THE NUMERICAL FACILITY PROJECT

Joseph Tal

JOHNSON O'CONNOR RESEARCH FOUNDATION
HUMAN ENGINEERING LABORATORY

Technical Report 1987-1
October, 1987

The Numerical Facility Project

Joseph Tal

ABSTRACT

An experimental test battery designed to measure numerical facility was administered to 1,451 Johnson O'Connor Research Foundation (JOCRF) examinees. The battery consisted of four worksamples: (a) Arithmetic; (b) Counting Backwards--alternate and sequential subtraction of two quantities from a given number; (c) Number Reasoning--arrangement of numbers so that they satisfy arithmetic equations; and (d) Rule Learning--induction and application of rules derived from a symbolic system consisting of letters. In addition, the Number Series worksample, involving induction of numerical relationships from ordered series of numbers, was administered to these same examinees as part of the standard JOCRF battery and included with the four experimental worksamples in the majority of the data analyses.

All five worksamples displayed moderately high internal reliabilities (.84-.89). Factor and item analyses revealed that each worksample measured primarily one dimension. A factor analysis of the experimental battery yielded one factor. Arithmetic and Counting Backwards displayed the strongest loadings on the Number factor, while Rule Learning and Number Series displayed lower, though still strong, loadings on this factor. Number Reasoning's loading on the Number factor was between that of the two pairs mentioned above.

Discriminant validity was examined by analyzing the experimental worksamples together with the other worksamples of the standard JOCRF battery. In general, the numerical facility tests displayed relatively low average correlations (.14-.30) with the cognitive worksamples of the JOCRF battery; their average correlations with the noncognitive worksamples were even lower (.06-.10). Arithmetic and Counting Backwards displayed lower correlations with the worksamples of the JOCRF battery than did Rule Learning and Number Series. As before, Number Reasoning displayed an average correlation with the other worksamples that was between that of the two pairs. Multiple regression revealed that between 59 and 76% of the variance in the experimental worksamples remained unexplained after all 25 JOCRF worksamples were partialled out. This indicates that the worksamples of the numerical facility battery provide information that cannot be obtained form the standard JOCRF battery alone.

Predictive validity was investigated by dividing examinees into four college-major groups (quantitative, business, social science, and humanities) and obtaining an average score for each group on each of the numerical facility tests. All five experimental worksamples discriminated to some degree between the groups. However, differences among the four college-major categories were largest for Rule Learning and Number Series and smallest for Arithmetic and Counting Backwards, a reversal of the results obtained for discriminant validity. Differences among the groups on Number Reasoning were between those of the two pairs.

It was concluded that based on internal structure, discriminant validity, and predictive validity, all five numerical facility worksamples could be considered as suitable measures of numerical facility.

CONTENTS

LIST OF TABLES

INTRODUCTION

Thurstone (1938) wrote that "the insistence of the numerical factor makes it almost certain that it represents a unique ability" (p. 83). He also felt that there was some evidence that this ability was genetic in nature. French (1951) wrote that "the number factor is the clearest of them all" (p. 115). The research summarized in this paper represents an effort to identify and to measure the construct of numerical facility. This was done using an experimental test battery and other tests from the aptitude battery of the Johnson O'Connor Research Foundation (JOCRF).

This paper is divided into four parts. The first section introduces the concept of numerical facility and reviews some of the relevant literature in this area. Much of the literature presented in this section was presented in greater detail in Technical Report 1985-3. The second section describes the tests that made up the experimental numerical facility battery and their administration. The third section reviews results of the administration of the experimental battery and its relationship to the aptitudes measured by the rest of the JOCRF battery. A summary of these results and conclusions based on this summary are presented in the fourth and final section.

The Construct of Numerical Facility

By definition, numerical facility involves an ability to work with numbers. However, there are many tasks that employ numbers and that are very different from one another. The most commonly encountered task that utilizes numbers is simple arithmetic. For many individuals, simple arithmetic requires little or no reasoning; it is simply an exercise in the recall of already existing knowledge structures (e.g., the multiplication table). Some tasks utilizing numbers involve primarily the manipulation of short-term memory (e.g., Digit Span on the Wechsler Adult Intelligence Scale; Wechsler, 1958). Still other tasks employing numbers demand complex reasoning of one form or another (e.g., number series tests).

Many studies have attempted to distinguish between various types of numerical tasks. Typically in these studies, several tests employing numbers were administered in conjunction with other aptitude tests. The correlation patterns among all these tests were then explored using data reduction techniques such as factor analysis. The following section presents the results of some of these studies.

Factor Analytic Studies

Thurstone's (1938) now classic primary mental abilities study was the first study to formally identify a numerical factor.

Thurstone administered 57 tests to students at the University of Chicago and found these tests to be best described by 13 factors. One of these he named the numerical factor (N). Ten of the 57 tests administered by Thurstone involved numbers. Nine of these loaded on (correlated with) N .35 or higher. However, while almost all the numerical tests displayed relatively high loadings on N, it was those tests involving simple arithmetic that loaded highest (.61-.81). Tests not restricted to simple arithmetic (e.g., number series) tended to display their highest loadings on other factors.

The results of factor analysis depend importantly on choices, often arbitrary ones, made by the researcher. This is especially true of the number and type of factors that are extracted from a particular correlation matrix. It is therefore important to note that Thurstone's solution was consistent with several other factor analytic solutions of the same correlation matrix (Kaiser, 1960). "More particularly," Kaiser wrote, "correlations between number factors defined by the different rotational solutions generally are higher than other correlations [.90-.97]" (p. 155). In other words, Thurstone's N was especially stable across the various factor analytic solutions.

Other factor analytic studies have displayed results that are consistent with those obtained by Thurstone. In 1942 and 1943, the Army Air Forces Aviation Psychology Program administered 15 batteries, most of which included numerical tests (Guilford & Lacey, 1947). N emerged in factor analyses of 10 of these batteries. In general, N appeared when a test involving simple arithmetic was present and did not appear when such a test was absent. Furthermore, tests of simple arithmetic tended to load exclusively on N. On the other hand, tests involving numerical reasoning (e.g., "word problems," number series) usually loaded about equally on N and a reasoning factor. Such tests also displayed moderate loadings on a verbal factor that is typically defined by tests of reading and vocabulary. Most other factor analytic studies exploring this topic obtained similar results (e.g., Chein, 1939; Comrey, 1949).

The research presented to this point indicates that there is indeed a unique ability associated with the manipulation of numbers. This ability correlates highest with tasks of simple arithmetic. Tasks involving arithmetic reasoning correlate highly with this ability but correlate about equally with reasoning tasks that do not employ numbers. The following section addresses the possibility of a reasoning component in N.

Reasoning and Numerical Facility

While tasks involving simple arithmetic are the primary defining tests of N, numerical tasks requiring reasoning also correlate with this factor. There are at least two possible explanations for the moderate correlations found between numerical reasoning tasks and N. The first focuses on the numerical component of these tasks. It could be argued that since both simple and complex numerical tasks require arithmetic computation, the variance in N that these two tasks share is accounted for by this common component. A second possibility focuses on the reasoning component. According to this argument, the reasoning that accounts for good performance on numerical reasoning tasks is similar to that which is used in simple arithmetic computation.

Of the two hypotheses offered above, the first is the more straightforward. However, it is difficult to believe that simple computation, which is typically a small part of most numerical reasoning tasks (number series excepted), accounts for enough variance in reasoning tasks to completely explain the correlation between them and arithmetic tasks. Individuals performing arithmetic tasks spend all, or almost all, of their time on simple computation. Consequently, variance on this task is due to individual differences in the ability to perform simple arithmetic problems quickly. Alternately, in numerical reasoning tasks the bulk of the time is spent on the reasoning component rather than on computation. Thus, it would be expected that almost all of the variance in such tasks is a result of individual differences in the ability to reason numerically. This rationale suggests that the correlation between numerical reasoning and arithmetic tasks is explained, at least in part, by a reasoning component that the two tasks share.

A study conducted by Forsyth and Ansly (1982) supports the hypothesis that simple arithmetic and arithmetic reasoning share a reasoning component. Forsyth and Ansly were interested in the degree to which computational skills were associated with differences in the performance of numerical reasoning problems. They administered these problems to 567 high school students. Students in one condition completed all the necessary computations by hand; students in the second condition used calculators. No significant differences were found between the two groups. This was taken to indicate the minor role computational skills play in the solution of mathematical problems.

Werdelin and Stjernberg (1971) demonstrated that numerical reasoning is associated with nonnumerical reasoning more than it is associated with tasks involving numbers. They gave subjects arithmetic problems, number series, and nonnumerical logical reasoning. They found that the more difficult a problem was, the higher it tended to load on a general reasoning factor. In other words, the more reasoning required by a numerical problem, the less it loaded on N. In the Kit of Factor-Referenced Cognitive Tests (Ekstrom et al., 1976), the General Reasoning factor is defined exclusively by tests involving the manipulation of numbers. However, none of these tests are purely computational.

No clear conclusion can be reached based on the available evidence. While simple arithmetic is the primary defining test of N, the possibility that reasoning explains some of the variance in N has not been ruled out.

The tasks described to this point all involve numbers. The following section explores whether numerical facility can be measured by tasks that do not employ numbers.

Is Numerical Facility Restricted to Numbers?

By definition, tasks loading highly on N differ from those not loading highly on N. One readily observable difference is that "N tasks" typically employ numbers while non-N tasks generally do not employ numbers. This would suggest that at least part of N's uniqueness is a result of its association with a number system. Nevertheless, a question remains as to whether this ability is specific to numbers, or whether it is helpful in the manipulation of symbolic systems in general.

Coombs (1941) hypothesized that arithmetic ability involves the capacity to manipulate symbols using specified rules. In order to test his hypothesis, Coombs constructed a task requiring the manipulation of letters using rules that were consistent with his conceptualization of numerical ability. He administered this task along with tests measuring seven of Thurstone's (1938) Primary Mental Abilities. In his study only simple arithmetic problems loaded highly on N. At the same time, the correlations between tests of letter manipulation and those involving arithmetic operations ranged between .27 and .52. Of 30 such correlations, 24 were .35 or higher.

Coombs (1941) concluded that "the number factor is most clearly identified by very simple number tests," and that his results "are in agreement with the hypothesis that number ability is characterized by a facility in manipulating a symbolic system according to a specified set of rules" (pp. 188-89). Vernon (1961) disputed Coombs's conclusions. He felt that the relatively small correlations between N and tasks not involving numbers indicated that number facility is number specific.

The disagreement between Vernon and Coombs may be a result of a common problem that arises in interpreting studies such as the one noted above: the distinction between aptitude and achievement. It is expected that the correlation between the two is high but not perfect. Arithmetic, which is highly practiced, is primarily a measure of achievement. Coombs's (1941) letter manipulation task is, by virtue of its novelty, primarily a measure of aptitude. The observed correlation between both tasks is the product of their respective correlations with an identical latent trait (e.g., numerical facility). Even if letter manipulation and arithmetic correlated .7 with a common latent trait, their observed correlation would be only moderate (.49). Consequently, moderate correlations between two tasks are not necessarily an indication that the two do not primarily measure the same latent trait.

While a clear conclusion cannot be reached from Coombs's (1941) study, his ideas are important in that they provide a possible framework for separating ability from achievement in the area of numerical facility.

An interesting result that emerged from Coombs's (1941) study was letter manipulation's higher correlation with arithmetic following practice on the former. This suggested that part of the uniqueness of arithmetic tasks may lie in their highly practiced nature. The following section explores this possibility.

