LONG-TERM STABILITY OF 11 APTITUDE TESTS

Janine K. Bethscheider

and

David H. Schroeder

JOHNSON O'CONNOR RESEARCH FOUNDATION, INC.

Technical Report 1997-1

December 1997

Long-Term Stability of 11 Aptitude Tests

Janine K. Bethscheider
and
David H. Schroeder

ABSTRACT

It is widely accepted that cognitive abilities as well as general intelligence and personality are highly stable during adulthood. To date, however, only a modest amount of research has focused on the enduring stability of *specific* cognitive abilities, in spite of their relevance not only to aptitude testing programs that provide educational and occupational guidance but also to our understanding of general cognitive functioning.

For a number of years the Johnson O'Connor Research Foundation has collected retest data for the aptitude tests in its standard battery. Recently we analyzed the data accumulated thus far, which encompassed a variety of client ages from adolescence through adulthood and test-retest intervals that ranged from less than one year to over ten years. In this report, we present the stability results of this Foundation research for the following 11 tests: Number Checking, Ideaphoria, Inductive Reasoning, Analytical Reasoning, Wiggly Block, Memory for Design, Silograms, Number Memory, Observation, Word Association, and Eye and Hand. Of particular interest were the overall levels of stability for the nine specific cognitive ability measures included in the study. We also considered whether age, test-retest interval, and sex moderate the degree of stability observed.

All the tests showed a very substantial degree of long-term stability, with overall stability coefficients ranging from .62 (for Ideaphoria and Observation) to .76 (for Number Checking). In addition to Number Checking, the tests that were most stable over time were three of the Foundation's memory tests--Memory for Design, Silograms, and Number Memory. There also appears to be a nontrivial amount of short-term instability in the test scores. Nevertheless, it is a fairly small proportion of the variation for most of the tests, and there appears to be little additional change over long periods of time.

Our results clearly indicate that cognitive abilities of the type measured by the Foundation's aptitude test battery remain largely stable from adolescence through adulthood. Day-to-day variations in performance on these tests may occur, but the underlying aptitudes are highly stable over the years. This confirms a core tenet of the Foundation's testing program, namely that one's aptitudes will remain consistent over time. The present evidence thus provides a strong foundation for applications of cognitive ability tests in educational and vocational guidance.

CONTENTS

LIST OF TABLES AND FIGURE

ACKNOWLEDGMENTS

We would like to acknowledge the assistance of the test administrators and directors of the Foundation's testing offices, who collected the data we report here, and staff of the Research Department, who assisted in the preparation of the data for analysis. Also, the officers of the Foundation have generously supported this project throughout its course. In addition, Mark Daniel, of American Guidance Service, provided valuable advice regarding forming regression terms to test for the moderating effects of intertest time interval and age at first testing.

A portion of this research was presented previously in a poster by David Schroeder and Janine Bethscheider titled *Temporal Stability of Specific Cognitive Abilities* at the annual meeting of the American Psychological Association in Los Angeles in August 1994.

INTRODUCTION

An important attribute of an aptitude test score is its enduring stability over the course of time. Indeed, the effectiveness of the Foundation's aptitude testing program depends to a great extent on our capacity to assess individuals' aptitudes and provide guidance that will continue to apply to their lives years after they are tested.

A primary strategy for studying the stability of aptitudes is to administer the same (or similar) aptitude tests to the same persons at different points in time (i.e., collecting longitudinal data). The correlation between the examinees' scores on the first and second administrations is an indication of the degree to which the underlying aptitude is stable. In examining data of this type, one must bear in mind several factors that influence test scores and thereby affect stability coefficients.

One factor is the effect of successive administrations on test performance (i.e., practice effects). Test items encountered for a second time may be much easier for examinees because examinees may recall many of their responses from their initial testings. This means that improvement in scores on retesting could be the result of practice rather than real change in ability. In general, though, practice effects are most evident for retest intervals that are fairly short (Anastasi, 1976; Schaie, 1996).

A second factor is the age of examinees at testing and retesting in relation to the age curve, or developmental pattern, of the abilities being studied. In order to study the stability of aptitudes across a range of ages, one must have information about the ages of peak level of performance for each ability (i.e., the adult plateau) as well as the age at which diminution begins and the rate of the corresponding decline. For many cognitive abilities, performance peaks in young adulthood and then remains level through most of the adult years; for some abilities, however, performance peaks earlier or later in life--during adolescence or middle age, for example (Bloom, 1964; Nesselroade & Baltes, 1974; Schaie & Willis, 1996, chap. 12; see also Foundation work on age curves, e.g., Schroeder & Nakajima, 1997). In addition, some abilities begin to decline at a fairly early age in adulthood, whereas other abilities do not start to diminish until well into late mid-life or early old age (Cunningham & Owens, 1983; Horn & Donaldson, 1980; Schaie, 1993). The rate of decline will also vary: some abilities tend to show steep decrements in performance level with advancing age while others decrease more slowly (McCrae & Costa, 1994; Schaie, 1983, 1994). As a consequence, performance may vary at different stages of development along the age curve (Hoyer & Rybash, 1994). Thus, unless adjustments are made to the data, test scores are likely to be affected by age whenever the aptitudes under study have distinct age curves or the sample encompasses a variety of test-retest ages that correspond to different stages along the developmental curve.

A related issue is that of cohort, or generational, effects. (Cohort effects may be in evidence when, for example, persons born in 1945 tend to score lower or higher on average on an ability measure than persons born in 1975.) The reasons for generational differences in abilities are many. Often the later-born cohorts are more advantaged on cognitive ability tests because of improvements in education, nutrition, and medicine as well as cultural advances in such fields as communications and computer technology (Anastasi, 1976; Neisser et al., 1996; Schaie & Willis, 1996, pp. 384-386). Typically, the more recently born cohorts also are more experienced in aptitude test-taking (Neisser et al., 1996). Nevertheless, later-born groups have been found to score lower on some ability measures when compared with earlier-born cohorts (Schaie, 1993, 1994). Needless to say, the impact of these cohort effects will need to be addressed in any stability study.

Other considerations in the assessment of stability are the extent to which an observed long-term test-retest correlation deviates from 1.00 and the sources of variation that contribute to this departure from unity. The long-term test-retest correlation represents the proportion of variance that is stable in both the long and short term (i.e., the long-term stable variance). The rest of the variance in longitudinal studies can be subdivided into variance associated with true change and variance unrelated to true change. We can determine whether any true change has occurred by subtracting the long-term correlation from the short-term correlation. This difference, which provides an estimate of the variance that is stable in the short run but not the long run (i.e., short-term variance), represents the percentage of variance corresponding to the true change in aptitude levels that has occurred over time (i.e., long-term change). The remaining variance is associated with short-term fluctuations in performance that are unrelated to true change (i.e., error variance). These variations in performance may arise, for example, because of changes in the testing environment, such as timing errors or distracting noises, or changes in the condition of the examinee, such as fatigue, illness, or temporary mood (Anastasi, 1976). To reiterate, the sources of variation that contribute to performance on successive administrations of an aptitude test can be separated into the following components: (a) true variance that is stable in the long run; (b) true variance that is associated with change over the long run; and (c) error variance. By way of an example, if the long-term stability coefficient is .75 and the short-term correlation is .85, then we can say that 85 percent of the variance is attributable to true variance, of which 75 percent is stable for the long term and 10 percent (.85 - .75 = .10) is associated with long-term change. The remaining 15 percent (1.00 - .75 - .10 = .15) we would interpret as error variance.

Reviews of research on cognitive performance in adulthood have devoted remarkably little attention to the stability of individual differences[1] in specific abilities (Alwin, 1994; Schaie, 1996; Schaie & Willis, 1996). When research is presented, it usually focuses on the stability of general intelligence, which has been shown to be relatively stable during adulthood (see, e.g., Anastasi, 1958, pp. 231-249; Botwinick, 1977; Brody, 1992; Campbell, 1965; Carroll, 1993; Siegler, 1983; Willerman, 1979). The stability of personality also is now widely accepted (see, e.g., Hogan, Hogan, & Roberts, 1996, p. 473; McCrae & Costa, 1994). Surprisingly few studies, however, have investigated the stability of specific cognitive abilities (Carroll, 1993), in spite of their relevance to guidance and counseling and to understanding general cognitive functioning (Lowman, 1991; Snow, Kyllonen, & Marshalek, 1984).

One of the few longitudinal studies to include measures of specific abilities is Schaie's Seattle Longitudinal Study (SLS; Schaie, 1983, 1994). Data for the SLS have been collected during the course of six testing cycles spaced at 7-year intervals, starting in 1956 and continuing through

[1] It should be emphasized that there are two types of stability: stability of mean level (across age) and stability of individual differences. In this technical report, we address stability of individual differences, which refers not to an individual's test performance in absolute terms but rather to the individual's performance relative to an age cohort. Findings regarding stability of mean level have been reviewed by Schaie (1993, 1994).

1963, 1970, 1977, 1984, and, most recently, 1991. As part of this study, participants have been administered five cognitive ability tests from the Schaie-Thurstone Adult Mental Abilities Test (Schaie, 1985), which are described below. On the first occasion, in 1956, for example, cognitive ability test scores and other data were obtained for 500 adults ranging in age from 22 to 70 (Schaie, 1983, p. 79). For each successive testing cycle, as many individuals as possible were retested, and new participants were added to the study. At present, the SLS database consists of 4,132 adults, of whom 1,785 have been retested at least once. Of those for whom test-retest data are available, 860 have been tested a total of two times, 416 three times, 256 four times, 182 five times, and 71 all six times (numbers derived from Figure 1 of Schaie, 1994, p. 305).

The ability tests used in the SLS are: Letter Series, Figure Rotation, Number Addition, Word Fluency, and Verbal Meaning. A brief description of the five subtests follows.

1. The Letter Series test is a measure of inductive reasoning, i.e., the ability to solve logical problems. The test lasts 6 minutes and consists of 30 multiple-choice items. Each item of the test consists of a series of alphabetic letters. The examinee must discover the rule underlying the letter series and choose the letter that comes next in the sequence.

2. The Figure Rotation test is a measure of spatial orientation, which Schaie (1983) describes as "the ability to imagine how an object or figure would look when it is rotated, to visualize objects in two or three dimensions, and to see the relations of an arrangement of objects in space" (p. 73). The Figure Rotation test has a time limit of 5 minutes and consists of 20 items. For each item of the test, the examinee is given a two-dimensional stimulus figure and six choices and must indicate all the response alternatives that are rotated versions of the stimulus figure rather than mirror images.

3. The Number Addition test measures numerical facility, i.e., the ability to perform basic arithmetic operations quickly and accurately. The test lasts 6 minutes and comprises 60 items. Each test item is a simple addition task presented as a vertical column of numbers with a sum for the column of figures given at the bottom of the column. The examinee must indicate whether the sum given is correct or incorrect.

4. The Word Fluency test measures a specific aspect of divergent thinking, namely word fluency, which is the facility to recall words that meet a specified criterion (such as beginning or ending with a given letter or affix). In the Word Fluency test the examinee has five minutes to write as many words as possible beginning with the letter *s*.

5. The Verbal Meaning test measures a specific ability within the language domain, namely word, or lexical, knowledge. The test has a time limit of 4 minutes and contains 50 items arranged in order of difficulty. For each item, the examinee is given a stimulus word and must select the response alternative that is closest in meaning to the stimulus word.

