1/SD2
Integrate Time	.050	.258	.997	.143	.721	1.045
Decision Time	.038	.265	1.009	.111	.808	1.055
Accuracy	-.017	-.142	.929	-.009	-.063	.999
Sum of Z-Scores	.056	.033	.901	-.043	-.022	1.025

The data from Sample E' can be used for a more careful evaluation of practice effects by subdividing the data into sequential blocks of eight trials. Each block of eight trials represents a single replication of the item design: 2 trials each of 3, 4, 5, and 6 piece problems with one member of each pair requiring a same response and the other member requiring a different response. This subdivision was conducted on both the first and second testing sessions thereby producing eight sequential blocks of eight trials. For each block, log Integrate Time, log Decision Time, and Accuracy were computed. These values are plotted in Figure 6. (To facilitate comparisons between latency and accuracy measures, Accuracy is plotted in the Figure using the 0.0 to 1.0 portion of the scale to represent proportion correct.) In the Figure, blocks 1 to 4 are from the first testing while blocks 5 to 8 are from the second testing. The Figure clearly reveals that the two latency measures become sequentially faster with practice, both within and across testing sessions. This is consistent with other data in the literature (e.g., Ackerman, 1986, 1987; Adams, 1987; Alderton, Pellegrino & Lydiatt, 1988; Fleishman, 1972; Fleishman & Hempel, 1955). There is some evidence that speed gains are leveling out between blocks 5 and 7 but both measures show an additional drop in the final block. For log Integrate and log Decision Times the null hypothesis for a practice effect was easily rejected (F's (7,2590) $\geq$ 181.94, $p < .001$. In contrast, Accuracy shows no change within or across testing sessions; a statistical test across all eight blocks did not reject the null hypothesis (F (7,2590) = 1.59, $p = .133$).

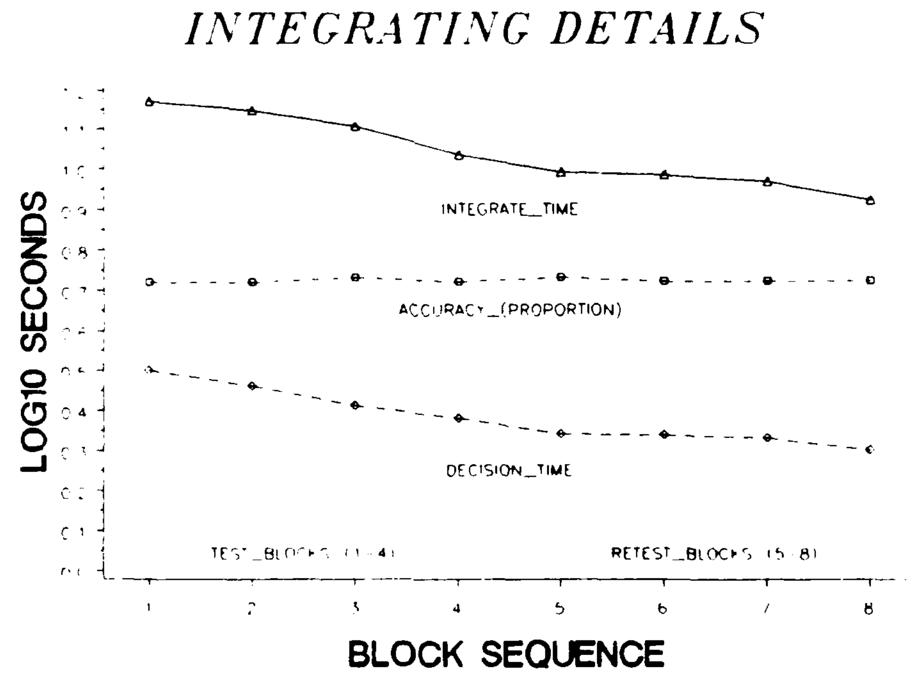


Figure 6. Plot of Integrating Details measures by block sequence (Sample E').

Construct Validity

In this report, the discussion of validity will be limited to construct validity. Two types of construct validity data to be reported. Internal validity, or construct representation (Embretson, 1983), refers to how well the test performs in terms of the performance model (specified earlier). External validity, or nomothetic span (Embretson, 1983), concerns the relationship between test performance and performance on other ability measures.

Internal Validity. The data from Sample D was used to evaluate the information processing performance mode for the test.[14] The performance model for Integrate Time states that individuals encode an element, search the array for the matching letter, then synthesize the two pieces at the correct location and store the product in memory. Therefore, Integrate Time should chiefly be a function of the number of iterations of the encode-search-synthesize sequence. Across the items, the correlation between the number of puzzle pieces and Integrate Time was .93 (N points = 60; $\underline{t}(58) = \underline{p} < .001$). In contrast, the model predicts that Decision Time should only be a minor function of the number of pieces synthesized. The correlation between Decision Time and the number of synthesized pieces was a modest .50 ($\underline{t}(58) = 5.21$, $\underline{p} < .001$). The model also predicts that one source of errors for these problems should be the number of times the encode-search-synthesize sequence must be executed. This was supported by the -.63 ($\underline{t}(58) = -6.11$, $\underline{p} < .001$) correlation between accuracy and the number of pieces.