Automatization

The initial learning of arithmetic involves reasoning (Werdelin & Stjernberg, 1969). After a time, numerical operations become highly practiced to the point that for most educated adults they are "automatized." Thus, individual differences in arithmetic (and therefore in numerical facility) may be due in part to individual differences in the ability to automatize. In other words, it is possible that automatization is one component of the latent trait of numerical facility.

As noted above, Coombs's (1941) results suggest that arithmetic correlates higher with a practiced response than an unpracticed response. Keats (1965) administered arithmetic, arithmetic reasoning, verbal, and perceptual tests to college students. He found a possible "automatic process" factor on which multiplication loaded .88. Because of the high loading, it is suspected that his automatic process factor is a variant of N.

In a factorial investigation of perceptual speed, Bechtold (1947) administered perceptual, verbal, and numerical tests to college students. In his oblique solution, the first order factor of N correlated .74 with the second order factor of Perceptual Speed. In another study of perceptual speed, Werdelin and Stjernberg (1969) found that practice increased the correlation of tests of perception with N. In other words, individual differences in the performance of arithmetic tasks may be associated with differences in speeded performance in general.

Differences in performance on speeded tasks have been attributed, in part, to age differences. In addition, differential performance on numerical facility tasks has been associated, in part, with sex differences. The following section examines some of these group differences.

Sex Differences

There is general agreement that there is a sex difference favoring males in mathematical reasoning (Maccoby & Jacklin, 1974). However, there is an ongoing debate concerning the magnitude and the reasons for this difference. After reviewing the relevant literature, Maccoby and Jacklin concluded that the male/female difference in numerical reasoning emerges at about the age of 13 and increases during the high school years. This difference tends to emerge earlier in gifted populations than in less gifted populations (Halpern, 1986). When the number of mathematics courses has been statistically controlled, the male/female difference has generally been reduced (Meece, Eccles, Parsons, Kaczala, Goff, & Futterman, 1982). In other words, the male/female difference in numerical reasoning among individuals that have taken an equal number of mathematics courses is smaller than it is in the general population. Still, males retain a slight superiority even after the number of mathematics courses has been taken into account.

Differences between the sexes favoring males have been found on the quantitative section of the Scholastic Aptitude Test (SAT; an average of 50 points; Halpern, 1986) and the numerical ability component of the Differential Aptitude Test (DAT; an average of .25 of a standard deviation; Bennett, Seashore, & Wesman, 1968). Becker (1983) conducted an item analysis of the SAT and concluded that the male/female difference was restricted to algebraic items but was not evident on arithmetic items. Thus, it appears that the males and females differ in their ability to reason mathematically but not in their ability to compute. It should be noted that females have been found to perform better than males on speeded clerical tasks (e.g., see Bennett et al., 1968). Female superiority in clerical speed may give them an advantage over males in the performance of speeded simple arithmetic tasks.

Age Differences

The results of studies investigating the relationship between age and numerical facility have been mixed (Salthouse, 1982). Salthouse reported relatively little decline up to the age of 50 in the arithmetic subtest of the Wechsler Adult Intelligence Scale (WAIS; Wechsler, 1958). The WAIS's arithmetic subtest is made up of 14 simple word problems. Other investigators have reported similar results (e.g., Owens, 1966; Thurstone & Thurstone, 1949). Bromley (1974, p. 188) has suggested that tests emphasizing "mechanical computation" tend to produce stable age functions, while those requiring "thoughtful reasoning" display sharper declines with age. This hypothesis is not consistent with the relatively small decline reported on the WAIS's arithmetic subtest, which is a test of reasoning. "Nevertheless," Salthouse (1982) concludes, "the mechanical-thoughtful distinction may be useful in characterizing much of the remaining data on age relationships in numerical abilities" (p. 60).

When discussing intellectual functioning and aging, it is useful to distinguish between fluid and crystallized abilities. Crystallized abilities (e.g., vocabulary) are more dependent on sociocultural influences, while fluid abilities (e.g., visuospatial aptitudes) are more dependent on genetic endowment and neurophysiological state. The curves for the decline of these abilities with age differ. Crystallized abilities display relatively little decline with age, while fluid abilities show a comparatively sharp decline with age.

Several intellectual functions may be associated with performance on numerical tests. It is useful to examine each of these with regard to the fluid/crystallized distinction:

(a) Reasoning--Schaie (1980) suggested that numerical reasoning is a fluid ability. As such it would be expected to display significant declines with age. This is consistent with Bromley's (1974) conclusion that thoughtful reasoning declines with age. Tasks requiring inductive reasoning (e.g., number series) are also considered measures of fluid abilities (Willerman, 1979).

(b) Speed--The decline with age in response speed is well documented (Welford, 1977). Many numerical tasks are speeded, and in some numerical tasks (e.g., arithmetic) the proportion of time spent on the motoric activity can be relatively large. Willerman (1979) pointed out that "the distinction between psychomotor tasks and other tasks is not easy to draw, since all psychological tasks require some central processing mechanisms." He added, however, that "the apparent simplicity or complexity of the task may have little to do with the actual underlying complexities of the central or peripheral mechanisms involved" (p. 418).

In any event, it appears that both motoric speed and information processing display significant declines with age (Willerman, 1979). Birren (1974) pointed out that reduced information processing speed means that information in short-term memory cannot be rehearsed as well and, as a result, may decay before it can be utilized. Consequently, less information has an opportunity to enter long-term memory. Of more importance for numerical tasks, problems may have to be attempted several times before the solution is found.

(c) Short-term memory--Willerman (1979) wrote: "One may conclude that there are declines in the efficiency of the short-term memory store with increasing age" (p. 416). As noted above, declines in short-term memory can be expected to most adversely affect those tasks which have memory demands and in which speed is important. Counting Backwards, one such task, is described later in this section. When speed is not important, one can compensate for short-term memory losses by increased rehearsal and, as a result, increased utilization of long-term memory.

Summary and Conclusions

The tasks that are the purest measures of numerical facility are those involving simple arithmetic operations. It appears that as numerical tasks increasingly involve reasoning, they load less on N and more on reasoning factors. Consequently, any test battery intending to study numerical facility must include a test of simple arithmetic. Arithmetic was the marker test (i.e., the test used as a reference for the other tests) for the numerical facility battery administered in this study.

While numerical facility does not primarily involve reasoning, there is some indication that it is not completely without a reasoning component. As a result, an arithmetic test that requires a minimal amount of reasoning is also appropriate as part of a numerical facility battery. The Number Reasoning test was intended to fulfill this function in the current battery.

It has also been hypothesized that arithmetic ability is one example of a more general ability to manipulate symbolic systems according to specified rules. If this were true, it would be useful to administer a test involving a novel task utilizing the manipulation of a nonnumeric system in addition to simple arithmetic. The advantage of using such a task is that all the examinees performing it will have had an equal degree of prior exposure to the task (namely, no exposure). Consequently, the task would be primarily a measure of aptitude rather than of achievement. The Rule Learning test constructed for the current battery was intended to fulfill this function.

A fourth test, Counting Backwards, was also administered as part of the experimental numerical facility battery. This test differs from the three other tests in that it appears to possess a relatively large memory component. Both Counting Backwards and Number Reasoning have been previously used by the Foundation.

The complete Johnson O'Connor Research Foundation battery includes a Number Series test. It was both convenient and instructive to regard Number Series as part of the numerical facility battery for many of the statistical analyses conducted in this study.

A more detailed description of the worksamples making up the experimental numerical facility battery can be found in the following section.

METHOD

Examinees

Examinees taking the JOCRF battery typically do so to obtain career guidance. Most of the 1,451 examinees who took the complete experimental numerical facility battery were college-educated or college-bound. The average age of those taking the numerical facility battery was 29 (SD = 10.3), and the median was 24.5. The JOCRF examinee is typically white and middle to upper-middle class.

The data for this study were collected from 12 testing centers across the United States over a period of five months. Approximately 50% of the data were obtained from the eastern United States (Boston, New York, Philadelphia, Washington, D.C.), 30% from the south (Atlanta, Dallas, Houston, Tulsa), and 20% from the western United States (Denver, Los Angeles, San Diego, Seattle).

Measures

The Complete JOCRF Aptitude Battery

The numerical facility battery was given as part of a larger battery administered by the JOCRF. Table 1 presents a brief description of the aptitudes measured by the standard worksamples in this battery. A more detailed description of the experimental numerical facility battery follows.

The Experimental Numerical Facility Battery

Arithmetic (Worksample 721 A*). The Arithmetic worksample was made up of three parts. The first part contained 30 problems, all of which involved adding up four single-digit numbers. Examinees were given one minute to work on these problems.

The second part of the worksample contained 26 subtraction problems. Twelve of these entailed subtracting three-digit numbers from three-digit numbers, while the remaining problems required subtracting three-digit numbers from four-digit numbers. There were no negative solutions to the subtraction problems. Examinees were given one minute and 40 seconds to complete these problems.

The third part of the worksample contained 29 multiplication problems. Each of the problems entailed multiplying a two-digit number by a single-digit number. One minute and 20 seconds were given for the completion of this section.

Counting Backwards (Worksample 420 AB). In the Counting Backwards worksample the examinee was asked to subtract seven from a given number and to then subtract six from the result. Subtracting seven and six alternately was continued in sequence until the examinee had reached a predetermined point. Each examinee was given two essentially identical trials.

Number Reasoning (Worksample 436 H*). In the Number Reasoning worksample the examinee was presented with a playing board with six spaces marked across the top and two arithmetic equations in the center. Hexagonal chips with numbers printed on them were placed across the top. When the last chip was placed,

Table 1

The Aptitudes Measured by the Standard JOCRF Test Battery

Name	Description
Graphoria	Speed and accuracy in noticing if pairs of numbers are the same or different.
Ideaphoria	Verbal fluency, the rate of flow of ideas.
Foresight	Ability to keep one's mind on a long-range goal.
Inductive Reasoning	Quickness in seeing a common element among separate facts, ideas, or observations.
Analytical Reasoning	Quickness in arranging ideas into logical sequence.
Wiggly Block	Structural visualization, an aptitude for visualizing three-dimensional forms. Measured by the ability to reconstruct a three-dimensional block.
Paper Folding	Structural visualization. Measured by the ability to rotate two-dimensional surfaces through three-dimensional space.
Personality	Tendency to react from a general, objective viewpoint versus reacting from a personal, subjective viewpoint. Describes how well suited a person is for work that is highly oriented toward person contact (objective) or toward individual performance (subjective).
Tonal Memory	Ability to remember sequences of tones.
Pitch Discrimination	Ability to differentiate fine differences in pitch.
Rhythm Memory	Ability to remember complex rhythmic patterns.
Memory for Design	Memory for straight-line patterns.
Silograms	Associative memory for English words paired with nonsense syllables.
Number Memory	Ability to remember several six-digit numbers simultaneously.

(ctd.)

Table 1 (continued)

Name	Description
Number Series	Ability to detect complex numerical patterns.
Observation	Quickness in recalling fine visual details.
Finger Dexterity	Speed and accuracy in manipulating small objects with one's fingers.
Tweezer Dexterity	Speed and accuracy in handling small objects with tweezers.
Reading Efficiency	Ability to read quickly and accurately.
Vocabulary	Knowledge of English words.

the examinee was required to quickly and accurately arrange the numbers in the spaces in the two equations in a way that satisfied the specified arithmetic relationships. Examinees were given a maximum of 30 seconds to complete each problem. It was expected that most examinees would have little difficulty completing each problem within the allotted time.

Rule Learning (Worksample 720 A*). The Rule Learning worksample was divided into three categories of problems: Induction, Practice, and Application. Examinees learned and practiced the task in the first two sections. Enough time was given for these sections so that most examinees would not feel time-pressured. Induction and Practice were designed as a preliminary to the more important Application section. It was expected that the Application section would yield the most meaningful scores of the three sections. Unlike the first two sections, Application was designed so that most examinees would not complete all the problems within the designated time limit.