Overall, all five ability tests showed high long-term stability, with stability coefficients ranging from .68 to .88, using data from Cycles 1 through 4[2] (Schaie, 1985). In general, the most

[2] To date, although a vast amount of data has been collected for the SLS, relatively few analyses of long-term stability have been reported, especially with regard to the data from Cycles 5 and 6.

stable tests were Number Addition and Letter Series and the least stable were Word Fluency and Figure Rotation. For test-retest intervals of seven years, Schaie (1985) has reported the following stability coefficients:[3] .82 to .86 for Letter Series, .72 to .81 for Figure Rotation, .80 to .82 for Number Addition, .75 to .78 for Word Fluency, and .79 to .81 for Verbal Meaning. The 14- and 21-year-interval data yielded similar results.[4] For retest intervals of 14 years, Schaie found stabilities of .82 to .83 for Letter Series, .68 to .70 for Figure Rotation, .81 to .88 for Number Addition, .70 to .71 for Word Fluency, and .77 to .78 for Verbal Meaning. The 21-year stabilities for the tests were as follows: .81 for Letter Series, .77 for Figure Rotation, .82 for Number Addition, .77 for Word Fluency, and .78 for Verbal Meaning.

Recently, a four-year longitudinal study of the stability of specific cognitive abilities was conducted by the Ball Foundation. For this study, Dawis, Goldman, and Sung (1992) tested 121 students at age 17-18 and again at age 21-22 on a set of tests that is quite similar to our (the Johnson O'Connor Research Foundation's; JOCRF's) battery of aptitude tests.

At the time of the study, the Ball Foundation battery comprised a total of 14 aptitude tests, including seven cognitive ability tests--Clerical, Idea Fluency (Uses), Ideaphoria, Inductive Reasoning, Analytical Reasoning, Shape Assembly, and Paper Folding--as well as Word Association and Vocabulary. (Measures of writing speed, finger dexterity, and grip were also part of the battery but are not discussed here because they have no direct bearing on this technical report.) Dawis et al. (1992) reported the following stabilities for these tests for the four-year retest interval: Clerical (Number Checking), .77; Idea Fluency, .57; Ideaphoria, .63; Inductive Reasoning, .41; Analytical Reasoning, .56; Shape Assembly (Wiggly Block), .60; Paper Folding, .80; Word Association, .59; and Vocabulary, .90. In short, the Vocabulary, Paper Folding, and Clerical tests proved to be the most stable, while Inductive Reasoning turned out to be the least stable over the time period.

The results of the SLS and the Ball Foundation study lend credence to the accepted notion that cognitive abilities are stable to a substantial extent over the long term. The findings, while promising, nevertheless are limited in several respects. The SLS findings, for example, are based on only five measures of specific abilities selected from the vast domain of cognitive ability tests, and the Ball Foundation findings are limited by the length of the retest interval and by the youthfulness of the adult sample. Therefore, any generalizations from these findings concerning the enduring stability of abilities during the adult years should be made with caution. That is to say, although the results of these two longitudinal studies have contributed greatly to the understanding that researchers now have regarding the long-term stability of specific abilities, further research is needed to extend and clarify this base of knowledge.

[3] Multiple values are presented because Schaie (1985) reported stabilities for three 7-year intervals: 1956-1963, 1963-1970, and 1970-1977.

[4] Because of the design of the SLS, a person retested at 14 years had been retested previously at 7 years, and a person retested at 21 years had been retested previously at 7 and 14 years. Schaie analyzed the effect of these successive administrations on test performance and reported the results in his 1996 book (pp. 215-218). He found that there were sizable effects for attrition in his study (up to one-half of an *SD*), but the effects of practice were minimal--no more than one-tenth of an *SD* in most cases.

For more than two decades, the JOCRF has been systematically collecting retest data for the aptitude tests in its standard battery. Preliminary analyses of the data collected in the 1970s were performed by Daniel (Statistical Bulletins 1976-20, 1977-25, and 1979-20) and Bethscheider (Statistical Bulletin 1989-5). Their findings, although for the most part supportive of the Foundation's contention that aptitudes are stable over time, were based on fairly small samples of examinees with retest intervals of one year or greater (*n*s ranged from 41 to 144). Since that time the Foundation has continued to gather retest data for its standard tests so that more-extensive analyses could be performed.

In this report, we present the results of this Foundation research for 11 aptitude tests, including nine specific cognitive ability measures, for samples that varied widely in age and test-retest interval. We analyzed the overall level of stability of scores and considered whether age, test-retest interval, and sex moderate the degree of stability observed.

METHOD

Samples

The samples were composed of clients of the JOCRF's aptitude-testing service, who paid a fee to receive assessment of their aptitudes, typically for purposes of educational planning and career guidance. The Foundation's client population is a relatively homogeneous group with respect to education and socioeconomic status and tends to be white, upper-middle-class, and college-bound or college-educated. Foundation clients can be presumed to be distributed across the ability range of that segment of the population. This represents a somewhat restricted range relative to the general population, and the stability coefficients for the general population are therefore likely to be somewhat higher than those we observed in this study.

Test-retest data for more than 4,500 Foundation clients were collected for this project. Because the focus of this study is the long-term stability of aptitudes, the primary samples were limited to those individuals with a test-retest interval of at least one year (approximately 77% of the cases collected). Samples composed of those individuals with shorter retest intervals were used as comparison groups for some analyses.

The following descriptive information helps characterize the primary samples, which ranged in size from 84 to 880 for individual tests. Approximately 39 to 55 percent of each sample was female. The age of examinees (clients) at first testing ranged from 14 to 58, and the median age of the samples was, with one exception, 19 or 20. The age-at-first-testing distributions were clearly skewed toward the younger end of the range, with approximately one-quarter of each sample initially tested at age 16 or 17. The number of months between test administrations ranged from 12 to 540, with the median interval length ranging from 60 to 89 months, with the exception of eye dominance. Although these values are spread across the interval range, the distributions of the test-retest interval are asymmetrical, with a greater representation of values at the low end of the distributions. Demographic data for the primary samples are provided in Table 1. As can be seen, these long-term samples are essentially comparable with regard to gender composition, age

Table 1

Descriptive Statistics for Samples[a]

		Age at first testing (years)			Test-retest interval (mos.)		
Test	% females	Median	*SD*	Range	Median	*SD*	Range
Number Checking	46.8	19	8.8	14-53	68	73.6	12-417
Ideaphoria	46.2	20	8.2	14-55	67	76.3	12-490
Inductive Reasoning	50.6	19	7.5	14-58	60	76.2	12-540
Analytical Reasoning	52.8	19	7.9	14-48	61	65.3	12-530
Wiggly Block	43.7	19	7.8	14-50	75	89.6	12-540
Memory for Design	47.0	20	9.0	14-50	79	71.0	12-373
Silograms	50.2	20	8.6	14-54	89	73.6	12-464
Number Memory	50.8	19	9.1	14-52	67	59.2	12-422
Observation	39.0	19	8.6	14-52	81	63.5	12-487
Word Association	46.7	19	8.0	14-54	65	80.4	12-464
Eye and Hand							
Eye	54.8	22	10.5	14-58	19	14.5	12-53
Hand	53.3	20	9.3	14-58	73	47.0	12-280

[a] Primary samples only (excluding examinees with intervals of less than one year).

of original testing, and length of retest interval. (The noticeable exception is the test-retest interval for the eyedness section of the Eye and Hand test. The reason for this much-lower median interval, discussed in more detail in the Measures section, is that only examinees tested in 1988 or later were included in this sample.) Compared with the general Foundation population of examinees, the primary samples tend to be younger.

With regard to the short-term samples (i.e., samples composed of those individuals with a test-retest interval of less than one year), size varied from 46 to 304 individuals. Females constituted approximately 41 to 59 percent of each sample. Age at original testing ranged from 15 to 57, with the median age of the samples falling in the 26-to-30-year-old range, with one exception. In general, the age distribution curves were uniform up to the late-thirties or early-forties, with a tapering off of cases thereafter. The test-retest intervals ranged from 0 to 11 months with median interval lengths between 7 and 9 months. For the most part, the interval distributions were spread fairly evenly across the entire range of values, except for a high concentration of cases at the upper end of the range, with approximately one-quarter to one-third of each sample retested 11 months after initial testing. (See Appendix A for demographic characteristics of the short-term samples.) The short-term samples are essentially comparable to each other with regard to gender composition, age of original testing, and length of retest interval. Compared with the long-term samples, the short-term samples tended to be older when initially tested.

Table 2 presents sample-size information for the primary samples as well as for several test-retest intervals and age-at-first-testing groups. Three subsamples based on retest interval were employed in this study. The first consisted of Foundation examinees with a test-retest interval of less than one year (i.e., the short-term samples). The other two interval groups were subsets of the primary samples, which were divided at the approximate median for all the tests, so that one group was composed of examinees with an interval of one year up to, but not including, six years and the other group consisted of examinees with an interval of six years or more. The comparison groups for age, which included only examinees in the primary sample, were (a) examinees first tested between the ages of 14 and 19 and (b) those initially tested at 20 years of age or older. Table 3 provides information on sample sizes separately for males and females.

Measures

The Foundation's multiple aptitude test battery covers a wide spectrum of aptitude areas including reasoning, memory, spatial abilities, idea production, auditory aptitudes, dexterity, personality, and laterality. In addition, the clients are given a knowledge test that measures their English vocabulary proficiency. The following 11 tests from the standard battery were selected for this study: Number Checking, Ideaphoria, Inductive Reasoning, Analytical Reasoning, Wiggly Block, Memory for Design, Silograms, Number Memory, Observation, Word Association, and Eye and Hand. Further information on these tests is provided in Table 4.

For each testing, the examinee was assigned a percentile score relative to Foundation examinees of his/her age. By using age-normed data, we eliminated (for the most part) population-wide age effects, which allowed us to isolate individual differences relative to age-characteristic performance. (The Ideaphoria test also utilized sex-based percentile norms to adjust for sex-related score differences on the measure.)

Table 2

Sample Sizes for Subsamples

Test	Overall[a]	Test-retest interval			Age at first testing[a]	
		Under 1 year	1 yr. - 6 yrs.[b]	6 yrs. & over	14-19	20+
Number Checking	314	92	165	149	166	148
Ideaphoria	880	304	466	414	405	475
Inductive Reasoning	473	147	275	198	256	217
Analytical Reasoning	235	46	136	99	127	108
Wiggly Block	254	71	120	134	139	115
Memory for Design	185	63	87	98	88	97
Silograms	209	59	85	124	104	105
Number Memory	187	62	103	84	100	87
Observation	231	70	94	137	125	106
Word Association	428	94	232	196	229	199
Eye and Hand						
Eye	84	87	84	0[c]	30	54
Hand	225	87	110	115	109	116

[a] Primary samples only (excluding examinees with intervals of less than one year).
[b] Interval extends up to, but does not include, 6 years.
[c] Because there were no examinees in the 6-years-and-over group, the overall group is the same as the 1-year-to-6-years interval group.