While these data provide substantial support for the validity of the model there is clearly room for improved prediction, especially for Accuracy and Decision time. As indicated in the discussion of the model, several additional factors should affect both Accuracy and the difficulty of deciding if the presented alternative matches the one constructed. One such factor is the complexity or irregularity of the completed object, which is a function of the number of vertices, the number of oblique lines, the number of indentations (or concavities), and the presence or absence of symmetry. Final object complexity should impact item solution in two ways. With highly complex objects, it should be more difficult to maintain an accurate and good quality surface structure representation which should increase the likelihood of making errors. More complex objects should also require more Decision Time because of the number of features that must be evaluated, this further increases the likelihood of making errors (based on inaccurate feature comparisons). For trials where the presented alternative could not be made from the puzzle pieces, the similarity between the correct and presented alternative should influence Accuracy and to some extent Decision Time. To test these hypotheses, a sample of 219 junior college students rated the complexity of the final shapes. Complexity ratings were obtained using a discrete seven point scale, with 1 being "very simple" and 7 being "very complex." For the 30 different response items, these same subjects rated the similarity between the 30 presented alternatives and the corresponding object that would have been made from the puzzle pieces. Similarity ratings were obtained using a discrete seven point scale, with 1 being "very similar" and 7 being "very dissimilar." These subjects never took the actual test. The split-half

[14] The data from Sample D was chosen because of the larger range of item difficulty (2-6 vs. 3-6) and because the items used for the sample represents a superset of all other samples.

reliability for the complexity ratings was .97; the reliability for the similarity ratings was .88.[15]

These complexity and similarity ratings were used to predict average Accuracy, Integrate Time, and Decision Time for the 60 items administered to the 427 subjects from Sample D. The data were analyzed once for the same and different trials separately, then a combined analysis was performed. (The separate analyses were required since similarity has no meaning for same judgement trials.) To facilitate comparisons, Table 19 contains the simple correlations between the three Integrating Details measures and the number of pieces, calculated separately for same and different trial items. (These are located in the second and last columns of Table 19. The entries in the three middle columns labeled "Same," "Same & Diff," and "Diff" contain adjusted multiple R values.) The three Integrating Details measures, derived from same response items, were separately entered into multiple regression equations with the number of pieces and final shape complexity as the predictors. As can be seen in Table 19, for Integrate Time, item complexity added little to the prediction over that provided by the number of pieces alone (.94 vs. .92). In contrast, the addition of item complexity greatly increased the prediction of Decision Time (.72 vs. .57). Item complexity ratings also improved the prediction of same trial Accuracy (.65 vs. -.58). The different response items were used in the second set of analyses. In these regressions, each Integrating Details measure was predicted by the number of pieces, the complexity of the alternative, and the similarity between the presented and correct alternatives. For Integrate Time there was a small increase in predictability by adding in complexity and similarity (.96 vs. .90). As expected, since the alternative had not yet been presented during this stage of processing, the contribution of similarity to Integrate Time was nonsignificant. In contrast, a large increase in predicting Decision Time (.81 vs. .59) was gained by adding in complexity and similarity (both received significant regression weights). Finally, the prediction of different item Accuracy was also substantially improved by adding in complexity and similarity (.85 vs. -.70). For the final set of analyses same and different trial items were combined. Each of the three Integrating Details measures were entered into regression equations with the number of pieces and final object complexity as predictors. Integrate Time was only slightly better predicted using optimal weights for the two predictors than using the number of pieces alone (.94 vs. .93). Decision Time prediction was substantially improved by adding in item complexity (.71 vs. .50). Accuracy was also better predicted by using the number of pieces and complexity (.70) as opposed to using the number of pieces alone (-63).

[15] The reliability for the complexity ratings was determined by dividing the items into odd and even halves based on the original item design characteristics. Specifically, the odd half contained the final objects for half of the 2, 3, 4, 5, and 6 piece same problems and half of the 2, 3, 4, 5, and 6 piece different problems; the even half was similarly constructed. Mean complexity ratings were then calculated for each half and for each subject. Finally, the odd and even halves were correlated, the raw correlation was then corrected using the Spearman-Brown half-length formula. The similarity rating were handled quite differently. Since there were only 30 pairs to be rate, it was decided that each correct and presented alternative pair would be administered twice--once with the presented alternative given as the first member of the pair and again with the presented alternative given as the second member of the pair; this created 60 items to be rated. The reliability for the similarity ratings was then calculated by correlating ratings of the presented-correct pair with the correct-presented pair. Thus, the reported reliability is the raw, uncorrected correlation.