In the Induction portion of the worksample, a set of examples was given for each of three rules in a symbolic system consisting of letters. The examinee was asked to discover each rule and to use that rule in a set of ten problems. In all, the Induction section consisted of 30 problems.

The Practice section followed each rule's Induction section. In the Practice section the rule was explained, and the examinee was given a set of 10 problems in which all the rules used to that point were to be applied. Thus, the Practice section consisted of 30 problems.

After all the rules had been learned and practiced, the examinee was taught the use of parentheses in the test (same principle as in conventional arithmetic expressions: operations within parentheses are performed first). The Application category that followed was made up of four problem sets in which parentheses were used.

In order to complete the problems in these sets correctly, the examinee was required to use the rules learned in the first two sections. The first and third problem sets contained 20 problems each. The second and fourth problem sets were made up of 13 problems each. In all, the Application section contained 66 problems.

To summarize, the Rule Learning test was made up of three categories: Induction, Application, and Practice. Each of the first two categories was made up of three sections. A Practice section followed each Induction section. The final category was made up of four sections. In all, examinees were given 10 sections containing a total of 126 problems.

The results of the administration of the numerical facility battery are reported in the following sections.

RESULTS

The Experimental Worksamples Considered Individually

Arithmetic

Scores. As noted earlier, the Arithmetic worksample was divided into three subtests: (a) addition, (b) subtraction, and (c) multiplication. Problems within each of these subtests were scored as to whether they were answered correctly. Examinees then received scores for each of the subtests based on the total number of problems each completed correctly. The score for the complete worksample was computed by adding up each examinee's subtest scores. Examinees were also given scores for each subtest and for the complete worksample based on the total number of problems each completed (including both correct and incorrect problems). Table 2 presents the means and standard deviations for each of the Arithmetic subtests.

Internal structure. The internal structure of highly speeded tests cannot be analyzed in the usual manner. Most of the items that are attempted on these tests are completed correctly. When this occurs, there is not sufficient variance at the item level to make an item-level reliability estimate meaningful. As can be observed from Table 2, the differences between the total number of correct problems and the total number attempted were very small (.63, 1.70, and 1.84 for addition, subtraction, and multiplication, respectively). Indeed, over 92% of all the problems that were attempted were completed correctly.

An alternative to estimating reliability at the item level involves using subtest scores. Typically, there is sufficient variance at the subtest level so that each subtest can be treated as an item for the purpose of a reliability analysis. The complete test is then viewed as being made up of as many items as there are subtests. The following formula computes an estimate of alpha reliability based on interitem correlations:

$$\underline{r}_{xx} = K^2\underline{r} \ / \ (K + K^2\underline{r} - K\underline{r}),$$

where $\underline{r}_{xx}$ is alpha, K is the number of items on the test (in the case of arithmetic three: addition, subtraction, multiplication), and $\underline{r}$ is the average correlation among the items.

Table 3 presents the correlations between the subtests for both the total attempted and total correct variables. When the variable being used was the total number of problems completed correctly, the average correlation among the subtests was .63, resulting in an alpha of .84. When the variable was the total number of problems attempted, the average correlation was .66, yielding an alpha of .85. As can be seen, the reliabilities derived from the two dependent variables are essentially the

Table 2

Descriptive Statistics for Arithmetic Worksample

Subtest	No. problems	Mean correct	SD	Mean attempted	SD
Addition	30	16.86	5.55	17.49	5.57
Subtraction	26	15.97	3.95	17.67	3.75
Multiplication	29	15.51	5.76	17.35	5.69
Total	85	48.34	13.26	52.51	13.19

Table 3

Correlations Between Arithmetic Subtests

Subtest	ADD[1]	SUB	MULT	ARIT[2]	ADDTOT	SUBTOT	MULTTOT	ARITTOT
ADD								
SUB	64							
MULT	63	61						
ARIT	88	83	88					
ADDTOT	98	63	64	88				
SUBTOT	67	93	65	84	67			
MULTTOT	63	60	96	86	64	66		
ARITTOT	88	79	87	98	89	85	89	

Note. *N* = 1,451. Decimal points omitted. All correlations significant at .001 level.

[1] Subtests with suffix "TOT" refer to total attempted, while those without suffix refer to total correct.

[2] "ARIT" refers to sum of ADD, SUB, and MULT.

same. This result was expected because of the high correlations between the total attempted and the total correct variables (see Table 3). Based on the reliability analysis alone, it could not be concluded that one of the variables was superior to the other. In the interest of simplicity, subsequent analyses are reported only for the total correct variable. Analyses using this variable yielded results that were essentially the same as those using the total attempted variable.

A high alpha coefficient usually indicates that a test is unidimensional (Cronbach, 1951). Nevertheless, there can be instances in which a test that is not unidimensional possesses a high alpha coefficient (e.g., when a large group of items within the test measures one dimension and a smaller group measures another dimension). To verify the unidimensionality of the Arithmetic worksample, it would have been necessary to conduct a factor analysis (or some other data reduction procedure) of the test's items. This could not be done because of the small variance at the item level.

While the more appropriate test of unidimensionality could not be conducted on this worksample, it is probably safe to assume that arithmetic is unidimensional. First, based on the literature that was reviewed in the introduction, it was expected that any test made up of simple arithmetic problems would measure only one dimension. Second, if Arithmetic were multidimensional, it would be expected that items measuring different dimensions would reside in different subtests. The high correlations between the subtests indicate that this probably is not the case. Analyses of the complete battery and of the bargraph data were expected to shed additional light on this issue.

Age effects. Table 4 presents a breakdown by age of examinees' scores on the complete worksample. Surprisingly, there was a tendency for older examinees to do better than younger ones. The correlation between age and total number correct was .25 (p < .001). This indicates that only about 6% of the variance in number correct was explained by age. However, this figure is artificially low because of the overrepresentation of younger examinees among those who took this worksample. When a correction for the restriction of range was employed (Cohen & Cohen, 1983) the correlation between number correct and age increases to .31.

Sex effects. The performance of males and females differed significantly on two of the three subtests. Males did better than females on the addition subtest, averaging 17.23 correctly completed problems to the females' average of 16.58, a difference of .12 of a standard deviation (p < .05; n = 1,264). However, this pattern was reversed for the subtraction subtest, where females averaged 16.44 correctly completed problems to the males' average of 15.62, a difference of .21 of a standard deviation (p < .001; n = 1,264). Because (a) the significant differences that were observed were in opposite directions, (b) the magnitude of

Table 4

Number of Arithmetic Problems Answered Correctly by Age

Age interval	No. correct	N
13 - 15	48.20	24
16 - 18	46.49	222
19 - 21	46.48	199
22 - 24	43.19	170
25 - 27	45.09	99
28 - 30	50.37	108
31 - 33	48.62	90
34 - 36	54.18	77
37 - 39	52.25	75
40 - 42	51.49	61
43 - 45	56.70	49
46 - 48	54.09	38
49 - 51	56.23	25
52 - 54	51.34	10
55 - 57	48.85	11
58 - 61[1]	59.36	8
Overall	48.50	1,266

[1] No 60-year-olds took the worksample.

these differences did not exceed .21 standard deviation, and (c) the sexes did not significantly differ in their complete worksample score, it cannot be concluded that, in general, males and females differ in their performance on this worksample.

Summary. (a) The Arithmetic worksample is moderately reliable, possessing an alpha of .84.

(b) While an item factor analysis could not be done, it is probably safe to say that the worksample is unidimensional.

(c) Surprisingly, the correlation between age and performance was positive, .25 (.31 when a correction for restricted age sampling was employed).

(d) No reliable sex effects were obtained.

Counting Backwards

Scores. As was noted earlier, the Counting Backwards worksample consists of two essentially identical parts. An examinee's speed score for each part was the time it took him or her to complete that part. The number of correct subtractions for each part was the accuracy score for that part. A speed score for the complete worksample was computed by adding together the examinee's speeds on Parts 1 and 2. Similarly, an accuracy score for the worksample was computed by adding the examinee's accuracies for Parts 1 and 2. Table 5 presents means and standard deviations for the dependent measures of the Counting Backwards worksample.

When, as in counting backwards, the dependent variable is the time to the completion of a task, a reciprocal transformation is usually the most appropriate. In addition to reducing the effect of extreme values, a reciprocal transformation on such data can be justified theoretically. Cohen and Cohen (1983) wrote:

> "Reciprocals arise quite naturally in the consideration of rate data. Imagine a . . . task presented in time limit form--all subjects are given a constant amount of time (*T*), during which they complete a varying number of units (*u*). One might express the scores in the form of rates as *u/T* but, because *T* is a constant, we may ignore *T* and simply use *u* as the score. Now consider the same task, but presented in work limit form--subjects are given a constant number of units to complete (*U*), and are scored as to the varying amounts of time (*t*) they take. Now if we express their performance as rates, it is *U/t* and, if we ignore the constant *U*, we are left with 1/*t*, not *t*. If rate is linearly related to some other variable *v*, then for the time limit task, *v* will be linearly related to *u*, but for the work limit task, *v* will be linearly related not to *t*, but to 1/*t*." (p. 263-64)

The reciprocal of the speed variable was computed for Parts 1 and 2 of the worksample, as was the sum of the two reciprocals. Most of the analyses in this section utilized transformed scores. So that the relationships between transformed time,

Table 5

Descriptive Statistics for Counting Backwards Items

Score	Untransformed mean	SD	Transformed[1] mean	SD
Part 1				
Speed[2]	1.16	.55	8.93	.49
Accuracy[3]	12.72	2.19		
Part 2				
Speed	1.16	.53	8.96	.47
Accuracy	13.15	2.19		
Complete worksample				
Speed	2.32	1.00	17.89	.90
Accuracy	25.87	3.70		

[1] Transformed speed = 10 - (1/speed).

[2] Time to completion in minutes.

[3] Number of subtractions completed correctly.

untransformed time, and other variables were in the same direction, each transformed score was subtracted from a constant, namely, 10.

Internal structure. Analysis of the internal structure of the Counting Backwards worksample was straightforward. Because there was no reasonable way to divide each of its two parts into smaller units, the worksample was viewed as being composed of two items. Table 6 presents the correlations between the various scores and subscores. As can be seen, the correlation between the speed variables was much higher than that between the accuracy variables (.71 untransformed speed, .77 transformed speed versus .43 for accuracy). This indicated that, in this task, speed was a more reliable score than accuracy. In addition, the stronger relationship between the two halves that was obtained using transformed speed as compared with untransformed speed suggested that the transformation was justified. The alpha reliability for transformed speed was .87 (based on a two-item test with a .77 correlation between the items).

A composite score based on speed and accuracy was computed by adding together the standardized scores of the two dependent measures within each part of the worksample (standardized transformed speed scores were used for this composite). The correlation between the two speed/accuracy variables was .48, much lower than the .77 obtained between the transformed speed scores of Parts 1 and 2. As a result, it was concluded that transformed speed alone was, psychometrically, a more appropriate score than a measure using both speed and accuracy.

Age effects. The correlation between age and total transformed speed was -.06 ($p < .05$), accounting for less than one percent of the variance of the test. As noted in the previous section, this is an underestimate because of the overrepresentation of younger examinees in the sample. When the correction for restriction of range was applied, the correlation was -.08. Thus, it appears that the relationship between age and speed on the Counting Backwards worksample is trivial.