Table 3

Sample Sizes for Subsamples for Males and Females Separately

Test	Overall[a]	Test-retest interval			Age at first testing[a]	
		Under 1 year	1 yr. - 6 yrs.[b]	6 yrs. & over	14-19	20+
Number Checking						
Males	167	54	84	83	99	68
Females	147	38	81	66	67	80
Ideaphoria						
Males	473	140	240	233	237	236
Females	406	164	226	180	168	238
Inductive Reasoning						
Males	234	81	121	113	133	101
Females	239	66	154	85	123	116
Analytical Reasoning						
Males	111	20	54	57	57	54
Females	124	26	82	42	70	54
Wiggly Block						
Males	142	41	65	77	80	62
Females	110	30	55	55	59	51
Memory for Design						
Males	98	30	45	53	49	49
Females	87	33	42	45	39	48
Silograms						
Males	104	31	38	66	52	52
Females	105	28	47	58	52	53

(*table continues*)

Test	Overall[a]	Test-retest interval			Age at first testing[a]	
		Under 1 year	1 yr. - 6 yrs.[b]	6 yrs. & over	14-19	20+
Number Memory						
Males	92	29	51	41	49	43
Females	95	33	52	43	51	44
Observation						
Males	141	32	63	78	76	65
Females	90	38	31	59	49	41
Word Association						
Males	228	39	112	116	135	93
Females	200	55	120	80	94	106
Eye and Hand						
Eye						
Males	38	39	38	0[c]	15	23
Females	46	48	46	0[c]	15	31
Hand						
Males	105	39	49	56	58	47
Females	120	48	61	59	51	69

[a] Primary samples only (excluding examinees with intervals of less than one year).

[b] Interval extends up to, but does not include, 6 years.

[c] Because there were no examinees in the 6-years-and-over group, the overall group is the same as the 1-year-to-6-years interval group.

Table 4

Aptitude Tests in the Study

Name	Internal-consistency relia.	Ability measured
Number Checking	.96	Clerical speed and accuracy. Test involves quickly comparing pairs of numbers to see whether they are the same or different.
Ideaphoria	.97	Rate of flow of ideas (ideational fluency).
Inductive Reasoning	.84	Quickness in seeing relationships among separate facts, ideas, or observations.
Analytical Reasoning	.75	Ability to arrange ideas into a logical sequence.
Wiggly Block	.73	Structural visualization: ability to visualize three-dimensional forms. Test involves reconstructing three-dimensional blocks.
Memory for Design	.80	Memory for straight-line patterns.
Silograms	.92	Associative memory for verbal material.
Number Memory	.82	Memory for numbers.
Observation	.62[a]	Memory for fine visual details.
Word Association	.89	Tendency to react to experience from a general, objective viewpoint versus a narrow, subjective viewpoint. Describes how well-suited a person is for working in a group (Objective) versus working on one's own as an individual (Subjective).
Eye and Hand	—	Eye dominance and hand dominance.

[a] Odd-even reliability rather than internal-consistency reliability.

The percentile scores obtained for each examinee in the study were based on the norms in use at the time of that examinee's test administration except when (a) sex-based norms were used with earlier but not more-recent forms of a test or (b) the norms were known to have been inaccurate (i.e., too hard or, more commonly, too easy). In the former cases, percentile scores based on separate-sex or males-only norms were replaced with scores based on combined-sex norms. In the latter cases, the percentile scores were altered to correspond to more-accurate norms than were available when the test was originally administered. Often we were able to use later, more-accurate norms because the tests had not changed substantively in the meantime. When this was not possible, however, we calculated new percentiles using procedures detailed in a 1975 report by Daniel and a 1994 report by Bethscheider. (These calculated percentiles were based on obtained scores from the particular time period rather than on analytically derived simulated scores.)

Because percentiles are not interval-level units of measurement, each percentile score was converted to the value in the normal distribution that corresponds to that percentile score. Such a conversion transformed the scores to fit a normal curve that has a mean of 0 and standard deviation of 1. This means examinees' scores were expressed in the same form as standard scores, or z-scores (i.e., with a mean of 0 and *SD* of 1), so that henceforth we will refer to them as z-scores. (In this type of transformation, a percentile score of 50 corresponds to a z-score of 0.) The purpose of this conversion was to place scores on an interval scale of measurement so that parametric methods of analysis could be applied to the data. All analyses performed for this project used these z-scores rather than percentile scores.

Word Association is scaled differently than the other tests in the Foundation battery. Individuals scoring in the upper three-quarters of the score distribution are considered objective, and those scoring in the bottom quarter of the distribution are considered subjective. This bidirectional, one-fourth/three-fourths distribution remained intact with the conversion of Word Association scores to z-scores. This means that the division between the subjective and objective ranges lies between z-scores of -.69 and -.63, and a z-score of 0 corresponds to approximately the 67th certile objective. (For more information on Word Association certiles and z-score conversions in general, refer to Bethscheider, 1986.)

Instead of percentile scores transformed to z-scores for Eye and Hand, raw scores on the test were converted to eyedness and handedness ratio scores. Handedness scores were computed as the ratio of the number of trials on which the right hand was used to the total number of right-hand and left-hand responses given on the handedness section of the Eye and Hand test[5], so that a handedness ratio of 0 indicated a left-handed individual, a ratio of 1 indicated a right-handed person, and a ratio between 0 and 1 indicated a variable-handed examinee. For Eye and Hand tests administered in 1988 or later, eyedness scores were computed as the ratio of the number of trials on which the right eye was used as the sighting eye in Part 1 of the eyedness portion of the test to the total number of trials administered in Part 1. An eyedness ratio of 0 identified an individual as completely left-eyed, a ratio of 1 as completely right-eyed, and a ratio between 0 and 1 as variable-eyed. Eye dominance tests administered prior to 1988 were scored on Parts 1-3 of the test, which means that individuals could be classified as variable-eyed because they

[5] On each trial of the handedness section of the Eye and Hand test, both a right-hand and left-hand response could be given, for a maximum of two points per trial.

spontaneously switched eyes in Parts 1 or 2 or they switched eyes when directed to use their nondominant eye in Part 3. Because there was no way of distinguishing the spontaneous switchers, or truly variable-eyed, from those who could switch when instructed to do so, all pre-1988 data were excluded from the analyses of the Eye section of the test.

Procedures

Data Collection

Each test was administered to at least 171 Foundation examinees on two separate occasions. At their initial testing, examinees were given the standard battery of Foundation aptitude tests. Then at a later date, these examinees (at their initiative) requested a "follow-up" discussion of their test results. At that time, they were asked to take a retest of a given test. For example, follow-up clients between May 1987 and January 1988 were asked to take a retest of our Analytical Reasoning test. Between 1987 and 1994, retests of each of the 11 tests in this study were administered for a period of time.[6] Retest data collected during the 1970s for five of the tests (Number Checking, Ideaphoria, Inductive Reasoning, Wiggly Block, and Word Association) also were included in the analyses. For the most part, then, each test had a different sample: overall, approximately 95% of the cases can be found in only one sample, and 5% in at least two samples; of the latter, 34 examinees retook the entire Foundation battery and are included in the sample for each test.

All testing was conducted at Foundation offices by trained test administrators. Oral instructions preceded all the tests. Inductive Reasoning, Analytical Reasoning, Wiggly Block, Observation, Word Association, and Eye and Hand were administered individually; Number Checking, Ideaphoria, Memory for Design, Silograms, and Number Memory were group-administered using taped and written instructions and, in some cases, slide presentations.

In general, examinees were retested on the same form or a slightly revised form of the test they took earlier. The exception to this pattern was Ideaphoria, for which about half the examinees retook the same form and half retook a parallel form. We analyzed the results separately for the two groups, and because the results were essentially the same, we report the combined results here.

[6] It should be noted that this data-collection design created some confounding of variables that are of interest in this study. Because the year of retesting was about the same for most examinees in each sample, the year of first testing and the test-retest interval are strongly related to each other (i.e., the earlier the year of first testing, the larger the interval). Furthermore, in earlier years the Foundation tested more males and younger examinees, so that age at first testing and, in some cases, sex also are confounded with year of first testing and test-retest interval. (This was borne out by analyses we performed comparing various groups of examinees--e.g., early age versus later age at first testing--on other variables such as sex, year of first testing, and length of test-retest interval.) Nevertheless, we do not feel that these factors compromise the validity of our findings because the scores we used are based on age-normed and, when appropriate, sex-normed percentiles (see Measures section).

Analyses

Our initial analyses examined practice effects (i.e., the average improvement from the first to the second administration on the tests) for the primary samples as well as for various subsamples. To be specific, *practice effect* was defined as the mean for the second administration minus the mean for the first administration standardized in terms of the Foundation testing population's standard deviations. (Because the Foundation population *SD*s are 1 in this study for the test scores based on z-scores, the practice effect is equal to the difference in scores between the two administrations for each Foundation test except Eye and Hand.)

We evaluated practice effects for the primary samples by comparing mean scores (in z-score units) at initial testing with those at retesting using *t*-tests for repeated measures, or paired samples. The size of the effect associated with each significant *t*-test was calculated using a population standard deviation of 1, thus making effect size the same as the difference between z-score means. (Note that for this study, the effect size for each sample is equivalent to the practice effect.) Differences in practice effects across the three interval groups (i.e., the three subsamples that varied with regard to the length of their retest interval) were evaluated using one-way analyses of variance (ANOVAs); *t* tests were employed to compare the practice effects for the two age groups.

Two additional sets of analyses of practice effects were undertaken. We examined practice effects for males and females separately. We also investigated practice effects separately for same-form and different-form retests on the three tests for which we had alternate-form data--namely, Number Checking, Ideaphoria, and Number Memory.

The primary focus of this study was the test-retest stabilities of the Foundation tests. We not only analyzed the overall level of stability of scores but also considered whether age, test-retest interval, and sex moderate the degree of stability observed.

Each stability coefficient was calculated as the correlation between scores from the first and second administrations. Differences in stability coefficients for the various subsamples of interest were assessed using tests of significance for differences in correlation coefficients from independent samples. This required applying Fisher's Z transformation to the stability coefficients and then computing the z statistic for the two-group comparisons (i.e., age at first testing, sex, and alternate form taken) and the *V* statistic for the three-group comparison (i.e., length of interval between testings). (See Hays, 1973, pp. 661-664, for more information on these statistics.) We also performed multiple regressions designed to evaluate the extent to which stability is affected by the intertest time interval and the examinee's age at first testing. Although these variables appear to affect stability modestly for most of the tests, in most cases the effect was too small to reach statistical significance with our sample sizes; thus, we only mention trends in the Results section.

The simple (uncorrected) correlation coefficients between initial test scores and retest scores are attenuated by the effect of short-term fluctuation in addition to long-term change on retest scores. To correct for this, we divided the overall long-term stability coefficients by the short-term stability coefficients, which yields estimates of the long-term stability we would observe if scores at both test administrations were free of error (i.e., true scores). These disattenuated coefficients reflect how much of the test-score variance that is stable in the short term continues to be stable in the long term. These values have been interpreted as the percentage of true-score variance in traits that is stable over time (McCrae & Costa, 1990).