Table 19

Prediction of Item Performance

Measure	Same No. Pieces (N = 30)	Same (N = 30)	Same & Diff (N = 60)	Diff (N = 30)	Diff No. Pieces (N = 30)
Integrate Time	.92	.94	.94	.96	.90
Decision Time	.57	.72	.71	.81	.59
Accuracy	-.58	.65	.70	.85	-.69
	Simple r	Adjusted Multiple R			Simple r

These analyses provide substantial support for the underlying information processing model, and the hypotheses concerning sources of item difficulty. As anticipated, Integrate Time is principally a function of the number of puzzle pieces which itself is a function of the number of encode-search-synthesize cycles. Decision Time is only poorly related to the number of puzzle pieces. However, substantial improvements in predicting Decision Time are gained by adding in the complexity of the final object that must be compared. Prediction is further improved when the similarity between the presented and correct alternative is added to the equation for different trials (where it is meaningful). Item response Accuracy is moderately accounted for by the number of encode-search-synthesize cycles but this prediction is ameliorated by adding in the complexity of the final object, and further improved when the similarity between the presented and correct alternative is taken into consideration (for different response items). It should be emphasized that the predictors used in these analyses were obtained from an independent sample that never saw a single test item.

The previous analyses validate the model at the group performance level but this does not necessarily mean that the model is capable of capturing the problem solving behavior of individual subjects. A separate set of analyses were conducted to determine the degree to which the model was capable of capturing the performance of individual subjects. The data from Sample D were used. Because of the inherent instability of single trial or item level data, the data for each subject was aggregated by the number of pieces across same and different trials. Thus for each subject, mean Integrate Time, Decision Time, and Accuracy was estimated for 2, 3, 4, 5, and 6 piece puzzle problems; each of the 5 points were determined by 12 trials. While this greatly increases the stability of the data, only questions concerning the effect of the number of puzzle pieces can be answered. The five Integrate Time means (for 2, 3, 4, 5, and 6 piece puzzles) were correlated with the number of puzzle pieces producing an estimate of model fit (*r*) for each subject. Across subjects, the average correlation between the number of pieces and Integrate Time was .904. Thus, on the average, the model was performing quite well at the individual subject level. To better understand the extent to which the model was fitting individual subjects, the model fit statistic (*r*) was recorded to a 0-1 variable with a 0 meaning that the subject's data was non-significantly fit by the model, and a 1 meaning that it was significantly fit (a simpler *r* > .805 is significant at the .05 level with three degrees-of-freedom). This new variable showed that over 89.2 percent of all individuals were significantly fit by the model. The same model was fit to Decision Time as well. Remember, Decision Time is expected to be only moderately correlated with the number of pieces. The average model fit was a modest .492, substantially lower than that for Integrate Time and comparable to that obtained at the group level (.50). The model fit statistic was recoded, as before, to a 0-1 variable and showed that only 27.4 percent of all subjects were significantly fit, further confirming the validity of the model. Finally, Accuracy was correlated with the number of puzzle elements for each subject. It was hypothesized that Accuracy should be less related to the number of puzzle elements than Integrate Time but more related to puzzle elements than Decision Time. This was exactly the case, the average r for Accuracy was -.58, similar to the value obtained at the group level (-.63), and 38.7 percent of all subjects' were significantly fit by the model. These values clearly fall between those of Integrate Time and Decision Time.

In summary, the information processing performance model developed for Integrating Details was substantially validated at both the group and individual subject levels. This type of construct validity is crucial for interpreting patterns of correlations between performance on this test and other criteria. Further work is required that will allow a strict quantification of "complexity" and "similarity" to set parameters for future item development and alternate form equating.

External Validity. External validity concerns the relationship between performance on a test and performance on tasks external to the test, such as other ability measures. It is important to show that Integrating Details is related to other measures of spatial ability (convergent validity) yet relatively unrelated to measures such as verbal and numerical ability (divergent validity).

An extensive evaluation of the external validity of Integrating Details is possible with the data from Sample A. This has been reported in detail elsewhere (Hunt et al., 1987, 1988) and will only be summarized here. Table 20 contains correlations between the Integrating Details measures and scores from seven paper-and-pencil tests. The seven tests are: Raven's Advanced Progressive Matrices (APM), Differential Aptitude Test battery's Space test (DAT), Guilford-Zimmerman Spatial Orientation test (GZ; Guilford & Zimmerman, 1947), Primary Mental Abilities Space test (PMA), Lansman's adaptation of the Vandenberg Mental Rotation test (Rotate; Lansman, 1981; Vandenberg & Kuse, 1978), Identical Pictures test (Ident.), and the Nelson-Denny vocabulary test (Voc; Brown, Bennett, & Hanna, 1981). Excluding the vocabulary test, the remaining tests are spatial tests and are ordered roughly in terms of the complexity of the items employed (complexity decreases from left to right). As expected, the correlation between Integrating Details Accuracy and the test scores decrease as the complexity of the test items decrease. This provides strong evidence that Integrating Details is a test of complex spatial problem solving. The correlation between the two latency measures and the accuracy scores from the other tests are generally small and inconsistent. Thus, these measures are relatively independent of spatial problem solving accuracy as traditionally measured. Note that the Integrating Details measures are nearly uncorrelated with the vocabulary test score which further supports the argument that Integrating Details requires principally spatial-analog processing and not verbal-propositional processing.