Sex effects. Males took an average of 1.04 minutes to complete each part of the worksample, while females took an average of 1.24 and 1.27 minutes to complete Parts 1 and 2, respectively. The difference in total transformed time between the sexes was significant at $p < .001$ (an effect size of .36 of a standard deviation). In addition, men were somewhat more consistent than women in the time it took them to complete each of the two parts; the correlation between speed in the two parts was .72 for women and .77 for men ($p < .001$).

Summary. (a) The most appropriate score for the Counting Backwards worksample was transformed speed.

(b) Alpha reliability of transformed speed was .87.

(c) The relationship between age and transformed speed was

Table 6

Correlations Within Counting Backwards Worksample

A. Using untransformed speed scores

Score	Speed1[1]	Accuracy1	Speed2	Accuracy2	Total Speed	Total Accuracy
Speed1						
Accuracy1	-19					
Speed2	71	-27				
Accuracy2	-20	43	-28			
Total Speed	93	-25	92	-26		
Total Accuracy	-23	85	-33	85	-30	

B. Using transformed speed scores

Score	Speed1	Accuracy1	Speed2	Accuracy2	Total Speed	Total Accuracy
Speed1						
Accuracy1	-16					
Speed2	77	-27				
Accuracy2	-17	43	-26			
Total Speed	94	-22	94	-23		
Total Accuracy	-20	85	-31	85	-27	

Note. *N* = 1,451. Decimal points omitted. All correlations significant at .001 level.

[1] The numbers 1 and 2 at the end of a variable name refer to Parts 1 and 2 of the worksample.

statistically significant but negligible.

(d) Men were about one-third of a standard deviation faster than women for the complete worksample.

Number Reasoning

Scores. The score assigned for each problem in the Number Reasoning worksample was the time it took to complete that problem correctly. The maximum time allowed for any particular problem was .50 minute. The score for the complete worksample was the sum of the scores of the individual problems. The number of problems not completed correctly was not sufficient for them to be discriminated from problems completed correctly in the maximum time, .50 minute. Examinees not completing a problem within the allotted time period were assigned a score of .50 for that problem.

Table 7 presents the means and standard deviations for the problems (items) in the Number Reasoning worksample. An average of 9% of the examinees received scores of .50 on any given problem. As can be seen from Table 7, the average time needed for the completion of a problem was .192. The average standard deviation across all the problems was .116. In other words, .50 is on the average 2.66 standard deviations away from the mean of any particular problem. Thus, roughly 9% of subjects were 2.66 standard deviations away from the mean on any particular problem.

The Number Reasoning and Counting Backwards worksamples were similar in that both required examinees to complete tasks with essentially no time limit. As noted in the section describing the results of the Counting Backwards worksample, when the dependent variable is speed of completion, a reciprocal transformation of the data is often necessary (Cohen & Cohen, 1983). In addition to being theoretically appropriate, this transformation minimized the effects of the extreme scores described in the previous paragraph. It should be noted that the scores for the complete worksample were obtained by summing the transformed scores of the individual items (and not by transforming the sum of the raw scores of the individual items). So that the relationships between transformed time, untransformed time, and other variables were in the same direction, each transformed score was subtracted from a constant, namely, 30.

Internal structure. Part A of Table 8 presents alpha reliability for the complete worksample. As expected, the reliability coefficients obtained using transformed scores were higher than those computed from the untransformed scores. In other words, interitem correlations were higher for transformed as compared with untransformed scores. This indicated that the reciprocal transformation was justified.

Part B of Table 8 displays the corrected item-total correlations for each item in the worksample. An item-total correlation is the correlation between examinees' scores on a

Table 7

Descriptive Statistics for Untransformed Number Reasoning Items

Item	Mean time[1] to completion	SD
1	.230	.149
2	.351	.148
3	.231	.129
4	.149	.095
5	.163	.112
6	.192	.122
7	.214	.135
8	.212	.134
9	.302	.155
10	.171	.103
11	.125	.077
12	.188	.128
13	.117	.076
14	.183	.124
15	.146	.093
16	.145	.094
17	.109	.076
18	.229	.145
Total[2]	192	.051[3]

[1] In fractions of one minute (e.g., .20 minute = 12 seconds).

[2] The transformed total score was 22.16, with a standard deviation of 2.18.

[3] This is the *SD* of examinees' means, not the mean of the *SD*s across all the problems (which is .116).

particular item and the total score for each examinee across all the items. It describes the degree to which an individual item fits the test; e.g., the degree to which the item measures the same latent trait that is measured by the other items on the test. The higher the item-total correlation, the better an item fits the test. Good-fitting items increase a test's reliability, while poorly fitting items tend to decrease reliability or to add very little to reliability; i.e., removal of a good item decreases the reliability of a test, while removal of a bad item increases reliability or leaves it the same.

None of the items of the Number Reasoning worksample had seriously low item-total correlations. However, Problems 1, 2, 8, 9, and 18 displayed relatively low item-total correlations. Reliability was also computed with individual items excluded in turn from the analysis. Other than for item 18, all such analyses resulted in decreased reliability coefficients. Excluding Problem 18 did not bring about any change in the worksample's reliability. In other words, while Problem 18 was not a good problem, including it in the worksample did not detract from the test. It is probable that Problem 18's lack of fit was due to its position on the worksample rather than to its content. If examinees were provided with cues that this was the final item, they may have approached it differently than they did the other items (e.g., by expending less effort). As a result, this problem may have measured a motivational variable (e.g., willingness to "slack off" at the end of a task) in addition to the trait measured by the other problems. Whenever a particular item measures an attribute that is not measured by most of the other items, its contribution to the test's overall reliability is expected to be relatively small or negative.

While Problem 18 was the only problem that contributed nothing to the worksample's reliability, Problems 1, 2, 8, and 9 contributed relatively little. The contribution of the first two problems to the reliability of Number Reasoning was probably limited regardless of their content. Since Number Reasoning was not a task that examinees were expected to have encountered before their testing at the Foundation, the first problems probably measured how well examinees were learning the task, as well as how good they were in manipulating numbers. As noted above, a problem measuring a variable that other problems do not measure contributes little to overall reliability. However, the relatively low item-total correlations of Problems 8 and 9 indicated that they may not have been good items.

The moderately good alpha reliability of the Number Reasoning (.84) suggested that this worksample was unidimensional--in other words, that it primarily measured a single attribute, or latent trait, of the examinee. This was confirmed by factor analysis. Initial principal axis factoring of the item correlation matrix yielded three factors with eigenvalues greater than 1.0. However, the eigenvalue of the first factor was more than four times the size of each of the two subsequent factors (the

Table 8

Reliability Statistics for Number Reasoning Worksample

A. Reliability coefficients for complete worksample.

Alpha reliability[1]

Untransformed items	Transformed items
.74	.83

B. Corrected[2,3] item-total correlations.

Item	Corrected item-total correlation
1	.30
2	.32
3	.45
4	.51
5	.45
6	.40
7	.44
8	.30
9	.31
10	.51
11	.43
12	.44
13	.46
14	.40
15	.46
16	.46
17	.55
18	.28

[1] Nonstandardized alpha.

[2] Based on transformed times.

[3] The time an examinee took to complete a particular item is excluded from the total for that item.

eigenvalues for the first three factors were 4.71, 1.14, and 1.02, respectively). This indicated that a one-factor solution was optimal for this worksample. Table 9 presents the loadings of the individual items in the one-factor solution. It should be noted that, as expected, the lowest loadings were displayed by those items that contributed least to the worksample's reliability.

Age effects. Table 10 presents a breakdown by age of the average time it took examinees to complete the worksample. Older examinees took, on the average, more time to complete the worksample. The correlation between age and average time to completion was .20 ($p < .001$). This indicated that only about 4% of the variance in time was explained by age. As noted in the previous section, this figure is artificially low because of the overrepresentation of younger examinees in this worksample. When a correction for the restriction of range was employed, the correlation between age and performance increased to .25.

Sex effects. Males and females did not significantly differ in their performance on the complete worksample. They differed significantly on only one of the 18 problems in the worksample (males took longer to complete Problem 10; $t < .05$). This result was about what would have been expected by chance.

Reliability and factor analyses done separately for each of the sexes yielded results that were essentially the same as those obtained for the complete sample. In other words, the internal structure of the worksample did not differ between the two sexes.

Summary. (a) The most appropriate score for the Number Reasoning worksample was examinees' transformed time.

(b) Using transformed time, the reliability of this worksample was moderately good, .84.

(c) Problems 8 and 9 displayed low item-total correlations and should probably be replaced. Other low item-total correlations were likely caused by the problems' position in the worksample rather than by their content. Test administrators should be careful not to give examinees cues regarding the length of the worksample.

(d) The correlation between age and performance was .20 (.25 when a correction for restriction of range was employed).

(e) No reliable sex effects were encountered in the worksample.

Rule Learning

Scores. As described earlier, the Rule Learning worksample was divided into 10 subtests. The subtests were classified into three categories: Induction, Practice, and Application. An examinee's score for a particular subtest consisted of the number of items he or she answered correctly within that subtest. Scores for each category were obtained by adding together the scores of the subtests within that category. While scores were

Table 9

Loadings of Individual Items[1] on Primary Factor of Number Reasoning Worksample[2]

Item	Loading
1	.33
2	.35
3	.51
4	.57
5	.51
6	.44
7	.49
8	.33
9	.35
10	.57
11	.48
12	.49
13	.51
14	.44
15	.50
16	.61
17	.61
18	.31

[1] Transformed item scores were used for the factor analysis.

[2] The primary factor displayed an eigenvalue of 4.71, accounting for 26.1% of the variance in the worksample.

Table 10

Mean Untransformed Completion Time per Item by Age Interval for Number Reasoning

Age interval	Mean minutes to completion	N
13-15	.190	24
16-18	.184	222
19-21	.178	199
22-24	.195	170
25-27	.190	99
28-30	.177	108
31-33	.196	90
34-36	.200	77
37-39	.207	75
40-42	.202	61
43-45	.207	49
46-48	.210	38
49-51	.212	25
52-54	.271	10
55-57	.228	11
58-61	.185	8
Total	.192	1,266

obtained for all 10 subtests and for the three categories, it was expected that the scores for the Application subtests would be the most meaningful. Table 11 presents the means and standard deviations for each of the subtests and for each of the categories.

Internal structure. Because it was anticipated that the Application category would be the most meaningful, its internal structure is reported first and in the greatest detail.

The Application category. As was expected, no examinee attempted all of the problems within the Application category. To derive alpha in the usual manner (at the item level), it was necessary to treat problems that were not attempted in the same manner as problems that were completed incorrectly. As noted in the section describing the Arithmetic results, this was expected to lead to an inflated reliability coefficient. An alpha of .95 across all the subtests was obtained when this procedure was used.

Two alternative estimates of alpha were also utilized. For the first, only those items in each Application subtest that were completed by 75% or more of the examinees were used to derive alpha. This yielded an alpha reliability of .89. The second method treated each subtest within Application as an item on a four-item test. Alpha was then computed using the formula described in the section reporting the results of the Arithmetic worksample. Given an average intersubtest correlation of .65, this procedure yielded an alpha of .88. Since the two alternative procedures produced similar alphas, it is reasonable to assume that Application's alpha is in the vicinity of .90.

As noted in the section describing the Arithmetic worksample, a high alpha indicates that a test is probably unidimensional. Item-level analyses were used to investigate this issue further. As noted above, this presented a problem because so many of the items were not completed. Consequently, only items completed by at least 75% of the examinees were examined. Table 12 presents item-total correlations for these problems. In addition, reliability coefficients were computed with individual items excluded in turn from each analysis.