Because ratio scores rather than z-scores were used for our laterality measures, we did not use correlations to assess these test-retest stabilities. Instead, as a measure of stability, we computed summary stability indices defined as the percent of examinees who remained stable in terms of their eyedness or handedness category. Ratio scores for eye dominance were partitioned into three eyedness categories: (a) completely left-eyed if the ratio was 0, (b) variable-eyed if the ratio was between 0 and 1, or (c) completely right-eyed if the ratio was 1. Scores for hand dominance were partitioned into five handedness categories: (a) essentially left-handed (for a ratio of 0 to .10), (b) primarily left-handed (for a ratio of .11 to .30), (c) variable-handed (for a ratio of .31 to .69), (d) primarily right-handed (for a ratio of .70 to .89), or (e) essentially right-handed (for a ratio of .90 to 1). By way of illustration, for hand dominance, examinees with ratio scores of 0 at initial testing and retest scores of .10 would be counted as examinees who remained stable in terms of their handedness category, because both their test and retest scores correspond to the "essentially left-handed" category. A score of .35 at first testing, which corresponds to the "variable-handed" category, and a score of 0 at second administration, which corresponds to the "essentially left-handed" category, would be considered an example of unstable handedness. Differences in stability indices were evaluated using tests of significance for differences between independent proportions.

The significance level was set at .05 for all statistical tests reported in this document. Unless otherwise stated, the *SPSS/PC+ Base* (Version 4.0; Norusis, 1990a) and *SPSS/PC+ Statistics* (Version 4.0; Norusis, 1990b) computer software packages were used for the analyses.

RESULTS

Descriptive Statistics

Table 5 presents the means and standard deviations of the z-scores at original testing and retesting for the primary samples. (See Appendix B for a similar table for the short-term samples.) In general, the means and *SD*s at initial testing were around zero and 1.00, respectively. With regard to the Word Association sample, 69% was objective and 31% subjective. On the Eye and Hand test, approximately 33% of the Eye sample was identified as completely left-eyed, 66% completely right-eyed, and 1% variable-eyed; about 5% of the Hand sample was identified as essentially left-handed, 3% as primarily left-handed, 5% as variable-handed, 5% as primarily right-handed, and 82% as essentially right-handed. This means that, in terms of test scores, the primary samples are fairly representative of the Foundation testing population, which, as noted previously, is somewhat different from the general population. The means for the short-term samples (Appendix B) ranged from -.18 for Number Memory to .37 for Silograms. A substantial portion of the deviations from zero, however, are probably due to sampling error.

Practice Effects

The differences in z-scores between the two test administrations, which can be interpreted as practice effects, can be found in the last column of Table 5. The *t*-test values for the differences in

Table 5

Descriptive Statistics for Test Scores[a]

Test	*n*	Z-score at Time 1		Z-score at Time 2			Standardized practice effect[b]
		Mean	*SD*	Mean	*SD*	*t*	
Number Checking	314	.03	.96	.13	.99	-2.54*	.10
Ideaphoria	880	.14	.97	.16	1.00	-.60	.02
Inductive Reasoning	473	.16	.92	.38	.97	-5.98*	.22
Analytical Reasoning	235	-.07	.88	.41	.93	-9.54*	.48
Wiggly Block	254	-.04	.87	.16	1.04	-3.88*	.20
Memory for Design	185	.12	1.05	.38	.97	-4.70*	.26
Silograms	209	.15	.88	.40	.99	-5.20*	.25
Number Memory	187	-.03	1.01	-.06	.95	.69	-.03
Observation	231	.05	.99	.29	.97	-4.31*	.24
Word Association	428	-.15	.99	-.15	.93	.00	.00
Eye and Hand							
Eye	84	.67	.47	.66	.47	.06	.00
Hand	225	.89	.26	.89	.26	.02	.00

[a] Primary samples only (excluding examinees with intervals of less than one year). With the exception of Word Association and Eye and Hand, mean test scores are based on percentile scores that were converted to standard scores, or z-scores. For Word Association, mean test scores are based on certile, rather than percentile, scores that were converted to z-scores, with high scores indicating Objectivity and low scores indicating Subjectivity. For Eye and Hand, mean test scores are based on raw scores that were converted to ratios, so that a ratio of 1 indicates a completely right-eyed or right-handed person and a ratio of 0 a left-eyed or left-handed individual.

[b] Effect size was calculated using a population standard deviation of 1, thus making effect size the same as the difference between the z-score means, except for Eye and Hand.

*$p < .05$

scores from the two administrations also are shown in this table along with estimates of the size of the practice effects in standardized units. (It bears repeating that practice effects and effect sizes are equivalent in this project. As mentioned earlier in the Analyses section, *practice effects* are the average improvement from the first to the second administration on the tests and in this study are equal to the differences in *z*-scores between the two administrations for each Foundation test except Eye and Hand. Effect sizes likewise are the same as the differences between means because the Foundation population *SD*s are 1 in this study, with the exception of Eye and Hand.)

Practice effects for the primary samples ranged from -.03 *SD* to .48 *SD*. As evidenced in the table, retest scores were significantly higher than initial scores for seven of the aptitude tests--Number Checking, Inductive Reasoning, Analytical Reasoning, Wiggly Block, Memory for Design, Silograms, and Observation--with Analytical Reasoning showing the greatest average improvement from first to second administration. On the other four Foundation tests, there were no statistically significant differences in performance on the two testings. Effect sizes for the seven score differences that were significant ranged from .10 to .48. Of these, the magnitude of the effect associated with each test was, with one exception, at least one-fifth of a standard deviation. Thus, in terms of Cohen's conventional criteria (1988, pp. 24-27), the effect sizes for Inductive Reasoning, Analytical Reasoning, Wiggly Block, Memory for Design, Silograms, and Observation can be considered in the small to medium range. For Number Checking, the difference in performance, albeit statistically significant, is too small to be of practical importance.

Table 6 augments the practice-effects data contained in Table 5 by including the three test-retest intervals and two age-at-first-testing groups in addition to the primary sample. (Thus, the overall practice-effects column in Table 6, which lists the practice effects for the primary samples, is the same as the last column in Table 5. Likewise, the second column in Table 6, which lists the practice effects for the short-term samples, is the same as the last column in Appendix B.) For examinees with a test-retest interval of less than one year, practice effects on the cognitive tests ranged from .18 to .57 *SD*s, whereas practice effects for Word Association and Eye and Hand were minimal (in the range of .04 to .06 *SD*s). Examinees with a retest interval of between one and six years likewise performed better on the second administration, with practice effects ranging from .07 to .56 *SD*s on the cognitive and Word Association tests and no effect for Eye and Hand. There was less evidence of a practice effect when the interval was six years or more. For examinees with a test-retest interval of at least six years, practice effects ranged from -.17 to .38 *SD*s on the cognitive and Word Association tests and again were negligible for Eye and Hand. Differences across interval groups were significant for all but Analytical Reasoning and Eye and Hand, with the practice effect more pronounced for shorter test-retest intervals.

With regard to age at first testing, the practice effect generally was more pronounced for examinees initially tested between the ages of 14 and 19. Only for Wiggly Block, however, was the difference between age groups significant.

In general, then, examinees performed better on the second administration than on initial testing by a few tenths of a standard deviation, and the size of the practice effect is affected by length of the retest interval: the greatest practice effect was found for examinees with a short-term retest interval, followed by those with a retest interval of 1 to 6 years and then those with a retest interval of 6 years or more. Thus, the effect appears to be mainly due to memory, although general familiarity with the test formats and content may also be factors here.

Table 6

Practice Effects for Aptitude Tests

		Test-retest interval			Age at first testing[a]	
Test	Overall[a]	Under 1 year	1 yr. - 6 yrs.[b]	6 yrs. & over	14-19	20+
Number Checking	.10	.20*	.19*	-.01*	.15	.03
Ideaphoria	.02	.18*	.10*	-.08*	.04	.00[c]
Inductive Reasoning	.22	.51*	.24*	.19*	.25	.19
Analytical Reasoning	.48	.33	.56	.38	.51	.45
Wiggly Block	.20	.57*	.34*	.07*	.33*	.04*
Memory for Design	.26	.49*	.41*	.12*	.33	.20
Silograms	.25	.53*	.34*	.18*	.20	.30
Number Memory	-.03	.46*	.07*	-.17*	.01	-.10
Observation	.24	.49*	.32*	.19*	.20	.29
Word Association	.00[c]	.04*	.10*	-.12*	.00[c]	.00[c]
Eye and Hand						
Eye	---[d]	.06	.00[c]	---[d]	.05	-.04
Hand	.00[c]	.04	.00[c]	.00[c]	.00[c]	.00[c]

Note. With the exception of Word Association and Eye and Hand, practice effects are based on percentile scores that were converted to standard scores, or z-scores. For Word Association, practice effects are based on certile, rather than percentile, scores that were converted to z-scores, with high scores indicating Objectivity and low scores indicating Subjectivity. For Eye and Hand, practice effects are based on raw scores that were converted to ratios, so that a ratio of 1 indicates a completely right-eyed or right-handed person and a ratio of 0 a left-eyed or left-handed individual.

[a] Primary samples only (excluding examinees with intervals of less than one year).
[b] Interval extends up to, but does not include, 6 years.
[c] The value is less than ±.005 and therefore is rounded to .00.
[d] Because there were no examinees in the 6-years-and-over group, the overall group is the same as the 1-year-to-6-years interval group.

*$p < .05$ for comparisons of practice effects across the three intervals (ANOVAs) or between the two age groups (*t*-tests)

Practice effects for males and females separately are presented in Tables 7 and 8, respectively. Overall, for both sexes, the greatest long-term practice effect was associated with the Analytical Reasoning test, followed by Observation, Silograms, Inductive Reasoning, and Wiggly Block for women and Memory for Design, Silograms, Observation, and Inductive Reasoning for men. The tests with the greatest short-term practice effect were Number Memory for men and Inductive Reasoning for women.

For males, differences across interval groups were significant for all but Number Checking, Memory for Design, eye dominance, and hand dominance. For females, differences were significant for only five of the tests: Number Checking, Ideaphoria, Inductive Reasoning, Memory for Design, and Number Memory. Differences between the two age-at-first-testing groups were significant for females on Analytical Reasoning, Wiggly Block, and hand dominance and on Wiggly Block for males, with the practice effect more pronounced in each case for examinees initially tested between the ages of 14 and 19 than for examinees 20 and over.

A comparison between male and female practice effects revealed several significant differences. For a retest interval of 1 to 6 years, males showed a substantially greater practice effect than females on Analytical Reasoning, $t(134) = -2.37, p < .05$. For a retest interval of at least 6 years, the practice effects of men and women were significantly different for Word Association, $t(194) = 2.53, p < .05$. In addition, for examinees first tested at age 20 or older, there were significant differences in the practice effects of males and females on Ideaphoria, $t(472) = 2.06$, $p < .05$, and hand dominance, $t(61.54) = -2.04, p < .05$, although these effects were small to begin with in both cases.

In general, the practice effects obtained separately for males and females were similar to those obtained for the full sample for each test.

Table 9 displays the practice-effects data separately for same-form and different-form retests on three tests in this study--Number Checking, Ideaphoria, and Number Memory. There were no significant differences in practice effects between examinees who were administered the same form at their second testing and those who were given an alternate form at their retesting. Because the results for all analyses were essentially the same regardless of the form taken, the stability coefficients for these tests (see below) are based on combined data.