In addition to the paper-and-pencil tests, the subjects from Sample A also solved four other computer administered spatial tests. These tests (referred to as "static" tasks in the original Hunt et al. reports) provided measures of item solution time as well as accuracy. The details of the tests are not pertinent but can be found in the original reports. What is pertinent is the result of the factor analysis that was performed on 12 measures, 7 latency measures and 5 accuracy measures, including the Integrating Details measures. A spatial accuracy (or power) factor and a spatial latency (or speed) factor cleanly emerged from the analysis. This is exactly what was expected based on the literature summary provided earlier. Importantly, Integrating Details Accuracy had the

Table 20

Correlations Between Integrating Details and Paper-and-Pencil Tests (N = 170)

Measure	APM	DAT	GZ	PMA	Rotate	Ident.	Voc
Integrate Time	.15	.04	.11	.02	.06	.11	.10
Decision Time	.03	-.18	.03	-.18	-.12	.20	.04
Accuracy	.57	.57	.41	.34	.30	.16	.18

second highest projection on the spatial accuracy factor, and the highest loadings on the spatial speed factor. This is additional evidence that Integrating Details is a complex spatial processing test. The two factor scores were also used in a set correlation to predict the intercorrelations among the paper-and-pencil spatial tests; $\underline{R}^2 = .82$.

To summarize the results from Sample A, Integrating Details was highly correlated with performance on paper-and-pencil test of spatial ability, and the highest correlations were with the most complex tests. The factor analysis of the accuracy and latency measures, from the computer-controlled spatial tests, revealed that Integrating Details helped to centrally define both the expected power (accuracy) and speed (latency) spatial factors. In turn, these two factors were capable of accounting for 82 percent of the common variance among the paper-and-pencil spatial tests. Finally, Integrating Details was virtually independent of verbal ability. Together, these data argue for the spatial nature of Integrating Details and for interpreting the test as a complex measure of spatial processing that is central to spatial ability.

Sample E provides additional data concerning external validity. Table 21 contains correlations between the Integrating Details measures, including the sum of z-scores, with latency and accuracy measures from a computer administered mental rotation spatial test, the accuracy score from a simple video game that required "shooting" a target moving across a computer monitor (Intercept; see Hunt et al. 1987, for details) and percentile standing in high school (high numbers are better). This is a 332 subject subset of Sample E for which high school standing was available from academic records. Notice that Integrating Details is significantly correlated (all $\underline{r} > .106$, have $\underline{p} < .05$, two-tailed) with performance on the mental rotation test and the dynamic spatial video game. The Integrating Details latency measures are more correlated with rotation speed than rotation accuracy but the reverse is true for Integrating Details Accuracy and the sum of z-scores. High school standing is highly correlated with Integrate Time, Accuracy, and the sum of z-scores. This is predictable given the complexity of Integrating Details and the previously reported correlation between it and Raven's APM, a measure of general intellectual ability. These data argue that Integrating Details is indeed a spatial processing test, and its correlation with high school standing attests to its complexity and general intellectual demands.

Table 22 contains the correlations between the Integrating Details measures and the Armed Service Vocational Aptitude Battery (ASVAB). The data are for 818 individuals, subsets of Samples C and D for which standardized ASVAB scores were available. The subtest abbreviations translate as follows: GS is General Science, AR is Arithmetic

Table 21

Correlations Between Integration Details, Two Spatial Processing Tests, and High School Standing (N = 332)

	Rotation			
Measure	Accuracy	Latency	Intercept	Standing
Integrate Time	.158	.196	.123	.242
Decision Time	.047	.263	-.097	-.015
Accuracy	.279	-.098	.248	.365
Sum of Z-Scores	.197	-.112	.237	.316

Table 22

Correlations Between Integrating Details and the ASVAB (N = 818)

		Math		Verbal		Technical			Speed		Factor	
Measure	GS	AR	MK	WK	PC	MC	AS	EI	NO	CS	G	S
Integrate Time	.19	.22	.23	.17	.13	.26	.14	.19	-.07	-.01	.26	-.05
Decision Time	-.07	-.03	-.00	-.09	-.13	-.08	-.11	-.04	-.06	-.05	-.09	-.06
Accuracy	.27	.32	.28	.26	.18	.37	.23	.29	-.02	.10	.37	.05
Sum of Z-Scores	.28	.30	.27	.27	.23	.37	.25	.27	-.02	.07	.37	.03

Reasoning, MK is Math Knowledge, WK is Word Knowledge, PC is Paragraph Comprehension, MC is Mechanical Reasoning, AS is Auto-Shop Information, EI is Electronics Information, NO is Numerical Operations, and CS is Coding Speed. The 10 subtests are grouped by content area, with General Science (GS) left alone. The last two columns of Table 22 contain unit-weighted factor scores, G, or General ability, is the sum of standardized scores from the eight power tests (GS through EI in Table 22), this score is typically indistinguishable from scores derived using the first principal component of the ASVAB. S, or Speed, is the sum of the standard scores from NO and CS, and is similar to a clerical speed and accuracy score (usually the second principal component of the ASVAB). Note, first of all, that none of the Integrating Details measures are related to the speeded subtests or the Speed factor score. Decision Time is unrelated to any of the ASVAB subtests or factor scores. Integrate Time, Accuracy, and the summed z-score measure are moderately correlated with each of the eight power tests but in a predictable pattern. The lowest correlations are between the Integrating Details measures and the highly verbal Paragraph Comprehension test, whereas the highest correlations are with Mechanical Comprehension, the only test with any spatial processing requirements. Integrate Time, Accuracy, and the z-score sum are also correlated with the general (G) factor score.