Items B1 and D1, two problems displaying relatively low item-total correlations, also brought about slight reductions in alpha. (Because the number of items included in this analysis was relatively large, each item had a negligible overall effect on the reliability. Thus, reductions in alpha could only be detected at the third decimal place.) Both these problems had the lowest item-total correlations within their respective subtests. It is important to note that each of these problems was also the first in its respective subtest. Similarly, A1 and C1 displayed relatively low item-total correlations (second from the lowest in their respective subtests). Other problems having low item-total correlations were generally found at the end of

Table 11

Descriptive Statistics for Rule Learning Subtests

Subtest	No. problems	Mean correct	SD	% of examinees answering all problems correctly
Induction A	10	8.82	2.57	76.5
Induction B	10	7.99	3.33	63.5
Induction C	10	8.38	2.78	61.2
Total Induction	30	23.77	7.08	32.9
Practice A	10	9.54	1.52	87.3
Practice B	10	9.51	1.43	84.5
Practice C	10	9.28	1.60	75.3
Total Practice	30	28.09	3.68	60.8
Application A	20	11.03	4.69	3.2
Application B	13	5.85	3.16	1.1
Application C	20	9.00	3.50	0.2
Application D	13	4.75	2.48	0.0
Total Application	66	30.36	12.12	0.0

Table 12

Item-Total Correlations for Application Problems[1]

Item	Corrected item-total correlation[2]
A1	.35
A2	.40
A3	.55
A4	.36
A5	.52
A6	.48
A7	.47
A8	.43
A9	.22
A10	.44
B1	.30
B2	.40
B3	.55
B4	.49
B5	.43
B6	.60
B7	.34
C1	.42
C2	.45
C3	.47
C4	.50
C5	.51
C6	.51
C7	.51
C8	.50
C9	.29
D1	.31
D2	.49
D3	.53
D4	.39
D5	.58

[1] An individual item was included only if 75% or more of the examinees completed it. Altogether, 684 examinees, or 47.1%, completed all 31 items. This table is based on these 684 examinees.

[2] The score for an examinee on a particular item is excluded from the total with which that item is correlated.

each subtest (end refers only to problems included in Table 12; i.e., these problems were the last problems completed by many examinees, but they were not actually the last problems within their respective subtests). This indicated that it was probably these problems' positions in the subtest, rather than their content, which caused them to be relatively bad items.

Table 13 presents the results of factor analysis using only those problems completed by 75% of the examinees. Principal axis factoring initially extracted seven factors with eigenvalues greater than one. The first factor accounted for 27.3% of the variance in the test, while the second accounted for 6.9%. None of the remaining factors accounted for more than 5% of the variance in the test. As a result, it was concluded that a one-factor solution was the most appropriate for the Application category. The loadings of the individual items on this first factor are displayed in Part A of Table 13. In general, items that were placed at the beginning and end of each subtest displayed lower loadings than those in the middle of the subtest. This was expected since these items displayed relatively low item-total correlations.

It is important to note that, in general, in the one-factor solution, items did not tend to "hang together" by subtest. In other words, problems from all the subtests could be found in the higher and lower ranges of factor loadings. This indicated that while the subtests were physically separated in the worksample, they measured the primary factor about equally. While problems from all the subtests could be found throughout Part A of Table 14, there did appear to be a slight overrepresentation of items from APL-C and APL-D in the higher range of loadings (and consequently an overrepresentation of APL-A and APL-B in the lower range). This was not surprising in light of the hypothesized role of "automatization" in the Rule Learning worksample (see Introduction).

Based on the results of the reliability and the one-factor factor analysis, it was hypothesized that "position" factors might emerge in a two-factor solution. That is, one factor would contain items primarily from the extremes of each subtest, while the other would contain items primarily from the center. Part B of Table 14 presents the results of the rotated two-factor solution. The hypothesized position factor did not emerge in this solution. However, it can be observed that the first factor contained all the APL-D items and all but one of the APL-B items (the only item from APL-B in the second factor displayed the lowest loading of any item on that factor and loaded almost equally on the first factor). It should be remembered that APL-B and APL-D contain complex items, while the other two subtests contain relatively simple items. However, since (a) the one-factor solution is optimal, and (b) the separation in the two-factor solution between the simple and complex problems is not clear cut (i.e., Factor 1 contains nine items from subtests APL-A

Table 13

Factor Loadings of Application Items Completed by at Least 75% of Examinees[1]

A. One-factor solution

Item	F1
A3	.66
D5	.64
B6	.62
A5	.61
D3	.61
C5	.61
C4	.58
C3	.57
D6	.57
B3	.56
C2	.56
C8	.54
B4	.53
C6	.52
A10	.52
A6	.51
C7	.50
D2	.50
A2	.46
B5	.45
A8	.45
A7	.45
D4	.44
C1	.43
A4	.42
B2	.38
B7	.38
A1	.36
C9	.33
D1	.33
B1	.31
A9	.27

B. Two-factor solution[2]

Item	F1	F2
B6	.69	.19
D5	.65	.25
D2	.63	.08
D6	.61	.20
D3	.58	.28
B3	.57	.22
B5	.53	.11
B2	.47	.06
A7	.44	.19
A1	.42	.09
C6	.41	.33
C7	.41	.29
A6	.39	.32
C8	.38	.38
B4	.38	.36
A8	.35	.29
C1	.31	.29
D4	.31	.31
D1	.29	.18
B7	.29	.24
A3	.22	.74
C4	.19	.65
A10	.12	.64
C3	.20	.62
A5	.28	.59
C5	.30	.56
C2	.25	.55
A4	.14	.47
A2	.21	.45
C9	.09	.38
A9	.05	.34
B1	.22	.23

[1] Principal axis factoring. Item D6 was included in factor analysis although it was completed by only 61.4% of examinees.

[2] Varimax rotation.

Table 14

Correlations Among Subtests of Rule Learning Worksample

A. Correlations among subtests

Subtest	IND-A[1]	IND-B	IND-C	PRA-A	PRA-B	PRA-C	APL-A	APL-B	APL-C	APL-D
IND-A										
IND-B	.18									
IND-C	.16	.24								
PRA-A	.40	.13	.15							
PRA-B	.13	.26	.21	.18						
PRA-C	.16	.19	.38	.19	.33					
APL-A	.17	.17	.28	.17	.24	.42				
APL-B	.16	.19	.27	.16	.19	.35	.60			
APL-C	.19	.22	.30	.21	.25	.42	.71	.66		
APL-D	.14	.20	.25	.16	.19	.31	.52	.75	.63	

B. Correlations among subtest composites[2]

Subtest	IND	PRA	APL
IND			
PRA	.33		
APL	.39	.21	

[1] The first three letters refer to the the type of problem set (IND = Induction, PRA = Practice, APL = Application) and the last letter to the subset within the problem type.

[2] Each composite is calculated by adding up the number of correct problems across all the subtests within a problem type.

and APL-B), a distinction between complex and simple problems is not justified.

The complete worksample. As noted in the section describing the numerical facility worksamples, the Induction and Practice categories were meant as a necessary introduction to the more important Application category. As such, they were not intended as tests whose scores would be used in measuring numerical facility. Nevertheless, it is useful to examine the internal structure of these categories and their relationship to the Application category.

Almost all of the items in the Induction and Practice categories that were attempted were answered correctly (.95 and .98 for Induction and Practice, respectively). Consequently, as in the case of the Arithmetic worksample, it was not appropriate to estimate reliability at the item level. However, an estimate of reliability could be obtained by treating the various subtests as items. Using this strategy, the alpha reliability of the Induction category was .42, and that of the Practice category was .48.

As was expected, the reliability of the Induction and Practice categories was not high. Even more instructive were the correlations between all the subtests presented in Part A of Table 14. Subtests in the Induction and Practice categories tended to correlate about equally as well within their respective categories as with subtests of the other categories. Furthermore, each of the Induction subtests displayed its highest correlation with subtests in Practice: IND-A correlated highest with PRA-A, IND-B with PRA-B, and IND-C with PRA-C. All of the above suggested that the problems within the first two categories measured the same construct.

The subtests within Application displayed generally much higher correlations with each other as compared to their correlations with the other subtests. This indicated that the Application subtests measured the same construct, and that this construct was not measured, or not measured as well, by the problems in the first two categories. This was confirmed by factor analysis. Table 15 presents the results of factor analyses conducted on the correlation matrix in Part A of Table 14.

Principal axis factoring yielded three factors with eigenvalues greater than 1.0 (3.60, 1.37, 1.09). The relative size of these three factors indicated that a one-factor solution was optimal. This solution is presented in Part A of Table 15. As can be seen, the four highest loading subtests all come from Application. Furthermore, the differences in loadings among the Application subtests were much smaller than those between the Application and the other subtests. This suggested that although all the subtests primarily measured the same construct, Application measured this construct best.

Table 15

Factor Analyses of Rule Learning Subtests[1]

A. One-factor solution

Subtest	Factor 1
APL-C	.817
APL-B	.795
APL-D	.736
APL-A	.731
PRA-C	.503
IND-C	.388
PRA-B	.305
IND-B	.282
IND-A	.262
PRA-A	.262

B. Factor loadings for rotated[2] three-factor solution

Subtest	Factor 1	Factor 2	Factor 3
APL-B	.847	.172	.092
APL-D	.766	.171	.084
APL-C	.746	.314	.115
APL-A	.650	.312	.097
PRA-C	.278	.532	.104
IND-C	.156	.513	.093
PRA-B	.109	.415	.077
IND-B	.104	.332	.150
IND-A	.068	.139	.684
PRA-A	.089	.148	.541

[1] Initial extraction using principal axis factoring.

[2] Varimax rotation.

While the three-factor solution was not optimal, it is presented in Part B of Table 15 to lend additional support to the point made above. As can be seen, all the primary subtests of Factor 1 came from Application. However, the primary subtests of Factors 2 and 3 came in equal parts from Induction and Practice. In other words, Application measured the first factor best, while Induction and Practice measured it equally poorly.

Age effects. Table 16 presents a breakdown by age of examinees' scores on the Application worksample. In general, the older the examinee the lower the score. The correlation between age and performance on Application was -.31 (p < .001). As noted in earlier sections, this is an underestimate because of the overrepresentation of younger examinees. When a correction for restriction of range was applied, this correlation increased to -.40.

Sex effects. Females significantly outperformed males on the Application category by .13 of a standard deviation (female M = 31.30, male M = 29.72; p < .05), a relatively small difference. There were no significant differences between the performance of males and females on the Induction and Practice categories.

There did not appear to be any significant differences in the correlations between subtests for males and females. The correlations between the subtests in each category were similar for the two sexes. Thus, the reliability of the Application category was similar for the two. The correlations between the categories were similar for the two sexes as well. All of the above indicates that the internal structure of the complete worksample and of the individual categories did not differ between the sexes.

Summary. (a) The score that should be used from this worksample is examinees' total score on the Application category. While the construct measured by Induction and Practice was similar to that measured by Application, the first two did not measure this construct nearly as well as the latter.

(b) The reliability of Application was high, about .9.

(c) The Application problems were unidimensional.

(d) On the average, younger examinees performed better than older ones. The correlation between age and performance on Application was -.31 (-.40 when corrected for restriction of range).

(e) Females outperformed males on Application by .13 of a standard deviation.

(f) The correlations within and between categories did not differ for the two sexes.

The Numerical Facility Battery

The previous sections examined the internal structure of the four experimental worksamples in the numerical facility battery.