Stability

The stability coefficients for the study are shown in Table 10. As can be seen from column 1, all the tests showed substantial long-term stability, with overall test-retest correlations ranging from .62 (for Ideaphoria and Observation) to .76 (for Number Checking). In addition to Number Checking, the tests that were most stable over time were three of the Foundation's memory tests--Memory for Design, Silograms, and Number Memory.

Stability coefficients for test-retest intervals of less than one year (i.e., short-term stability) ranged from .56 (for Observation) to .85 (for Number Checking). In addition to Number Checking, the most-stable tests in the short term were Wiggly Block, Silograms, and Word Association--all with coefficients above .80. For retest intervals between one and six years, the stability coefficients ranged from .60 (for Analytical Reasoning) to .79 (for Number Checking); for retest intervals of six or more years, they ranged from .52 (for Ideaphoria) to .74 (for Number Checking). For the most part, the length of the test-retest interval affected the degree of stability,

Table 7

Practice Effects for Aptitude Tests (Males)

		Test-retest interval			Age at first testing[a]	
Test	Overall[a]	Under 1 year	1 yr.- 6 yrs.[b]	6 yrs. & over	14-19	20+
Number Checking	.12	.13	.18	.06	.16	.06
Ideaphoria	-.01	.14*	.08*	-.11*	.05	-.08
Inductive Reasoning	.19	.46*	.21*	.15*	.25	.10
Analytical Reasoning	.52	.28*	.75*	.29*	.44	.60
Wiggly Block	.19	.63*	.41*	.00*[c]	.32*	.01*
Memory for Design	.34	.51	.48	.22	.45	.23
Silograms	.23	.56*	.32*	.17*	.09	.36
Number Memory	-.01	.59*	.16*	-.22*	-.01	-.01
Observation	.18	.58*	.22*	.14*	.09	.29
Word Association	-.04	.03*	.18*	-.26*	-.09	.03
Eye and Hand						
Eye	—[d]	.08	.01	—[d]	.00[c]	.01
Hand	.03	.05	.00[c]	.04	-.02	.11

Note. With the exception of Word Association and Eye and Hand, practice effects are based on percentile scores that were converted to standard scores, or z-scores. For Word Association, practice effects are based on certile, rather than percentile, scores that were converted to z-scores, with high scores indicating Objectivity and low scores indicating Subjectivity. For Eye and Hand, practice effects are based on raw scores that were converted to ratios, so that a ratio of 1 indicates a completely right-eyed or right-handed person and a ratio of 0 a left-eyed or left-handed individual.

[a] Primary samples only (excluding examinees with intervals of less than one year).
[b] Interval extends up to, but does not include, 6 years.
[c] The value is less than ±.005 and therefore is rounded to .00.
[d] Because there were no examinees in the 6-years-and-over group, the overall group is the same as the 1-year-to-6-years interval group.

*p < .05 for comparisons of practice effects across the three intervals (ANOVAs) or between the two age groups (t-tests)

Table 8

Practice Effects for Aptitude Tests (Females)

		Test-retest interval			Age at first testing[a]	
Test	Overall[a]	Under 1 year	1 yr.- 6 yrs.[b]	6 yrs. & over	14-19	20+
Number Checking	.07	.30*	.21*	-.09*	.14	.01
Ideaphoria	.05	.20*	.13*	-.04*	.03	.07
Inductive Reasoning	.26	.59*	.26*	.25*	.25	.26
Analytical Reasoning	.45	.36	.43	.50	.57*	.30*
Wiggly Block	.20	.48	.27	.13	.34*	.04*
Memory for Design	.17	.47*	.34*	.01*	.17	.17
Silograms	.27	.49	.36	.20	.30	.24
Number Memory	-.07	.33*	-.02*	-.12*	.04	-.19
Observation	.34	.41	.53	.24	.38	.30
Word Association	.05	.05	.03	.08	.14	-.03
Eye and Hand						
Eye	—[c]	.04	-.01	—[c]	.10	-.07
Hand	-.03	.03	.00[d]	-.06	.05*	-.09*

Note. With the exception of Word Association and Eye and Hand, practice effects are based on percentile scores that were converted to standard scores, or z-scores. For Word Association, practice effects are based on certile, rather than percentile, scores that were converted to z-scores, with high scores indicating Objectivity and low scores indicating Subjectivity. For Eye and Hand, practice effects are based on raw scores that were converted to ratios, so that a ratio of 1 indicates a completely right-eyed or right-handed person and a ratio of 0 a left-eyed or left-handed individual.

[a] Primary samples only (excluding examinees with intervals of less than one year).
[b] Interval extends up to, but does not include, 6 years.
[c] Because there were no examinees in the 6-years-and-over group, the overall group is the same as the 1-year-to-6-years interval group.
[d] The value is less than ±.005 and therefore is rounded to .00.

*$p < .05$ for comparisons of practice effects across the three intervals (ANOVAs) or between the two age groups (t-tests)

Table 9

Practice Effects Based on Form Taken

		Test-retest interval			Age at first testing[a]	
Test	Overall[a]	Under 1 year	1 yr.- 6 yrs.[b]	6 yrs. & over	14-19	20+
Number Checking, same form (380 → 380; 703 → 703)						
Practice effect	.17	.20	.21	.03	.17	.16
n	206	91	160	46	108	98
Number Checking, different form (380 → 613, 703)						
Practice effect	.11	–	–	.18	.24	-.17
n	24	0	2	22	16	8
Ideaphoria, same form (A → A; C → C)						
Practice effect	.05	.17	.15	-.07	.06	.04
n	347	124	190	157	148	199
Ideaphoria, different form (A → C; C → A)						
Practice effect	.05	.18	.06	.03	.08	.02
n	438	180	270	168	192	246
Number Memory, same form (A → A; B → B)						
Practice effect	-.03	.45	.08	-.17	.03	-.09
n	166	56	95	71	89	77
Number Memory, different form (A → B; B → A)						
Practice effect	-.13	.46	-.08	-.16	-.08	-.17
n	19	6	8	11	9	10

Note. Mean test scores are based on percentile scores that were converted to standard scores, or z-scores. None of the comparisons of practice effects between same-form and different-form retests (*t*-tests) was significant.

[a] Primary samples only (excluding examinees with intervals of less than one year).
[b] Interval extends up to, but does not include, 6 years.

Table 10

Stability Coefficients for Aptitude Tests

		Test-retest interval			Age at first testing[a]		
Test	Overall[a]	Under 1 year	1 yr.- 6 yrs.[b]	6 yrs. & over	14-19	20+	Disatten. coef.[c]
Number Checking							
Stability coef.	.76	.85	.79	.74	.74	.78	.89
n	314	92	165	149	166	148	
Ideaphoria							
Stability coef.	.62	.71*	.71*	.52*	.53*	.67*	.87
n	880	304	466	414	405	475	
Inductive Reasoning							
Stability coef.	.64	.67*	.71*	.54*	.58*	.70*	.96
n	473	147	275	198	256	217	
Analytical Reasoning							
Stability coef.	.63	.65	.60	.69	.65	.61	.97
n	235	46	136	99	127	108	
Wiggly Block							
Stability coef.	.65	.82*	.72*	.62*	.61	.70	.79
n	254	71	120	134	139	115	
Memory for Design							
Stability coef.	.73	.77	.74	.73	.74	.71	.95
n	185	63	87	98	88	97	

(table continues)

		Test-retest interval			Age at first testing[a]		
Test	Overall[a]	Under 1 year	1 yr.- 6 yrs.[b]	6 yrs. & over	14-19	20+	Disatten. coef.[c]
Silograms							
Stability coef.	.73	.81	.78	.68	.77	.69	.90
n	209	59	85	124	104	105	
Number Memory							
Stability coef.	.69	.73	.72	.67	.67	.72	.95
n	187	62	103	84	100	87	
Observation							
Stability coef.	.62	.56	.61	.63	.63	.61	1.00
n	231	70	94	137	125	106	(calculated value = 1.11)
Word Association							
Stability coef.	.63	.81*	.69*	.59*	.54*	.74*	.78
n	428	94	232	196	229	199	

Note. With the exception of Word Association, stability coefficients are based on percentile scores that were converted to standard scores, or z-scores. For Word Association, stability coefficients are based on certile, rather than percentile, scores that were converted to z-scores, with high scores indicating Objectivity and low scores indicating Subjectivity.

[a] Primary samples only (excluding examinees with intervals of less than one year).
[b] Interval extends up to, but does not include, 6 years.
[c] The disattenuated coefficient is the overall long-term stability coefficient (first column) divided by the short-term stability coefficient (second column).

*There are significant differences in the coefficients in this group (the three test-retest intervals or the two age-at-first-testing groups), $p < .05$.

with the highest correlations observed for intervals of less than one year and the lowest correlations for intervals of six or more years. Nevertheless, the lowest correlation for the samples with a retest interval of at least six years is .52. For example, for Number Checking, the stability coefficients are .85 for retest intervals of less than one year, .79 for intervals of one year up to six years, and .74 for intervals of six years or more. Differences across the three interval groups were statistically significant for Ideaphoria, Inductive Reasoning, Wiggly Block, and Word Association.

The stability coefficients were also affected by the age of the examinees when initially tested. The stabilities generally tended to be higher for examinees originally tested at age 20 or older than for younger examinees. For Number Checking, for example, the stability coefficient was .74 when age at first testing was between 14 and 19 years old and slightly higher, .78, when age at initial testing was 20 years or older. Differences between the two age-at-first-testing groups, however, were statistically significant for only three tests--Ideaphoria, Inductive Reasoning, and Word Association.

The stabilities for males and females separately were very similar for the primary samples. The lone exception was Analytical Reasoning, on which females showed significantly higher stability than males: overall $r_{\text{females}} = .71$ ($n = 124$), overall $r_{\text{males}} = .53$ ($n = 111$); $z = 2.25, p < .05$.

Not surprisingly, test stabilities are higher for retest intervals of less than one year than for intervals of one year or more. (The one exception was for Observation, where the long-term correlation was actually higher than the short-term correlation, presumably due to sampling error.) Changes in test stabilities over time (i.e., differences between long-term and short-term stability coefficients) were relatively small, however. For instance, for Number Checking, the difference between the short-term correlation of .85 and the long-term correlation of .76 was .09, which means that stability did decline with time, although not by very much. In brief, for the 11 aptitude tests in this study, changes in stabilities over time ranged from .02 for Analytical Reasoning to .18 for Word Association, with the median difference being .06. (Again, the one exception was Observation, for which a negative value was obtained because of sampling error.) The tests with the greatest decline in stability over time were Wiggly Block and Word Association; the tests with the least decline in stability over time were Inductive Reasoning, Analytical Reasoning, Memory for Design, and Number Memory.

The disattenuated coefficients also are displayed in Table 10 (see the right-hand column). As can be seen, most of the true-score variance on these tests is stable, with values ranging from .78 for Word Association to 1.00 for Observation (literally, 1.11, given sampling error). With regard to Number Checking, for example, it has already been pointed out that the long-term coefficient is not very different from the short-term coefficient, and this is reflected in the disattenuated coefficient of .89 for the test, indicating that 89% of the true-score variance was stable between administrations over the time period. The disattenuated coefficients demonstrate strong stability in the underlying abilities here, although there is also a nontrivial amount of short-term instability in test scores. But even though the variance that is due to short-term fluctuation in scores is larger than we might prefer in some cases, it is still a fairly small proportion of the variation for most of the tests. By far the greatest part of the true-score variance in the traits is stable.