While these simple correlations are informative an overall estimate of the degree to which *Integrating Details* is unique from the ASVAB collectively is desirable. This is crucial data since, if adopted, *Integrating Details* would be used for selection and classification along with the ASVAB. Such a statistic can be generated by first predicting each of the *Integrating Details* measures with the ASVAB tests, and then subtracting the resulting R^2 value from the reliability of the dependent measure (i.e., $r_{xx} - R^2 =$ Uniqueness). This uniqueness statistic is the estimated proportion of true score (or reliable) variance that is unique from the ASVAB. This is also the pool of variance that is available for predicting criterion performance independent of the ASVAB. The necessary R^2 value can be obtained either by using all 10 ASVAB subtests, or by using the two unit-weighted factor scores (G and S). In general, the latter procedure is preferable since the individual ASVAB tests are highly intercorrelated. Such multicollinearity creates enormous instability in the regression coefficients, made even worse by a high degree of capitalization on chance fluctuations in the correlations. Together, this results in generally unreproducable regression solutions, both for samples of the same size but especially for smaller samples (e.g., Cohen & Cohen, 1983). Therefore, the R^2 values used to compute uniqueness are those derived by predicting the *Integrating Details* measures from the two unit-weighted factor scores (G and S); the actual R^2 values used are also adjusted for shrinkage.[16]

Table 23 contains the uniqueness statistics for Samples D and C. For Sample C uniqueness is reported based on split-half and retest reliabilities. Note that even though the split-half reliabilities for the two samples are quite different (see Table 14), the uniqueness values for all of the *Integrating Details* measures are nearly identical across the two samples. As expected, when based on split-half reliability, the latency measures are substantially more unique from the ASVAB than are either the Accuracy score or the z-score sum, but the retest uniqueness is more comparable. For the Accuracy score, nearly 50 percent of the reliable variance is unaccounted for by the ASVAB, this is true for split-half and retest based uniqueness estimates. The z-score sum also shows substantial ASVAB uniqueness based on both split-half and retest reliabilities. These values importantly demonstrate that roughly half of the reliable variance in *Integrating Details* is unique from anything that the ASVAB is measuring, and thus there is ample opportunity for the test to augment the predictive validity of the ASVAB.

The external validity of *Integrating Details* is quite promising. The test is chiefly correlated with measures of spatial ability and most highly correlated with complex spatial tests. *Integrating Details* is relatively independent of verbal ability. Thus, across a number of samples, the test demonstrates both convergent and divergent validity. The complexity of the test is heartily supported by the correlations with measures of general intellectual ability such as Raven's APM, high school standing and the ASVAB G score. Importantly, roughly half of the reliable variance in *Integrating Details* was shown to be independent of what the ASVAB is measuring, thus there is ample opportunity for the test to augment the predictive validity of the ASVAB.

[16] Note that the actual difference between the unadjusted R^2 value based on 10 predictors and the adjusted R^2 value based on 2 predictors (the extreme cases) ranges from .026 to .033 in the data to be reported in Table 23. Therefore, if uniqueness comparisons for unadjusted R^2 values based on 10 predictors are needed, subtracting a constant .03 from the values reported in Table 23 will be quite close.

Table 23

ASVAB Uniqueness Statistics for Samples C, C', and D

Measure	Using Split-Half		Using Retest
	Sample D	Sample C	Sample C'
Integrate Time	.89	.89	.54
Decision Time	.93	.93	.63
Accuracy	.48	.48	.46
Sum of Z-Scores	.71	.77	.58

Results Summary

Integrating Details has been extensively evaluated on a number of samples comprising a broad cross section of individuals. Across the samples, initial results showed that performance, as indicated by means and standard deviations, was quite consistent. However, a number of changes in the test were called for including: altering the instructions, setting minimum response times, replacing several items, adding simpler items, and reordering items. Other more administrative changes were also made, such as, changing the computer used for presentation and limiting the number of other tests given in the same time period. Empirical investigations demonstrated that test scoring should be changed. Specifically, the data suggested that for both Integrate and Decision Times, raw latencies should be logarithmically transformed (common logarithms were selected) and that all trials should be used in estimating mean latency performance for each subject, whether the correct answer was given or not. The log transform improved the distributional properties of the measures while using all trials improved interpretability; both changes increased reliability. The reliability of all three measures were found to be adequate. The latency measures showed a fairly large difference between split-half and retest reliability estimates; it was argued that this may not be a problem with the test but, rather, it is the nature of latency scores in general. The retest reliability of the Accuracy measure was (nearly) identical to its split-half estimates. Using the retest data, small practice effects were observed but they only reached significance for the latency measures in one sample. Unlike some spatial tests, Integrating Details showed only a small gender bias, favoring males. The construct validity of the test was extensively evaluated. In terms of the underlying information processing model, the model proved to be quite valid at both the group and individual subject levels. A number of a priori hypotheses concerning performance on the test were tested and supported. The pattern of correlations between Integrating Details and other ability measures support the interpretation of the test as a complex spatial problem solving test. The test was highly correlated with other spatial tests, especially the most complex ones that helped define both the speed and power spatial dimensions, and was relatively uncorrelated with verbal ability. Integrating Details was found to be uncorrelated with the speed dimension of the ASVAB but correlated with the general ability factor. However, roughly 50 percent of the reliable variance in each of the test's measures were found to be independent of the ASVAB; this is crucial, since this is the pool of variance available for capturing school and job performance not predictable from the ASVAB.