Table 16

Number of Application Items Answered Correctly by Age

Age interval	No. correct	N
13 - 15	31.30	24
16 - 18	34.33	222
19 - 21	35.34	199
22 - 24	31.89	170
25 - 27	30.77	99
28 - 30	32.82	108
31 - 33	28.34	90
34 - 36	25.99	77
37 - 39	27.84	75
40 - 42	24.23	61
43 - 45	23.82	49
46 - 48	22.74	38
49 - 51	21.48	25
52 - 54	17.40	10
55 - 57	18.57	11
58 - 61[1]	17.75	8
Total sample	30.50	1,266

[1] No 60-year-olds took the worksample.

It was demonstrated that each of the worksamples displayed two essential features of psychometrically sound measures. First, each was found to be relatively consistent internally, with reliability estimates ranging between .83 for Number Reasoning and .90 for Rule Learning. Second, each of the worksamples appeared to be unidimensional (i.e., measuring primarily _one_ and only one latent variable).

The next issue that needs to be addressed is the structure of the numerical facility battery as a whole. In the same way that each worksample was assessed by examining the pattern of correlations among the items within the worksample, the structure of the battery was assessed by examining the pattern of correlations among the worksamples within the battery.

As noted in the Introduction, the Johnson O'Connor test battery already contained the Number Series worksample, a measure hypothesized to be related to measures of numerical facility. In order to assess the degree of this relationship, most of the analyses described in this section included examinees' scores on Number Series along with their scores on the four tests of the numerical facility battery. In other words, Number Series was considered a part of the numerical facility battery for the purposes of most of the statistical analyses described in this section. The reliability of the Number Series worksample is .87. A more detailed description of the Number Series worksample and its internal structure can be found in the Number Series test manual.

Part A of Table 17 presents the zero-order correlations (simple Pearson coefficients) among the five numerical facility worksamples. As was noted in the introduction, Arithmetic has traditionally been the marker test for the numerical factor. Arithmetic correlated highest with Counting Backwards and lowest with Rule Learning. Arithmetic's low correlation with Rule Learning (.22) suggested that the two tests tended to measure different latent traits. Rule Learning also displayed a low correlation with Counting Backwards (.28). This was not surprising in light of the latter's strong relationship to Arithmetic.

Rule Learning displayed relatively strong relationships with Number Reasoning and Number Series, which, in turn, correlated only moderately with Arithmetic. This pattern of relationships suggests that, while all the worksamples of the battery shared some variance, there appeared to be two somewhat distinct types of problems in the battery. One type primarily involved simple computation (Arithmetic and Counting Backwards), while the other required a degree of reasoning as well (Rule Learning, Number Reasoning, and Number Series).

However, there was reason to believe that the correlations in Part A of Table 17 were influenced by spurious factors. This experimental battery was intended to measure numerical facility.

Table 17

Correlations Among Numerical Facility Worksamples

A. Zero-order correlations

Worksample	Arithmetic	Counting Backwards	Number Reasoning	Rule Learning	Number Series
Arithmetic					
Counting Backwards	49				
Number Reasoning	34	34			
Rule Learning	22	28	43		
Number Series	32	35	34	42	

B. Correlations with sex and age partialled out

Worksample	Arithmetic	Counting Backwards	Number Reasoning	Rule Learning	Number Series
Arithmetic					
Counting Backwards	50				
Number Reasoning	42	38			
Rule Learning	34	37	39		
Number Series	35	36	34	43	

Note. Ns = 1,244 for Part A and 1,242 for Part B. Decimal points omitted. All correlations significant at .001 level.

Consequently, only the shared "numerical facility variance" of these worksamples is of interest in this study. Differential correlations with age and sex (or any other variables that are not intended to measure numerical facility) are attributable to these worksamples' unique, nonnumerical components rather than to their shared variance. Partialling out this unique variance should provide a clearer picture of the interrelationships of these worksamples as they relate to numerical facility.

As was reported in earlier sections, all four experimental numerical facility worksamples displayed correlations with age that differed significantly from chance ($\underline{p} < .05$). On the average, older examinees did not perform as well as younger ones on all but the Arithmetic worksample, where this relationship was reversed. The correspondence between age and performance on Rule Learning was especially strong (-.31; -.40 when a correction for the restriction of range was employed). Some sex differences were also present in this battery. Males and females significantly differed in their performance on Rule Learning and Counting Backwards. On the average, females outperformed males on Counting Backwards, while the reverse was true for Rule Learning.

In order to control for sex and age, these variables were partialled out of examinees' scores on the various worksamples. The results of this procedure are displayed in Part B of Table 17. As can be seen, all but one of the coefficients were increased. The two relatively weak correlations involving Rule Learning were replaced by moderate correlations. Furthermore, the initial hypothesized distinction between computational and reasoning worksamples was blurred. While Counting Backwards and Arithmetic still displayed a strong relationship, the correlations among the reasoning tests did not tend to be stronger than those between the reasoning and computational tests.

It should be noted that the correlations resulting from the partialling of age alone were very similar to those displayed in Part B of Table 17 (which resulted from the partialling of age and sex). In other words, age was the variable that was primarily responsible for the discrepancies between the two matrices displayed in Table 17.

Table 18 presents the results of factor analyses of the correlation matrices in Table 17. In both analyses only one factor with an eigenvalue greater than one emerged. Thus, while a possible distinction between computational and reasoning tests emerged from the examination of the zero-order matrix, this distinction was not strong enough to produce a two-factor solution. Arithmetic's loading on the primary factor increased perceptibly after age and sex were partialled out. The other worksamples displayed similar loadings in both parts of Table 18. The primary factor in Part A of Table 18 explained slightly

Table 18

Factor Loadings of Numerical Facility Worksamples Including Number Series

A. Factor analysis[1] of zero-order correlation matrix

Worksample	Factor 1[2]
Counting Backwards	.62
Number Reasoning	.62
Number Series	.60
Arithmetic	.57
Rule Learning	.57

B. Factor analysis of partialled[3] correlation matrix

Worksample	Factor 1[4]
Counting Backwards	.66
Arithmetic	.66
Number Reasoning	.61
Rule Learning	.60
Number Series	.58

Note. *N*s = 1,244 for Part A and 1,242 for Part B.

[1] Factor analyses used principal axis extraction.

[2] Accounted for 48.3% of variance.

[3] Partialled for sex and age.

[4] Accounted for 51.1% of variance.

less variance than the primary factor in Part B of that table. This was a consequence of the generally lower correlations in the zero-order matrix as compared with the partialled matrix.

Table 19 presents the results of factor analyses that excluded Number Series. Generally, these did not differ greatly from the analyses presented in Table 18. The exception to this was Rule Learning, which displayed a comparatively low loading in the factoring of the zero-order correlation. While this loading was boosted when the partialled matrix was factored, it remained the lowest among the four worksamples. Rule Learning's slightly higher loadings in the presence of Number Series were a function of its relatively high correlation with that worksample (.43 after age and sex were partialled out). This may in fact have been due to a reasoning component that the two worksamples shared with each other but not with the other worksamples. However, the differences between the factor analyses that included Number Series and those that did not were too small to be of any practical significance.

Summary

(a) All the worksamples in the numerical facility battery were reliable.
(b) The numerical facility battery was unidimensional (even with the inclusion of Number Series).
(c) Age and sex suppressed some of the correlations between the worksamples in the battery. Once these variables were partialled out, the battery appeared more cohesive.

The Numerical Facility Battery and its Relationship to the Rest of the JOCRF Battery

The results reported in the previous section demonstrate that the worksamples of the numerical facility battery measure primarily one factor. This section will assess whether this factor is unique to the worksamples of the numerical facility battery or whether it is also measured by other worksamples in the JOCRF battery. Most of the analyses reported in this section used scores that were corrected for sex and age effects (see previous section for rationale).

Numerical Facility and the Cognitive Worksamples

Table 20 presents the correlations of the numerical facility worksamples with the cognitive tests of the JOCRF battery. Of the five numerical facility worksamples, Arithmetic showed the lowest average correlation with the cognitive worksamples (.14). In other words, Arithmetic was the most discriminantly valid of the numerical facility worksamples (with respect to the cognitive tests of the JOCRF battery). Rule Learning displayed the lowest discriminant validity; on the average, it correlated highest with

Table 19

Factor Loadings[1] of Numerical Facility Worksamples Excluding Number Series

A. Factor analysis of zero-order correlation matrix[2]

Worksample	Factor 1[3]
Counting Backwards	.65
Number Reasoning	.62
Arithmetic	.60
Rule Learning	.50

B. Factor analysis of correlation matrix partialled for sex and age[4]

Worksample	Factor 1[5]
Arithmetic	.69
Counting Backwards	.68
Number Reasoning	.61
Rule Learning	.55

[1] Principal axis factoring.

[2] *N* = 1,244.

[3] Accounted for 51.3% of variance in battery.

[4] *N* = 1,242.

[5] Accounted for 55.1% of variance in battery.

Table 20

Correlations Between Numerical Facility Worksamples and Cognitive Worksamples of JOCRF Battery

	Cognitive worksamples					
Numerical facility worksamples	Graphoria	Ideaphoria	Foresight	Ind. Reas.	Ana. Reas.	Wiggly Block
Arithmetic	47	15	09	08	11	01
Counting Backwards	31	12	09	14	25	19
Number Reasoning	35	18	13	27	35	22
Rule Learning	34	18	13	27	33	32
Number Series	27	17	15	16	36	32

	Cognitive worksamples					
Numerical facility worksamples	Paper Folding	Mem. Design	Silograms	Number Memory	Observation	Avg. corr.
Arithmetic	04	06	22	28	07	14
Counting Backwards	24	21	25	35	13	21
Number Reasoning	24	23	19	30	14	24
Rule Learning	39	38	27	38	29	30
Number Series	39	33	31	34	13	27

Note. *N* = 1,134. All decimals omitted. .08 < *r* < .11 is significant at the .01 level; *r* ≥ .11 is significant at the .001 level.

the other cognitive tests (.30). Number Series also demonstrated a relatively high average correlation with the other cognitive tests (.27).

Examining Table 20, one can see that all the numerical facility worksamples correlated moderately with Graphoria. While the correlation with Graphoria was not initially predicted, it was not surprising. First, the Graphoria worksample uses numbers. Second, Graphoria is a test of perceptual speed. It was noted in the Introduction that the Number factor has been found to correlate substantially with the Perceptual Speed factor. The correlation pattern between Graphoria and the numerical facility worksamples suggests that a perceptual component may have been responsible, in part, for this result. Graphoria correlated higher with the three worksamples requiring more perceptual speed (.46, .35, and .34, with Arithmetic, Number Reasoning, and Rule Learning, respectively), as compared with those requiring less perceptual speed (.31 and .27 with Counting Backwards and Number Series, respectively). However, the fact that Graphoria demonstrated a moderate correlation with Counting Backwards, a worksample with no apparent perceptual demands, indicated that perception alone could not completely explain Graphoria's relationship to the Number factor.

All the numerical facility worksamples also correlated moderately with Number Memory. This too was not surprising, because the ability to process numbers in short-term memory is common to most numerical tasks, including Number Memory.

The relatively high average correlations displayed by Rule Learning and Number Series are primarily due to these worksamples' correlations with the reasoning and spatial tests. Rule Learning correlated .33 with Analytical Reasoning and .27 with Inductive Reasoning, while Number Series correlated .36 with Analytical Reasoning. It should be noted that Number Reasoning also displayed moderate correlations with the two reasoning worksamples (.35 and .27 with Analytical Reasoning and Inductive Reasoning, respectively). Both Rule Learning and Number Series averaged .36 correlations with the two spatial tests, Wiggly Block and Paper Folding.