For Number Checking, Ideaphoria, and Number Memory, we also looked at stability coefficients based on form taken. There were no significant differences between overall test-retest correlations for same-form and different-form retests. A comparison of same-form and different-form stability coefficients by interval group revealed one significant difference. Not unexpectedly, the short-term stability coefficient for Ideaphoria was significantly higher for examinees who were given the same topic at successive administrations than for examinees who received a different topic at their retesting: $r_{\text{same form}} = .78$ ($n = 124$), $r_{\text{different form}} = .64$ ($n = 180$); $z = 2.36$, $p < .05$. None of the comparisons by age group was significant. In addition, we compared stability coefficients for certain forms that were taken twice by examinees. Specifically, for Number Checking, we looked at the stabilities obtained when Worksample 380 was taken twice versus when Worksample 703 was taken twice; for Ideaphoria, we compared the stabilities of Form A versus Form C; and for Number Memory, we examined the stabilities of Form A versus Form B. None of the comparisons between these same-form retests was significant, although it should be noted that sample sizes were too small for analyses in three cases: for a test-retest interval under one year for Number Checking, Worksample 380, and for Number Memory, Form B; and for an interval of six or more years for Number Checking, Worksample 703.

For purposes of illustration, a typical degree of stability for a Foundation aptitude test is presented graphically in Figure 1. This figure depicts in bar-chart form the amount of change in examinees' percentile scores from first to second administration, using the long-term stability data for Inductive Reasoning. As can be seen from Figure 1, examinees' scores are for the most part quite stable over the long term, with approximately 31% changing by 5 or fewer percentile points and about 26% changing by 6 to 15 percentile points. Furthermore, as one would expect, the percent of examinees drops off dramatically as the difference between scores at testing and retesting increases (i.e., the degree of stability decreases). As an additional illustration, we put together Table 11, which presents the stability data for Inductive Reasoning within the framework of a two-way table of percentile scores at first testing by percentile scores at retesting. By referring to the table, one can determine, for example, that of the 62 examinees who scored in the 71-to-80 percentile range at initial testing, 24.2% scored in the same range at retesting, 25.8% scored in the 81-to-90 percentile range, 16.1% in the 91-to-99 range, 11.3% in the 61-to-70 range, and so on. That is to say, the majority of scores at retesting fell in the same or an adjacent percentile range, which typifies the good long-term stability associated with the Foundation's tests.

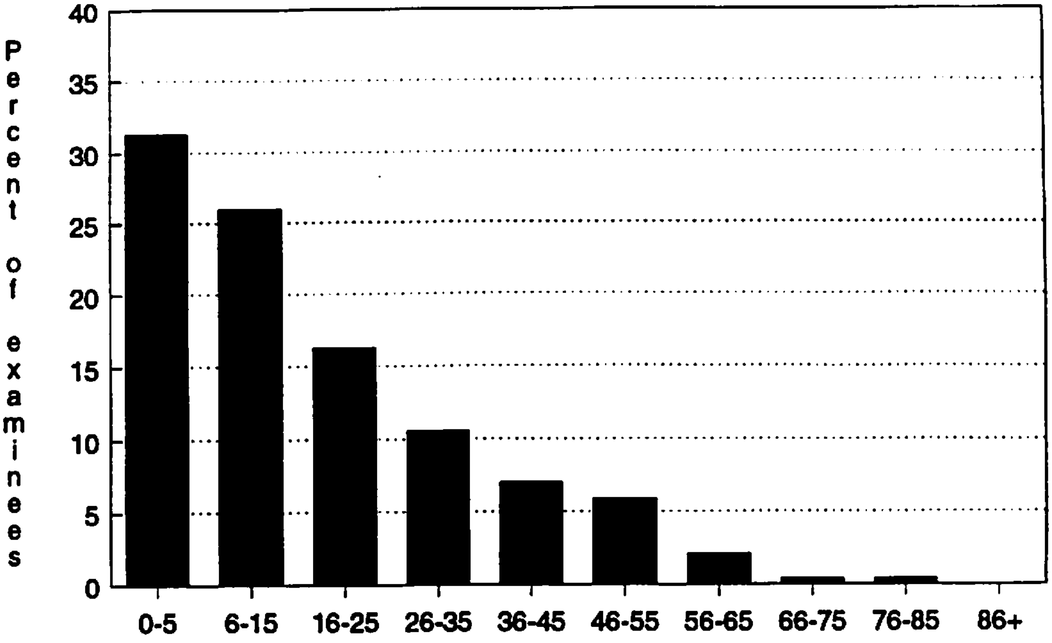
Test-Specific Findings

Some additional comments are in order with regard to our stability findings for individual tests. Some tests, for instance, have stabilities that significantly declined as the retest interval increased, whereas other tests showed very little decline in stability over time; some are significantly more stable for older examinees than for examinees in their teens, whereas others have similar stabilities for both age groups; and one test has shown significantly higher stability for females than for males, whereas the other tests have similar stabilities for the two sexes. For some tests, almost all of the true-score variance between test administrations is stable, so that there is little change in test performance from one occasion to another beyond any short-term fluctuations; for other tests, a smaller proportion of true-score variance is stable so that a correspondingly greater proportion of true-score variance is associated with long-term change. Needless to say, in the process of reviewing the stabilities for each Foundation test, we observed a striking variety of stability patterns among these aptitude measures.

Number Checking proved to be the aptitude test with the best stabilities all-around, with not only the highest long-term stability of the 10 tests but also the highest short-term stability,

Figure 1

Long-Term Stability of Inductive Reasoning



Percentile-point difference in scores from first testing to retesting

Table 11

Inductive Reasoning Stability: Distribution of Scores at First Testing Across the Range of Retest Scores

Percentile score	Percentile score at retesting										
at first testing	1-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80	81-90	91-99	*n*
1-10	45.5	13.6	15.9	9.1	6.8	4.5	0	0	4.5	0	44
11-20	11.8	17.6	5.9	20.6	11.8	5.9	5.9	17.6	2.9	0	34
21-30	6.1	9.1	24.2	9.1	18.2	12.1	9.1	3.0	9.1	0	33
31-40	2.5	12.5	22.5	12.5	2.5	10.0	12.5	7.5	15.0	2.5	40
41-50	5.7	5.7	9.4	13.2	13.2	15.1	5.7	7.5	11.3	13.2	53
51-60	0	4.9	1.6	13.1	11.5	8.2	14.8	13.1	18.0	14.8	61
61-70	0	1.9	7.4	7.4	7.4	13.0	14.8	20.4	20.4	7.4	54
71-80	0	3.2	4.8	4.8	1.6	8.1	11.3	24.2	25.8	16.1	62
81-90	0	0	3.6	1.8	3.6	3.6	5.4	17.9	32.1	32.1	56
91-99	0	0	0	0	2.8	0	5.6	8.3	27.8	55.6	36

Note. Table values are row percents; primary sample only (excluding examinees with intervals of less than one year), *N* = 473.

1-to-6-year stability, and 6-year-and-over stability. In addition, the test has the second highest stability for 14-to-19-year-olds and the highest for 20-year-olds and older. There were no significant differences in stabilities across interval groups or between age groups. Moreover, there were no significant differences between same-form and different-form stabilities. Based on a disattenuated coefficient of .89, we can say that there is little instability (or change) in the underlying ability beyond any short-term fluctuations.

The next best long-term stabilities belong to three of the Foundation's memory tests: Memory for Design, Silograms, and Number Memory. Furthermore, these three tests are stable in the short-run, especially Silograms, one of four Foundation tests with a short-term stability coefficient above .80. These memory tests also have good stabilities for both the 1-to-6-year and the 6-year-and-over intervals and for both younger and older examinees; indeed, Silograms' stability for 14-to-19-year-olds was the highest of all 10 tests. None of these three memory tests have stabilities that were significantly different across interval groups or between age groups. All three have disattenuated coefficients in the .90s, indicating very strong stability in these aptitudes.

By way of contrast, Observation, the fourth memory test in the Foundation battery and the only one of the four that is individually administered, has the lowest short-term stability of the 10 tests and ties (with Ideaphoria) for lowest long-term stability. Observation has the further distinction of being the only test with a short-term stability coefficient that is (a) in the .50s (the next lowest is .65) and (b) lower than its long-term stability coefficient, presumably due to sampling error. In addition, this test has the second lowest stability for the 1-to-6-year retest interval. Surprisingly, stability tends to increase rather than decrease across the intervals, although not significantly. There were no age differences in stabilities. For Observation, a preponderance of the true-score variance that is stable in the short term continues to be stable in the long term, so that very little true-score variance is associated with change; nevertheless, Observation also has a considerable amount of measured variance that is error.

Ideaphoria, which ties Observation for lowest overall stability, also has the lowest stability for the retest interval encompassing six or more years. The test's (identical) short-term and 1-to-6-year stabilities, though, fall in the middle range for Foundation tests and are significantly higher than this 6-year-and-over stability. At the same time, Ideaphoria is the least stable of the 10 tests for examinees in their teens; in fact, it is one of only three tests with significantly lower stability for younger examinees than for older examinees. One difference between same-form and different-form stabilities was observed: Not surprisingly, the short-term stability coefficient is significantly higher for examinees who were given the same topic at both administrations. Ideaphoria's disattenuated coefficient of .87 reflects little true-score change in the underlying ability beyond any short-term fluctuations.

With regard to the Foundation's reasoning tests, Inductive Reasoning and Analytical Reasoning have similar overall stability coefficients as well as short-term coefficients. Moreover, both tests have short-term stabilities that are only slightly higher than their long-term stabilities, so that their disattenuated coefficients are two of the highest of the 10 Foundation tests. In other words, they are substantially stable for the long term with very little true-score variance associated with true change. Nonetheless, in both cases, about one-third of their measured variance is associated with short-term fluctuations in performance that are unrelated to true change. These two tests have distinct patterns of stabilities across intervals and between age groups, however. Inductive Reasoning's pattern of stabilities is similar to Ideaphoria's in that (a) its stability for the 6-years-and-over interval is significantly lower than for the under-1-year and 1-to-6-year intervals, (b) it has one of the lowest stability coefficients for examinees in their teens, and (c) this coefficient for teens is significantly lower than the coefficient for examinees in their 20s and older. By way of contrast, Analytical Reasoning's stability coefficient for the 6-years-and-over retest interval is one of the highest of the 10 tests, whereas its stability coefficient for the 1-to-6-year interval is the lowest, although neither the stabilities across intervals nor between age groups are significantly different. Analytical Reasoning is the only Foundation test for which the overall stabilities of males and females differed significantly, with females obtaining higher test-retest correlations than males.

With respect to Wiggly Block, this test ranks second best in terms of short-term stability. It is significantly more stable for shorter retest intervals, but there is no difference between age groups. Of the 10 Foundation tests, Wiggly Block has the second-poorest long-term stability relative to its short-term stability; still, 79% of its true-score variance was stable in the long run, although some long-term change also occurred.