RECOMMENDATIONS

The reported work on Integrating Details demonstrates that the test has a strong theoretical basis and reliably measures a dimension of human intelligence that has been shown valid for both school and job performance, and that this is an ability class that is not presently assessed by the ASVAB. Therefore, the Integrating Details test is strongly recommended for advanced development and validation.

Based on the research reported herein, a final 40 item version of the test has been developed and is currently ready for administration on Hewlett-Packard Integral Personal Computers; this is the version that is most strongly recommended for further work. The 40 items are the best performing subset of items used to date. The items are divided into four blocks of 10 items each. Each block contains two each of 2, 3, 4, 5, and 6 piece puzzle problems, with one member of each pair requiring a different response. Each block begins with a 2-piece puzzle problem to engender motivation and a sense of accomplishment. Total time (test plus instructions) is approximately 25 minutes. Actual problem solving time is about 16.5 minutes but there is a large range; from about 10 to 35 minutes. Instruction time, plus five practice problems, takes from 4 to 10 minutes. This version of the test incorporates the various changes in instructions, noted elsewhere, and contains the 1.25 and .5 minimum response times set for Integrate and Decision Times, respectively. When the test is scored, all 40 item latencies, correct and incorrect, are to be transformed using common logarithms and used in computing mean estimates for each subject.

Future Work

There are four areas of future work that must be undertaken for Integrating Details to mature to a test suitable for implementation. First, it is crucial that research be undertaken to develop an optimal measure that combines the information from all three dependent measures, Integrate Time, Decision Time, and Accuracy. The sum of z-scores is only a starting point, but it demonstrates the desirable qualities of a combined summary score. These qualities include, the retention of large amounts of information from the original measures, the elimination of easily coachable strategies while directing examinee's toward the optimal solution strategy, reduced practice and gender effects, and increased reliability. Additional measures are currently under consideration but this may require solicitation of outside expert assistance.

More research is required to understand the details of test performance and item difficulty. Specifically, research is needed to elucidate the decision portion of item solution. This research should help quantify the dimensions of item "complexity" and "similarity" shown important earlier. This information is crucial for the development of additional items and parallel alternate forms. Currently, two lines of effort are directed at this problem. Beginning in July 1988 an in-house study will be conducted to collect verbal solution protocols from subjects solving Integrating Details items. This should provide important insights about the sequence of mental operations and sources of item difficulty in all stages of problem solution. A second line of inquiry involves scoring the Physical characteristics of each item, in terms of area, number of oblique lines, cardinality of shape, number of convolutions, etc. to determine what dimensions predict ratings of item complexity and similarity; these same dimensions will then be used to predict item difficulty. Collectively, these data will be used to expand the model of item solution and develop alternate forms of Integrating Details.

The third area of research concerns practice on spatial processing tasks. There is some evidence in the psychological literature that practice improves performance. Since some changes with practice were found on Integrating Details, additional research is needed to determine the nature and durability of these effects. Because of the difficulties of collecting multiple session data on Navy recruits, a research contract should be awarded to assess the locus and durability of practice effects on Integrating Details.

The final thrust for further research is the collection of large amounts of data to provide accurate normative scores and to assess the predictive validity of Integrating Details for both school and job performance. Some of this work is underway and being conducted in cooperation with other Navy and Joint-Service projects. However, to meet reasonable time lines a greatly increased effort is required, but this is contingent on future funding.

REFERENCES

Ackerman, P. L. (1986). Individual differences in information processing: An investigation of intellectual abilities and task performance during practice. Intelligence, 10, 101-139.

Adams, J. A. (1987). Historical review and appraisal of research on the learning, retention, and transfer of human motor skills. Psychological Bulletin, 101, 47-74.

Alderton, D. L. (1986). Individual differences in surface development problem solving: Analyses of the contribution of latency, accuracy, and strategy (Unpublished doctorial dissertation). Santa Barbara, CA: University of California, Santa Barbara.

Alderton, D. L., & Pellegrino, J. W. (1985, August). Sex differences in spatial ability: Componential analyses of process differences. Paper presented at the Annual Meeting of the American Psychological Association, Los Angeles, CA.

Alderton, D. L., & Pellegrino, J. W. (1984). Analysis of mental paper folding (Unpublished manuscript). Santa Barbara, CA: University of California, Santa Barbara.