Table 21 presents a factor analysis of the cognitive worksamples in the Johnson O'Connor battery, including the worksamples of the experimental numerical facility battery. Five factors with eigenvalues larger than one emerged (5.2, 1.8, 1.5, 1.2, 1.1). Because of the relative sizes of the eigenvalues, this would have typically resulted in a one- or two-factor solution. However, the current study was less concerned with the ideal factorial solution than with the relationships among the various worksamples. Specifically, it was interested in whether the numerical facility worksamples "hang together" when pooled with other aptitude tests. It is for this reason that all five factors are reported in Table 21.

Table 21

Rotated Five-Factor Solution[1] for Worksamples in JOCRF Battery

Worksample	F1	F2	F3	F4	F5
Arithmetic	.84	-.08	.06	.07	.01
Counting Backwards	.58	.18	.18	.15	.00
Graphoria	.53	.00	.07	.14	.22
Number Reasoning	.52	.23	.12	.10	.22
Rule Learning	.44	.36	.11	.27	.20
Number Series	.41	.38	.29	.17	.00
Paper Folding	.09	.79	.13	.17	.03
Wiggly Block	.06	.68	.02	.12	.16
Memory for Design	.08	.51	.08	.49	.06
Analytical Reasoning	.15	.40	.32	.20	.28
English Vocabulary	.08	.15	.88	.10	.00
Reading Efficiency	.18	.11	.58	.11	.27
Foresight	.09	.04	.27	.03	.25
Number Memory	.32	.17	.11	.64	-.01
Silograms	.17	.08	.32	.59	.02
Observation	.05	.21	-.06	.46	.20
Inductive Reasoning	.09	.23	.12	.10	.60
Ideaphoria	.16	-.03	.28	.02	.30

[1] Principal axis factoring with varimax rotation (*N* = 1,136).

Table 21 shows that the numerical facility worksamples remained together even when factored with the other cognitive JOCRF worksamples. As expected, the worksample with the largest loading on the Number factor was Arithmetic. In addition to the five numerical facility worksamples, Graphoria also emerged as a major test on this factor. The relatively strong relationship between Graphoria and the numerical facility worksamples was dealt with earlier in this section.

It is very important to note that, at the same time that the numerical facility worksamples generally did not load on other factors, other worksamples generally did not load on the Number factor. This suggests that the moderate correlations between the numerical facility worksamples and some of the other cognitive tests were primarily due to individual worksamples' unique variance and not to the variance that the numerical facility worksamples share with one another. The Number Memory worksample was an exception to this, displaying a moderate loading on the Number factor (.32).

The rotated solution in Table 21 indicates that Arithmetic, Counting Backwards, Graphoria, and Number Reasoning tended to load almost exclusively on the Number factor (none demonstrate a loading higher than .23 on any other factor). Rule Learning and Number Series displayed moderate loadings on F2 (.36 and .38, respectively), a Spatial factor.

It is known from previous research done at the Foundation (Technical Report 1983-2) that Graphoria generally does not correlate strongly with the other aptitude tests. Consequently, Graphoria has always been viewed as a singlet--a test that stands alone among the other aptitude tests. There is generally little utility in including singlets in factor analyses. Table 22 presents the results of a factor analysis that excluded Graphoria. As in the first factor analysis, the numerical facility worksamples remained together. Also as in the earlier solution, Arithmetic, Counting Backwards, and Number Reasoning appeared to measure numerical facility exclusively, while Rule Learning and Number Series loaded moderately on the Spatial factor.

There is no readily available explanation of why Rule Learning and Number Series correlated with the Spatial factor while the other numerical facility worksamples did not. It is possible that the two tests, due to their strong reasoning component, demanded the use of a greater number of aptitudes as compared to the other numerical facility worksamples. Some support for this hypothesis can be found by observing the loadings of the numerical facility worksamples on factors three through five. While none of the numerical facility worksamples load highly on these factors, Rule Learning and Number Series demonstrated generally higher loadings than the other numerical facility worksamples. Indeed, Rule Learning and Number Series loaded highest of all the numerical facility worksamples on F4 (a Memory factor) and F3 (a Verbal factor), respectively.

Table 22

Rotated Five-Factor Solution[1] for Worksamples in JOCRF Battery, Excluding Graphoria

Worksample	F1	F2	F3	F4	F5
Paper Folding	.79	.11	.13	.17	.03
Wiggly Block	.67	.08	.02	.13	.17
Memory for Design	.51	.08	.08	.49	.06
Analytical Reasoning	.39	.16	.32	.20	.30
Arithmetic	-.11	.81	.06	.09	.04
Counting Backwards	.15	.60	.16	.16	.03
Number Reasoning	.20	.52	.13	.11	.25
Rule Learning	.34	.44	.12	.28	.21
Number Series	.36	.43	.28	.18	.02
English Vocabulary	.15	.10	.87	.10	-.02
Reading Efficiency	.11	.16	.59	.12	.24
Foresight	.03	.08	.29	.04	.24
Number Memory	.16	.32	.10	.65	.00
Silograms	.08	.16	.32	.59	.01
Observation	.21	.04	-.05	.46	.20
Inductive Reasoning	.22	.07	.14	.10	.62
Ideaphoria	-.04	.15	.29	.03	.29

[1] Principal axis factoring with varimax rotation (*N* = 1,136).

Numerical Facility and Noncognitive Worksamples

Table 23 displays the average correlations between the worksamples of the numerical facility battery and the 11 noncognitive worksamples of the JOCRF battery. The tests of the numerical facility battery generally displayed low correlations with these worksamples. The correlation between Mathematics Vocabulary and the numerical facility worksamples is given separately. It was expected that Mathematics Vocabulary would display relatively large correlations with numerical facility worksamples.

As can be seen from Table 23, the numerical facility worksamples generally displayed low correlations with the noncognitive worksamples of the JOCRF battery. Rule Learning and Number Series showed the highest average correlations, as they did with the cognitive worksamples.

Predicting Numerical Facility Scores From Scores on the Other Worksamples

It has been shown in previous sections that: (a) each numerical facility worksample measures primarily a single construct, and (b) all the numerical facility worksamples primarily measure the same construct. Points (a) and (b) suggest that the numerical facility battery is psychometrically sound. However, they do not address the question of whether examinees' numerical facility scores could be obtained more efficiently--that is, without having to administer the numerical facility battery.

The relatively low correlations between the numerical facility worksamples and the other worksamples of the JOCRF battery suggest that the construct measured by the numerical facility battery is unique to that battery. In other words, it suggests that numerical facility scores cannot be obtained from scores on the other worksamples. However, the moderate correlations between some numerical facility worksamples and other worksamples indicate that some overlap exists between the numerical facility construct and those constructs that are measured by the rest of the JOCRF battery.

To address the question of overlap between numerical facility and nonnumerical facility worksamples, each numerical facility worksample was regressed on all the other worksamples in the JOCRF battery. Table 24 presents the results of these regression analyses. In order to obtain additional information, the nonnumerical facility tests (independent variables) were entered into the analysis hierarchically in four steps: (a) cognitive tests excluding Graphoria, (b) noncognitive tests, (c) Graphoria, and (d) Mathematics Vocabulary.

Several interesting results emerged from the regression analyses. First, as was expected, the greatest overlap between

Table 23

Relationships Between Numerical Facility Battery and Noncognitive[1] Worksamples in JOCRF Battery

Numerical facility worksample	Avg. corr.[2] with noncognitive wks.	Corr. with Math. Vocabulary
Arithmetic	.09 (.02-.20)[3]	.31
Counting Backwards	.06 (.00-.13)	.32
Number Reasoning	.06 (.00-.13)	.34
Rule Learning	.10 (.01-.21)	.37
Number Series	.10 (.00-.25)	.53

[1] Color Perception, Personality, Tonal Memory, Pitch Discrimination, Rhythm Memory, Tweezer Dexterity, Writing Speed, Writing Hand, Finger Dexterity, Eyedness, Handedness.

[2] Average based on absolute values. All coefficients partialled for sex and age.

[3] Numbers in parentheses represent range of correlations on which average correlation is based.

Table 24

Regression Analyses Predicting Numerical Facility Worksamples Using Worksamples of JOCRF Battery[1]

Numerical facility worksample	Step[2]	Multiple R	R-squared
Arithmetic	Step 1	.37 (.35)[3]	.14
	Step 2	.40 (.36)	.16
	Step 3	.53 (.51)	.28
	Step 4	.56 (.53)	.31
Counting Backwards	Step 1	.42 (.40)	.18
	Step 2	.45 (.41)	.20
	Step 3	.48 (.45)	.23
	Step 4	.49 (.45)	.24
Number Reasoning	Step 1	.47 (.46)	.22
	Step 2	.48 (.45)	.23
	Step 3	.53 (.50)	.28
	Step 4	.56 (.53)	.31
Number Series	Step 1	.55 (.54)	.30
	Step 2	.57 (.55)	.33
	Step 3	.59 (.56)	.34
	Step 4	.63 (.61)	.40
Rule Learning	Step 1	.60 (.59)	.36
	Step 2	.61 (.59)	.37
	Step 3	.63 (.61)	.39
	Step 4	.64 (.62)	.41

[1] Sex and age were not partialled out of the variables used in this table. It was expected that this would not affect the results of the multiple regression.

[2] Step 1 - Cognitive worksamples excluding Graphoria
Step 2 - Noncognitive worksamples
Step 3 - Graphoria
Step 4 - Mathematics Vocabulary

[3] Number in parenthesis is multiple correlation adjusted for shrinkage.

nonnumerical and numerical facility worksamples was found for the Rule Learning and Number Series tests. Nevertheless, only about 40% of the variance in these two tests is explained by all the other JOCRF battery worksamples combined. Even less of the variance in the other three numerical facility tests was explained by the other JOCRF worksamples. In other words, the construct of numerical facility could not be measured without administering at least one worksample from the numerical facility battery.

Table 24 also indicates that the noncognitive tests explained relatively little variance in the numerical facility worksamples, after the variance explained by the cognitive tests was removed. In general, Steps 3 and 4 also explained relatively little variance in the numerical facility worksamples above and beyond that which was explained by the previous steps. This was not surprising in light of the fact that only one variable was entered in each of Steps 3 and 4, while a total of 28 variables were entered in the previous two steps. An exception to this was Arithmetic, where Graphoria (Step 3) added 12% in explained variance above and beyond that explained in Steps 1 and 2. There is no ready explanation for this result.

Summary

(a) In a factor analysis with the cognitive worksamples of the JOCRF battery, all the numerical facility worksamples loaded primarily on one factor.

(b) Of the five numerical facility worksamples, Rule Learning and Number Series generally showed the highest loadings on the other factors.

(c) Based on (a) and (b) it can be said that the numerical facility worksamples showed discriminant validity, but that Rule Learning and Number Series were not quite as discriminantly valid as the other three worksamples.

(d) When included in the factor analysis, Graphoria emerged as a major test on the Number factor.

(e) Outside of the five numerical facility worksamples and Graphoria, Number Memory was the only worksample from the regular JOCRF battery that showed even a moderate loading on the Number factor (.32).

(f) The correlations between the numerical facility worksamples and the noncognitive worksamples were generally very low.

(g) Multiple regression analyses showed that between 24 and 40% of the variance in the numerical facility worksamples can be explained by the cognitive and noncognitive tests of the JOCRF battery. Since about 85% of the variance in the numerical facility tests is reliable, between 45 and 61% of the reliable variance in the numerical facility worksamples cannot be explained by the other worksamples in the battery. This is further evidence that the numerical facility worksamples are discriminantly valid.

Numerical Facility and Choice of Major

Up to this point it has been shown that the numerical facility battery is internally reliable and discriminantly valid. The question remains whether it has criterion-related validity. That is, does it discriminate between individuals who differ in their numerical facility? One variable that was available for testing this question was examinees' choice of college major. Variance in choice of college major is associated, at least in part, with variance in ability. For example, it would be expected that an individual excelling in numerical facility would tend to choose a major in which he would have the opportunity to use this ability. An individual low on this ability would be expected to avoid a field in which manipulation of numbers is important.