Word Association's short-term stability is also one of the highest, but its stabilities for both the 1-to-6-year and 6-year-and-over intervals are among the lowest of the Foundation tests. Furthermore, the test is significantly more stable for shorter intervals than for longer time periods and significantly less stable for 14-to-19-year-olds than for examinees aged 20 and older. Even though it proved to be the Foundation test with the lowest disattenuated coefficient, Word Association, like Wiggly Block, demonstrated reasonably good stability in the underlying ability, although there is also a nontrivial amount of short-term instability in its scores.

Stability of Laterality Measures

As noted in the Analyses section, summary stability indices rather than test-retest correlations were used as measures of stability for eye and hand dominance. To reiterate, these stability indices were the percents of examinees who remained stable in terms of their eyedness and handedness categories, respectively. The stability indices for Eye and Hand are shown in Table 12.

For Eyedness, 91.7% of the primary sample remained stable compared with 94.3% of the short-term sample. Stability indices based on age at first testing likewise were similar: 90.0% for those originally tested between the ages of 14 and 19 and 92.6% for those first tested at age 20 or older.

Separate indices for completely left-eyed and completely right-eyed examinees also are displayed in Table 12. There were no significant differences in stability indices between the long-term and short-term samples or between the two age groups for either the left-eyed or right-eyed groups. A comparison between the stability indices of left-eyed and right-eyed examinees revealed one significant difference. For a retest interval of less than one year, stability was significantly higher for right-eyed examinees than for left-eyed examinees, $z = 2.13, p < .05$.

For Handedness, 88.9% of the primary sample's scores remained stable. The stabilities for handedness tended to decline with the length of the interval, although not by very much. The stability indices based on age at first testing were similar to the overall stability: 88.1% for those originally tested between the ages of 14 and 19 and 89.7% for those first tested at age 20 or older.

Table 12

Stability Indices for Eyedness and Handedness[a]

		Test-retest interval			Age at first testing[b]	
Category	Overall[b]	Under 1 year	1 yr.- 6 yrs.[c]	6 yrs. & over	14-19	20+
All eyedness categories						
n	84	87	84		30	54
Percent stable	91.7	94.3	91.7	–	90.0	92.6
Completely left-eyed at first testing						
n	28	31	28		12	16
Percent stable	89.3	87.1*	89.3	–	83.3	93.8
Completely right-eyed at first testing						
n	55	56	55		18	37
Percent stable	94.5	98.2*	94.5	–	94.4	94.6
All handedness categories						
n	225	87	110	115	109	116
Percent stable	88.9	92.0	89.1	88.7	88.1	89.7
Essentially left-handed at first testing						
n	11	3	6	5	7	4
Percent stable	90.9	100.0	100.0	80.0	100.0	75.0
Essentially right-handed at first testing						
n	185	71	91	94	86	99
Percent stable	94.1	97.2	94.5	93.6	94.2	93.9

[a] *Stability indices* were the percents of examinees who remained stable in terms of their eyedness and handedness categories, respectively. Ratio scores for eye dominance were partitioned into three eyedness categories: (a) completely left-eyed if the ratio was 0, (b) variable-eyed if the ratio was between 0 and 1, or (c) completely right-eyed if the ratio was 1. Ratio scores for hand dominance were partitioned into five handedness categories: (a) essentially left-handed (for a ratio of 0 to .10), (b) primarily left-handed (for a ratio of .11 to .30), (c) variable-handed (for a ratio of .31 to .69), (d) primarily right-handed (for a ratio of .70 to .89), or (e) essentially right-handed (for a ratio of .90 to 1).

[b] Primary samples only (excluding examinees with intervals of less than one year).

[c] Interval extends up to, but does not include, 6 years.

*There are significant differences in the stability indices for left-sidedness and right-sidedness for this group, $p < .05$.

None of the separate indices for essentially left-handed or essentially right-handed at first testing were significantly different between the long-term and short-term samples or between the two age groups. Likewise, none of the comparisons between right-handed examinees and left-handed examinees were significant, although it should be noted that the sample of left-handed examinees was very small.

We were also interested in looking at the degree of instability for the examinees with unstable eyedness or handedness scores ("switchers"). For instance, three of the 12 examinees classified as unstable for eyedness were not very unstable: one examinee used his left eye for one trial and his right eye for eight trials at first administration and his right eye for all four trials at second administration; and two examinees' scores for each eye changed by only 1 point (e.g., at first administration, they used their left eyes for all four trials of Part 1, and at second administration, they used their left eyes for three trials and their right eyes for one trial). Of the remaining nine examinees with unstable eyedness scores, three examinees' scores changed by 3 points, and six examinees' scores changed by 4 points (i.e., they used one eye for all trials at first administration and the other eye for all trials at second administration).

We further examined the magnitude of instability by comparing the percentage of examinees from each category who were switchers to determine if certain categories (e.g., right-eyedness or right-handedness at first testing) tended to be more stable than other categories (e.g., left-eyedness or variable-handedness at first testing). In the case of eyedness, the percentage of examinees who originally were left-eyed and switched to the "completely right-eyed" category was not significantly different from the percentage of right-eyed examinees who switched to the "completely left-eyed" category (5.1% [*n* = 3] vs. 2.7% [*n* = 3], respectively). There was, however, a trend for a greater percentage of switchers in the primary sample than in the short-term sample to switch from using one eye for all four trials to using the other eye for all four trials (71.4% [*n* = 5] vs. 20.0% [*n* = 1], respectively), although this should be interpreted with caution because of the small samples sizes involved.

With respect to handedness, many of the 32 cases classified as unstable were not especially unstable. For instance, seven examinees classified as unstable performed the same number of activities with their dominant hand at both testings but performed one or two additional activities with their nondominant hand at one administration relative to the other. Another ten examinees classified as unstable had scores for each hand that changed by only 1 point from first to second testing (e.g., at first administration, they used their left hands for two activities and their right hands for ten activities, and at second administration, they used their left hands for one activity and their right hands for 11 activities); ten had scores that changed by 2 points, and four had scores that changed by 3 points. Only one case was extremely unstable--an examinee who used his left hand exclusively when first tested at age 21 and his right hand for all but two activities when retested 23 years later (although his right hand was used for writing on both occasions).

The percentage of switchers from each handedness group varied considerably from category to category, with the "essentially right-handed" and "essentially left-handed" categories showing significantly more stability than the other three handedness categories. Specifically, only 5.1% of examinees initially considered "essentially right-handed" and 7.1% of "essentially left-handed" examinees moved to another category at second testing. In contrast to this 5.1% from the "essentially right-handed" group (*n* = 13), the percentages of switchers from the "variable-handed," "primarily left-handed," and "primarily right-handed" categories were significantly higher: 26.7% of "variable-handed" examinees switched ($n = 4$), $z = 3.33$, $p < .01$; 40.0% of "primarily left-handed" examinees ($n = 4$), $z = 4.43$, $p < .01$; and 58.8% of "primarily right-handed" examinees ($n = 10$), $z = 7.75$, $p < .01$. All but three of the handedness switchers moved to an adjacent category at second administration, with 22 out of 29 of the one-category switchers moving from "essentially right-handed" to "primarily right-handed" or vice-versa. (Of the remaining seven one-category switchers, three moved from "primarily left-handed" to "essentially left-handed," two moved from "primarily left-handed" or "primarily right-handed" to "variable-handed," and two moved from "variable-handed" to "primarily left-handed" or "primarily right-handed.") With regard to the three examinees who moved at least two categories, two examinees switched from "variable-handed" to "essentially left-handed" or "essentially right-handed," and the other examinee moved from "essentially left-handed" to "primarily right-handed."

There were four additional questions regarding laterality that we wished to consider in this study:

1. Do changes in one modality (i.e., eyedness or handedness) show corresponding changes in the other modality?

2. Do changes in handedness tend to become congruent with eyedness and not vice versa?

3. Are the findings for writing hand comparable to the findings for handedness?

4. Do individuals who are cross-dominant at initial testing tend to remain cross-dominant over time, or do they tend to move closer to their dominant eye or their dominant hand?

Unfortunately, the number of examinees classified as unstable for one modality for whom we also had valid data for the other modality was small ($n = 12$ for eyedness switchers and 14 for handedness switchers). Nevertheless, with regard to the first question, the answer appears to be "No," with only 25.0% of the eyedness switchers also showing changes in their handedness category and 21.4% of the handedness switchers also showing changes in their eyedness category.

With respect to the second question, the percentage of handedness changes that became congruent with eyedness at retesting was not significantly different from the percentage of eyedness changes that became congruent with handedness (54.6% [$n = 6$] vs. 62.5% [$n = 5$], respectively), although this finding should be interpreted with caution because of the small samples sizes involved. Of interest, too, are the comparisons between right- and left-stable individuals. For example, all of the eye switchers who were right-hand stable had eyedness scores that became more congruent with right-sidedness at retesting ($n = 4$), but only 25.0% who were left-hand stable had eyedness scores that became more congruent with left-sidedness ($n = 1$), $z = 2.19$, $p < .05$. A similar trend emerged for hand switchers, with 80.0% of hand switchers who were right-eye stable having handedness scores that became more congruent with right-sidedness at retesting ($n = 4$) and 33.3% who were left-eye stable having handedness scores that became more congruent with left-sidedness ($n = 2$). In all, 88.9% of switchers who were right-stable for one modality became more congruent with the stable modality, or right-sidedness, whereas 30.0% of switchers who were left-stable for one modality became more congruent with the stable modality, or left-sidedness. That is to say, switchers of either modality tended to move closer to right-sidedness at retesting regardless of whether their stable modality was left- or right-dominant. There were also three individuals who were both eyedness and handedness switchers. At retesting, the eyedness scores of all three became more congruent with their handedness at first administration, and their handedness scores became more congruent with their eyedness at original testing.

As for writing hand, 99.6% of the primary sample remained stable compared with 98.9% of the short-term sample. The stability indices based on age at first testing were similar to the overall stability and identical to each other--99.1% both for those originally tested between the ages of 14 and 19 and for those first tested at age 20 or older. In fact, only two examinees in the sample switched writing hands at retesting; and, in both cases, they switched between their right hand and ambidexterity (rather than between their right and left hands). Of these two with unstable writing hand, one examinee, who was tested twice at age 28, switched from using his right hand for writing to being ambidextrous but remained left-eyed and right-handed. The other examinee, who was tested at ages 17 and 19, switched from being ambidextrous to using her right hand for writing and remained left-eyed and variable-handed, although her handedness score moved slightly in the direction of being more right-handed at retesting. As to whether changes in handedness or eyedness tend to become congruent with writing hand, 75% of eyedness changes and 57% of handedness changes became more congruent with writing hand at retesting.

Regarding cross-dominance, at the first administration of the Eye and Hand test, 29.8% of the primary sample and 28.7% of the short-term sample were found to be cross-dominant, of which approximately 92% of each group were left-eyed and right-handed and the remaining 8% right-eyed and left-handed. The stability indices for cross-dominance were 80.0% for the primary sample and 84.0% for the short-term sample. Two-thirds of cross-dominant examinees initially tested between the ages of 14 and 19 remained cross-dominant at second testing, and 92.3% of examinees 20 and over remained cross-dominant. Of the nine cross-dominant individuals with unstable eyedness or handedness scores, five switched eyedness categories at retesting to become more congruent with handedness, two switched handedness categories to become more congruent with eyedness, and two changed both eye and hand categories. Of the eyedness-only switchers, one went from "completely right-eyed" to "completely left-eyed," one went from "completely left-eyed" to "completely right-eyed," and three moved from "completely left-eyed" to "variable-eyed." Both handedness-only switchers moved from the "essentially right-handed" group to the "primarily right-handed" group. As to the two examinees who were both eyedness and handedness switchers, one went from "completely left-eyed" to "completely right-eyed," one went from "completely left-eyed" to "variable-eyed," and both moved from the "essentially right-handed" category to the "primarily right-handed" category.