Alderton, D. L., Pellegrino, J. W., & Lydiatt, D. (1988). Effects of extended practice on spatial information processing and spatial aptitude (Unpublished manuscript). San Diego: Navy Personnel Research and Development Center.

Anderson, J. R. (983). The architecture of cognition. Cambridge, MA: Harvard University Press.

Bennett, F. K., Seashore, H. G., & Wesman, A. G. (1974). Manual for the Differential Aptitude Test (5th ed.). New York: The Psychological Corporation.

Brown, J. L., Bennett, J. M., & Hanna, G. (1981). Nelson-Denny Reading Test. Riverside, CA: Riverside Publishing Co.

Burt, C. (1949a). The structure of the mind: A review of the results of factor analysis (Part I). British Journal of Educational Psychology, 19, 100-111).

Burt, C. (1949b). The structure of the mind: A review of the results of factor analysis (Part II). British Journal of Educational Psychology, 19, 176-199.

Carroll, J. B. (1976). Psychometric tests as cognitive tasks: A new "structure of intellect." In L. B. Resnick (Ed.), The nature of intelligence, 27-56. Hillsdale, NJ: Erlbaum.

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

Chase, W. G. (1986). Visual information processing. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), Handbook of perception and human performance: Cognitive processes and performance, II, 28:1-28:71. New York: Wiley.

Cohen, J., & Cohen, P. (1983). Applied multiple regression/correlation analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Cronbach, L. J. (1970). Essentials of psychological testing (3rd ed.). New York: Harper and Row.

Ekstrom, R. B., French, J. W., & Harman, H. H. (1976). Manual for the kit of factor-referenced cognitive tests. Princeton, NJ: Educational Testing Service.

Eliot, J. C., & Smith, I. M. (1983). International directory of spatial tests. Windsor, England: NFER-Nelson.

Embretson (Whitely), S. (1983). Construct validity: Construct representation versus nomothetic span. Psychological Bulletin, 93, 179-197.

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

Farah, M. J. (1984). The neurological basis of mental imagery: A computational analysis. Cognition, 18, 245-271.

Farah, M. J. (1985). Psychophysical evidence for a shared representational medium for mental images and percepts. Journal of Experimental Psychology: General, 114, 91-103.

Finke, R. A. (1980). Levels of equivalence in imagery and perception. Psychological Review, 87, 113-132.

Finke, R. A. (1985). Theories relating mental imagery to perception. Psychological Bulletin, 98, 236-259.

Finke, R. A., & Shepard, R. N. (1986). Visual functions of mental imagery. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), Handbook of perception and human performance: Cognitive processes and performance, II, 37:1-37:55. New York: Wiley.

Fleishman, E. A. (1972). On the relationship between abilities, learning, and human performance. American Psychologist, 27, 1018-1032.

Fleishman, E. A., & Hempel, W. E. (1955). The relation in a visual discrimination task between abilities and improvement with practice. Journal of Experimental Psychology, 49, 301-310.

Ghiselli, E. E. (1973). The validity of aptitude tests in personnel selection. Personnel Psychology, 26, 461-477.

Guilford, J. P., & Zimmerman, W. S. (1947). Guilford-Zimmerman aptitude survey. Orange, CA: Sheridan Psychological Services.

Guttman, L. (1954). A new approach to factor analysis: The radex. In P. F. Lazerfield (Ed.), Mathematical thinking in the social sciences, 258-348. Glencoe, IL: Free Press.

Hunt, E., Pellegrino, J. W., Abate, R., Alderton, D. L., Farr, S. A., Frick, R. W., & McDonald, T. P. (1987). Computer-controlled testing of visual-spatial ability (NPRDC TR 87-31). San Diego: Navy Personnel Research and Development Center.

Hunt, E., Pellegrino, J. W., Frick, R. W., Farr, S. A., & Alderton, D. L. (1988). The ability to reason about movement in the visual field. Intelligence, 12, 77-100.

Kelley, T. L. (1928). Crossroads in the mind of man. Stanford, CA: Stanford University Press.

Kosslyn, S. M. (1980). Image and mind. Cambridge, MA: Harvard University Press.

Kosslyn, S. M. (1981). The medium and the message in mental imagery: A theory. Psychological Review, 88, 46-66.

Kosslyn, S. M. (1985). Visual hemispheric specialization: A computational theory (Tech. Rep. #7). Arlington, VA: Office of Naval Research, Personnel and Training Program.

Kosslyn, S. M., Brunn, J. L., Cave, C. R., & Wallach, R. W. (1984). Individual differences in mental imagery ability: A computational analysis. Cognition, 18, 195-243.

Kosslyn, S. M., Holtzman, J. D., Farah, M. J., & Gazzaniga, M. S. (1985). A computational analysis of mental image generation: Evidence from functional dissociations in split-brain patients. Journal of Experimental Psychology: General, 114, 311-341.

Kosslyn, S. M., & Shwartz, S. P. (1977). A simulation of visual imagery. Cognitive Science, 1, 265-295.

Kosslyn, S. M., & Shwartz, S. P. (1978). Visual images as spatial representations in active memory. In E. M. Riseman & A. R. Hanson (Eds.), Computer vision systems, New York: Academic Press.