Table 25 displays the scores of examinees on the numerical facility tests as a function of their college major. As can be seen, examinees in the "quantitative" majors scored higher on all the numerical facility worksamples than any of the other groups. The difference on the Arithmetic worksample between quantitative examinees and those choosing business or social science approached significance ($p < .10$). All other differences between quantitative and nonquantitative majors were significant at the .01 level. Some significant differences also emerged from the comparison of the nonquantitative majors: (a) business and social science majors scored higher ($p < .01$) than humanities majors on the Arithmetic and Counting Backwards worksamples, and (b) social science majors scored higher ($p < .05$) than business and humanities majors on the Rule Learning worksample. *T*-tests were used for all pairwise comparisons.

Summary

Examinees choosing majors that require quantitative thinking scored higher on the numerical facility worksamples than those choosing other majors. In addition, business and social science majors tended to score higher on these worksamples than humanities majors. In other words, with respect to the choice of college major, the numerical facility battery displayed criterion-related validity.

Further Development of the Numerical Facility Worksamples

As noted earlier, the alpha reliabilities of the numerical facility worksamples are moderately high. However, if a further increase in reliability is needed, it can generally be achieved by increasing the number of items in the worksample. In speeded tests such as Arithmetic and the Application section of the Rule Learning worksample, a proportional increase in time-on-task is also required.

Caution must be exercised when adding items to a particular worksample. Such items must be designed so that their

Table 25

Breakdown of Scores on Numerical Facility Worksamples by Major Field

	Difference from sample mean in standard deviation units			
Worksample	Quantitative[1]	Business	Social science	Humanities
Arithmetic	.24	.14	.13	-.11
Counting Backwards	.30	.12	.12	-.05
Number Reasoning	.34	-.03	.02	-.03
Rule Learning	.43	-.05	.06	-.06
Number Series	.44	.00	.05	-.04
N	76	178	154	165

Note. *N* on which overall sample mean is based = 771.

[1] Computer science, engineering, physical sciences, mathematics.

correlation with earlier items is maximized. It is important to note that factors such as fatigue, frustration, and boredom may affect the relationship between earlier and later items in a particular worksample. Also, item analyses must be conducted to verify the effectiveness of any items that are added to a worksample.

It was noted in the section investigating the internal structure of the Number Reasoning worksample that two items (8 and 9) displayed relatively low item-total correlations. These items should be replaced with items having higher item-total correlations.

The score used in the Counting Backwards worksample was speed rather than accuracy; speed displayed a much higher reliability than accuracy. Combining the two scores resulted in a reliability coefficient that was lower than that of speed alone. It is recommended that the role of accuracy in Counting Backwards be further investigated.

OVERALL SUMMARY

The earlier sections of this paper examined separately each of the worksamples in the numerical facility battery. It was concluded that each of the numerical facility worksamples measures primarily one construct. This was based upon: (a) moderately good reliabilities, ranging between .84 and .89, and (b) where appropriate, factor analyses that revealed an underlying unidimensional factor structure. Thus, based on internal structure alone, it cannot be said that any one worksample is significantly superior to the others.

After it was established that each of the numerical facility worksamples is moderately reliable and unidimensional, the worksamples were examined jointly. A factor analysis of the numerical facility worksamples yielded one factor. This indicates that the construct measured reliably by each of the individual worksamples is the same across the set. In other words, taken together, the numerical facility worksamples constitute a cohesive battery measuring a Number factor. Nevertheless, the worksamples differed in their loadings on the Number factor. Arithmetic and Counting Backwards displayed the highest loadings (.66), while Number Series, Rule Learning, and Number Reasoning displayed lower loadings (.61, .60, and .58, respectively). This suggests that Arithmetic and Counting Backwards measure the Number factor somewhat better than the other worksamples.

The next set of analyses examined the discriminant validity of the numerical facility worksamples with respect to the other worksamples in the JOCRF battery. In general, the numerical facility worksamples displayed little relationship with the worksamples of the standard battery, indicating good discriminant validity. The only exception to this was Graphoria, a test of clerical speed utilizing numbers, which displayed moderate correlations with the worksamples of the numerical facility battery. These analyses indicate that, in general, the Number factor is not measured by the other worksamples in the JOCRF battery.

A joint factor analysis of the cognitive worksamples of the standard battery and the numerical facility worksamples yielded a distinct Number factor. All of the numerical facility worksamples were primary defining items of this factor. Graphoria also emerged as a major item on the Number factor. When the factor analysis was repeated excluding Graphoria, the worksamples of the numerical facility battery remained in one factor. Furthermore, no other worksamples emerged as primary defining items of this factor.

While all the numerical facility worksamples loaded on one factor, they differed on: (a) the degree to which they loaded on the Number factor and (b) the degree to which they loaded on other factors. Arithmetic and Counting Backwards loaded highest on the Number factor (.84 and .58, respectively), while displaying relatively low loadings on nonnumerical factors. Rule Learning and Number Series displayed the lowest loadings on the Number factor (.44 and .41, respectively), while generally loading higher than the other numerical facility worksamples on nonnumerical factors. In other words, Arithmetic and Counting Backwards are more discriminantly valid than Rule Learning and Number Series.

A similar pattern, though less pronounced, emerged when the numerical facility worksamples were correlated with the noncognitive worksamples of the JOCRF battery. Arithmetic and Counting Backwards displayed lower average correlations with the noncognitive worksamples (.09 and .06, respectively), as compared with Rule Learning and Number Series (each correlating on average .10 with the noncognitive worksamples). Number Reasoning's discriminant validity was somewhat better than that of Rule Learning and Number Series but not quite as good as that of Arithmetic and Counting Backwards.

The final set of analyses examined the criterion-related validity of each of the numerical facility worksamples. The external criterion used in this study was examinees' choice of college major. Four categories of college major were identified: quantitative, business, social science, and humanities. The average score of examinees in each category of major was computed for each of the five numerical facility worksamples. Quantitative majors performed significantly better than humanities majors on every numerical facility worksample. However, the degree to which quantitative majors differed in their performance from nonquantitative majors varied across the numerical facility worksamples. The pattern that emerged here was a reversal of the pattern for discriminant validity. Arithmetic displayed the weakest criterion-related validity--quantitative majors did not perform significantly better on Arithmetic than business and social science majors (which was not the case for any of the other numerical facility worksamples). Next to Arithmetic, Counting Backwards displayed the weakest criterion-related validity, showing relatively small differences in performance between quantitative and nonquantitative majors. Rule Learning and Number Series displayed the strongest criterion-related validity. Number Reasoning showed criterion-related validity that was superior to Arithmetic and Counting Backwards but not as strong as that displayed by Rule Learning and Number Series.

Conclusion

Arithmetic and Counting Backwards are, psychometrically, the purest numerical facility measures in the five-worksample set. While their reliability is similar to the that of the other numerical facility worksamples, they display higher loadings on the Number factor and better discriminant validity. While least pure in the psychometric sense, Rule Learning and Number Series display the best criterion-related validity of the numerical facility worksamples. Number Reasoning shows discriminant and criterion-related validities that are in between those of the two pairs of worksamples noted above.

It is important to remember that all five worksamples are moderately reliable and that all measure the Number factor at least moderately well. In addition, all show at least moderate discriminant and criterion-related validities. Consequently, all can be considered to be suitable measures of numerical facility.

REFERENCES

Bechtold, H. P. (1947). Factorial investigation of perceptual speed factor. *American Psychologist, 2,* 300-305.

Becker, B. J. (1983). *Item characteristics and sex differences on SAT-M for mathematically able youths.* Paper presented at the annual meeting of the American Education Research Association, Montreal.

Bennett, G. K., Seashore, H. G., & Wesman, A. G. (1968). *Differential Aptitude Test* (4th ed.). New York: The Psychological Corporation.

Birren, J. E. (1974). Translations in gerontology--from lab to life: Psychophysiology and speed of response. *American Psychologist, 29,* 808-815.

Bromley, D. B. (1974). *The psychology of human ageing.* Middlesex, England: Penguin.

Chein, I. (1939). An experimental study of verbal, numerical and spatial factors in mental organization. *Psychological Record, 3,* 71-94.

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

Comrey, A. L. (1949). A factorial study of achievement in West Point courses. *Educational and Psychological Measurement, 9,* 193-209.

Coombs, C. H. (1941). A factorial study of number ability. *Psychometrika, 6,* 161-189.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16,* 297-334.

Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976). *Manual for Kit of Factor-Referenced Cognitive Tests.* Princeton, NJ: Educational Testing Service.

Forsyth, R. A., & Ansly, T. N. (1982). The importance of computational skills for answering items in a mathematical problem solving test: Implications for construct validity. *Educational and Psychological Measurement, 42,* 257-263.

French, J. W. (1951). *The description of aptitude and achievement tests in terms of rotated factors* (Psychometric Monographs No. 5). Chicago: University of Chicago Press.

Guilford, J. P., & Lacey, J. I. (Eds.). (1947). *Printed Classification Tests.* Army Air Forces Aviation Psychology Program Research Reports, No. 5. Washington, DC: United States Government Printing Office.

Halpern, D. F. (1986). *Sex differences in cognitive abilities.* Hillsdale, NJ: Erlbaum.

Kaiser, H. F. (1960). Varimax solution of primary mental abilities. *Psychometrika, 25,* 153-158.

Keats, J. A. (1965). An experimental study of cognitive factors. *Australian Journal of Psychology, 17,* 52-57.

Maccoby, E. E., & Jacklin, C. N. (1974). *The psychology of sex differences.* Stanford, CA: Stanford University Press.

Meece, J. L., Eccles, J. S., Parsons, J., Kaczala, C. M., Goff, R. B., & Futterman, R. (1982). Sex differences in mathematics achievement: Toward a model of academic choice. *Psychological Bulletin, 91,* 324-348.

Owens, W. A. (1966). Age and mental abilities: A second follow-up. *Journal of Educational Psychology, 57,* 311-325.

Salthouse, T. A. (1982). *Adult cognition.* New York: Springer-Verlag.

Schaie, K. W. (1980). Age change in intelligence. In R. L. Sprott (Ed.), *Age, learning ability, and intelligence.* New York: Van Nostrand Reinhold Company.

Technical Report 1983-2. *Large-sample test intercorrelations.* M. Daniel. Boston: Johnson O'Connor Research Foundation.

Technical Report 1985-3. *Numerical facility: A review of the literature.* J. Tal. Chicago: Johnson O'Connor Research Foundation.

Thurstone, L. L. (1938). *Primary mental abilities* (Psychometric Monographs No. 1). Chicago: University of Chicago Press.

Thurstone, L. L., & Thurstone, T. G. (1949). *Examiner manual for the SRA Primary Mental Abilities test.* Chicago: Science Research Associates.

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

Wechsler, D. (1958). *The measurement and appraisal of adult intelligence* (4th ed.). Baltimore: Williams and Wilkins.

Welford, A. T. (1977). Motor performance. In J. E. Birren & K. W. Schaie (Eds.), Handbook of psychology and aging. New York: Van Nostrand Reinhold Company.

Werdelin, I., & Stjernberg, G. (1969). On the nature of the perceptual speed factor. Scandinavian Journal of Psychology, 10, 185-192.

Werdelin, I., & Stjernberg, G. (1971). The relationship between factor loadings of some visual-perceptual tests. Scandinavian Journal of Psychology, 12, 21-28.

Willerman, L. (1979). The psychology of individual differences. San Francisco: Freeman.