Kosslyn, S. M., & Shwartz, S. P. (1981). Empirical constraints on theories of mental imagery. In A. D. Baddeley & J. B. Long (Eds.), Attention and Performance, 9. Hillsdale, NJ: Erlbaum.

Lansman, M. (1981). Ability factors and speed of information processing. In M. P. Friedman, J. P. Das, & N. O'Connor (Eds.), Intelligence and earning, 441-457. New York: Plenum Press.

Likert, R., & Quasha, W. H. (1970). Manual for the revised Minnesota Paper Form Board Test. New York: The Psychological Corporation.

Linn, M. C., & Petersen, A. C. (1986). A meta-analysis of gender differences in spatial ability: Implications for mathematics and science achievement. In J. S. Hyde & M. C. Linn (Eds.), The psychology of gender: Advances through meta-analysis, 67-101. Baltimore: Johns Hopkins University Press.

Lohman, D. F. (1979). Spatial ability: A review and reanalysis of the correlational literature (TR No. 8). Palo Alto, CA; Stanford University, Aptitude Research Project, School of Education.

Lohman, D. F., Pellegrino, J. W., Alderton, D. L., & Regian, J. W. (1987). Dimensions and components of individual differences in spatial abilities. In S. H. Irvine & S. E. Newstead (Eds.), NATO ASI series D: Behavioral and social sciences-no. 38: Intelligence and cognition: Contemporary frames of reference, 253-312. Boston: Martinus Nijhoff.

Marr, D. (1982). Vision. San Francisco: Freeman.

McGee, M. G. (1979). Human spatial abilities: Sources of sex differences. New York: Praeger.

Mumaw, R. J., & Pellegrino, J. W. (1984). Individual differences in complex spatial processing. Journal of Educational Psychology, 76, 920-939.

Olson, D. R., & Bialystok, E. (1983). Spatial cognition: The structure and development of mental representations of spatial relations. Hillsdale, NJ: Erlbaum.

Pellegrino, J. W., Alderton, D. L., & Shute, V. J. (1984). Understanding spatial ability. Educational Psychologist, 19, 239-253.

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

Pellegrino, J. W., & Glaser, R. (1980). Components of inductive reasoning. In R. E. Snow, P-A. Federico, & W. E. Montague (Eds.), Aptitude, learning,a nd instruction: Cognitive process analyses of aptitude, 1, 177-218. Hillsdale, NJ: Erlbaum.

Pellegrino, J. W., & Glaser, R. (1982). Analyzing aptitudes for learning: Inductive reasoning. In R. Glaser (Ed.), Advances in instructional psychology, 2, 269-345. Hillsdale, NJ: Erlbaum.

Pellegrino, J. W., & Kail, R. (1982). Process analyses of spatial aptitude. In R. J. Sternberg (Ed.), Advances in the psychology of human intelligence, 1, 311-356. Hillsdale, NJ: Erlbaum.

Pinker, S. (1984). Visual cognition: An introduction. Cognition, 18, 1-63.

Poltrock, S. E., & Brown, P. (1984). Individual differences in visual imagery and spatial ability. Intelligence, 8, 93-138.

Raven, J. C. (1962). Advanced progressive matrices. London: H. K. Lewis.

Sacks, O. (1985). The man who mistook his wife for a hat. New York: Harper and Row.

Shepard, R. N. (1975). Form, formation, and transformation of internal representations. In R. L. Solso (Ed.), Information processing and cognition: The Loyola symposium, 87-122. Hillsdale, NJ: Erlbaum.

Shepard, R. N. (1981). Psychophysical complementarity. In M. Kubovy & J. R. Pomerantz (Eds.), Perceptual Organization, 279-341. Hillsdale, NJ: Erlbaum.

Shepard, R. N. (1984). Ecological constraints on internal representation: Resonant kinematics of perceiving, imagining, thinking, and dreaming. Psychological Review, 91, 417-447.

Smith, I. M. (1964). Spatial ability: Its educational and social significance. London: University of London Press.

Spearman, C. (1904). "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: Erlbaum.

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

Vandenberg, S. G., & Kuse, A. R. (1978). Mental rotations: A group test of three-dimensional spatial visualization. *Perceptual and Motor Skills*, *47*, 599-604.

Vernon, P. E. (1950). *The structure of human abilities*. London: Methuen.

DISTRIBUTION LIST

Distribution:
OCNR-222
OASD (FM&P/MN&PP)
OP-135L, OP-01B2
DTIC (2)

Copy to:
ARI-USAREUR (LIB) (2)
AFHRL (STINFO)
TSRL/Technical Library (FL 2870)
Director of Research, U.S. Naval Academy
Center for Naval Analyses, Acquisition Unit
Director, Office of Naval Research (OCNR-10)
Commanding Officer, Naval Aerospace Medical Research Laboratory, Pensacola, FL
Technical Director, U.S. ARI, Behavioral and Social Sciences, Alexandria, VA (PERI-ZT)
Commanding Officer, U.S. Coast Guard Research and Development Center, Groton, CT
Superintendent, Naval Postgraduate School