DISCUSSION

The primary purpose of this study was to assess the long-term stability of 11 aptitude tests in the Foundation's standard battery. Of particular interest were the overall levels of stability for nine specific cognitive ability measures. Our results clearly indicate that cognitive abilities of the type measured by the Foundation's aptitude test battery are largely stable over periods of six years and more. The disattenuated coefficients in this study were very high. There also appears to be a nontrivial amount of short-term instability in test scores, with Inductive and Analytical Reasoning showing short-term coefficients of less than .70. Nevertheless, there appears to be little additional change over long periods of time.

Some additional comments are in order with regard to our research findings. We begin this section with a comparison of our findings to those from the other major longitudinal studies that included measures of specific abilities. This is followed by discussions of the limitations of our study and implications of the current findings.

Comparison of Findings With Other Longitudinal Studies of the Stability of Specific Cognitive Abilities

The results from our study can be compared with the results of studies by the Ball Foundation (Dawis et al., 1992) and Schaie (1985). As mentioned in the Introduction, Dawis et al. tested 121 students at ages 17-18 and 21-22 on a set of tests quite similar to the Johnson O'Connor Research Foundation's tests. As in our study, they found considerable stability, with coefficients very similar to ours for Clerical (Number Checking) and Ideaphoria; somewhat lower coefficients (by an average of .05 to .07) for Analytical Reasoning, Shape Assembly (Wiggly Block), and Word Association; and a much lower coefficient for Inductive Reasoning.[7]

Schaie (1985) found even stronger results for adults who were tested at 7-year intervals on five tests from the Schaie-Thurstone Adult Mental Abilities Test. Overall, all five ability tests showed substantial long-term stability coefficients that ranged from .68 to .88 for intervals of 7, 14, and 21 years. Compared to our study, the stability coefficients obtained by Schaie were higher: .70 to .78 for Word Fluency compared with .52 to .67 for Ideaphoria; .81 to .86 for Letter Series compared with .54 to .70 for Inductive Reasoning; and .68 to .81 for Figure Rotation compared with .62 to .70 for Wiggly Block.[8] None of the Schaie-Thurstone and JOCRF tests are completely comparable, however. Word Fluency and Ideaphoria, for instance, are measures of distinct abilities within the domain of idea production, namely word fluency and ideational fluency, respectively. Figure Rotation is a paper-and-pencil, multiple-choice test of spatial ability that uses two-dimensional stimuli and response alternatives; in contrast, Wiggly Block is a performance, or assembly, test that utilizes three-dimensional pieces. Letter Series and the Foundation's Inductive Reasoning test, while both measures of inductive reasoning, involve very different types of inductive tasks.

[7] For purposes of comparison, in addition to using our overall long-term stability coefficients, we calculated stabilities for (a) all examinees with a retest interval of four to eight years and (b) examinees originally tested at 18 years of age or older with a retest interval of four or more years. These stabilities were very similar to our overall long-term stabilities. More-direct comparisons could not be made because our sample sizes needed to be over 100 to yield sufficiently precise stability coefficients.

[8] In addition to using our overall long-term stability coefficients, for comparison purposes we calculated stabilities for (a) all examinees with a retest interval of at least six years and (b) examinees 20 years of age or older with a retest interval of at least one year. These stabilities were very similar to our overall long-term stability coefficients. Again, more-direct comparisons could not be made because our sample sizes needed to be over 100 in order for us to be confident of our numbers.

The results of Schaie's and the Ball Foundation's research have contributed greatly to the understanding researchers now have regarding the enduring stabilities of cognitive abilities during the adult years. Our findings, which are consistent with these previous longitudinal studies, also make a valuable contribution to the modest, but growing, body of literature documenting the stability of specific cognitive abilities. First, the results of our investigation provide additional confirmatory evidence that cognitive abilities, like general intelligence, can be considered highly stable in adulthood. Second, our findings extend the current knowledge-base regarding the stability of specific abilities in several ways. To wit, because a broader range of measures was used in our study than in the Schaie and Dawis et al. studies, we have been able to identify additional cognitive abilities that are stable for the long term. We are referring specifically to the four distinct abilities within the memory domain that are measured by the JOCRF battery of tests (i.e., Memory for Design, Silograms, Number Memory, and Observation)--abilities for which we obtained some of our highest long-term stability coefficients, excepting Observation. Furthermore, we have extended the previous findings of Dawis et al. for the six Ball Foundation measures that resemble JOCRF measures because we based our research on samples that encompassed not only a wider range of test-retest intervals (rather than just four years) but also a more-extensive range of ages at original testing and retesting (rather than just young adults). We thereby have been able to document stabilities for examinees with intervals of six and more years as well as for those with initial testing ages of 20 years and older.

Limitations of the Study

Despite its contributions to research on the stability of specific cognitive abilities, the present study is not without limitations, the most salient of which pertain to issues of sample-selection that may qualify the findings. One limitation of the study is the relative homogeneity of our samples of Foundation clients with respect to education, ethnicity, and economic background, as noted previously in the Method section. Because our stability coefficients derive from samples in which members of ethnic minorities and individuals from the lower socioeconomic levels are underrepresented, there is a restriction of range relative to the general population. For this reason, the stability coefficients we report may be slightly lower than in the population as a whole. In addition, this restriction of range may limit the generalizability of the findings so that some caution should be used in applying the study findings beyond populations like the present samples. For example, persons who do not work in cognitively complex professional jobs may show greater decline in abilities during adulthood than persons who do (Schaie, 1996). This would lead to lower stability coefficients.

Another limitation pertains to potentially influential differences within the Foundation population itself. The reader should remember that the retest data used in this study were collected exclusively from Foundation clients who returned (at their own initiative) for a "follow-up" discussion of their test results. This raises the question of whether there may be some differences between Foundation clients who request follow-up appointments and those who do not that would have an impact on the results. One point to consider is that individuals tested after college may be underrepresented in our samples of follow-up clients, given the fact that age at first testing for the long-term samples was lower than the median for the general Foundation population. Even if true, however, we believe the plausible impact of level of education on test-score stability would likely be negligible. More potentially troublesome are the possibilities that (a) persons who change a great deal are less likely to request follow-ups and (b) persons whose aptitudes were not measured very well, for whatever reason, may not come back for follow-ups.

Because neither of these considerations has been addressed by Foundation research, we cannot completely rule them out. Nonetheless, although we acknowledge that they may be possible, we have no evidence to suggest either is probable or that they are likely to materially affect our results.

Implications of Our Findings

With the above limitations in mind, our findings regarding the stability of aptitudes have several important implications for the use of cognitive ability test batteries, most notably with respect to long-term planning and goal-setting, particularly in educational and occupational guidance situations (Lowman, 1991; Trembly, 1992). We also think it worthwhile to note that aptitude stability plays a substantive role in supporting a person's self-concept as well as increasing that person's understanding of himself and others. In addition, there are some implications specific to the Foundation and its testing program that will be mentioned here.

Aptitude testing is the cornerstone of many educational and vocational guidance programs, including the Foundation's. Of course, the effectiveness of these testing programs is contingent on the enduring nature of the aptitudes measured, for without aptitude stability, the information obtained from cognitive ability testing would be useful only for a limited period of time. It is only because we are fairly certain the pattern of one's aptitudes will remain consistent over time that we can have our aptitudes measured only once, and on that basis, make reasonable decisions regarding our future and confidently set long-range educational and occupational goals that will make effective use of our natural abilities. Thus, our confirmation of the tendency for aptitudes to remain stable during adulthood provides a strong foundation for applications of cognitive ability tests in educational and vocational guidance.

Moreover, the results of longitudinal studies on the stability of aptitudes, in conjunction with findings on personality stability (McCrae & Costa, 1990), provide clear evidence that such stability can provide an objective grounding for the individual's self-concept and sense of identity. That is to say, knowledge gained about one's specific abilities coupled with the realization that those aptitudes will characterize a person for years to come can provide a person with an enduring sense of self. This increased awareness of self, or who one is, is possible because the key aspects of one's make-up, that is, the elements most central to one's identity, will stay much the same throughout one's life.

In addition, any knowledge we acquire about the nature of aptitudes and how they affect human behavior enables us to not only understand but also appreciate persons with aptitude patterns different from our own. Such insights can be particularly relevant with regard to family members, for once we are able to recognize our spouses' and children's aptitudes and accept the enduring nature of those aptitudes, we can then progress toward allowing these individuals to develop and fully use their natural abilities rather than trying to force them to be or do something for which they are not well suited. (See Broadley, 1986, chap. 20, for a more-detailed discussion of this topic.)

Our findings from this study have implications specific to the Foundation's testing program as well, namely with respect to the age at which we think aptitudes are stable and therefore a person could be tested. One way of addressing this is to look at the degree to which a person's scores might vary because of long-term change. To this end, we (a) calculated stability coefficients for the various ages of initial testing that, after some smoothing, were used as best estimates of the population stability values and then (b) estimated the amount of change in scores that might occur over time using standard errors (*SE*s) of measurement relative to these population stability values.[9]

At age 22, for instance, the long-term stability coefficient is .68 and the *SE* is .56. From this we can infer that an individual's score on a given test would change approximately .56 *SD*s on retesting after a long intervening time period--e.g., 57 months for the average 22-year-old in our sample. (By way of comparison, over the short term one could expect a change in test scores of .51 *SD*s for 22-year-olds.) At age 16, the amount of change one could expect from one administration of a given test to another over the long run would be .64 *SD*s more or less (compared with .48 *SD*s for a retesting within one year). At ages 14 and 12, a person's score on a given test would change approximately .68 and .75 *SD*s, respectively, with a long-term retest interval. (There were no short-term cases in our files for examinees under the age of 15.)

With regard to testing people in their early to mid-teens, then, the data clearly indicate that stability improves with age. That is to say, the younger the age at first testing, the larger the average amount of change at retesting one can expect, i.e., the less stable the test scores. The *SE*s for tests taken at age 14 or 15 are slightly larger than the *SE*s for tests taken at age 16 or older, and the *SE*s for tests taken at age 12 or 13 are moderately larger than the *SE*s for tests taken at age 14 or 15. By the same token, there is a modest decline in stability when the age of testing is lowered from 16 or older to 14 or 15 and a somewhat larger drop in stability when the age of testing is further lowered to 12 or 13. In short, we found that when testing occurs before the age of 14, aptitudes appear to be too unstable for reliable guidance, whereas testing at age 14 is more-or-less acceptable and testing later in the teens is even better than testing at age 14.

Conclusion

To conclude, this study indicates that the aptitudes we studied here are largely stable from adolescence through adulthood for this population. This confirms the Foundation's view that these abilities are indeed *aptitudes*, that is, "natural talents, special abilities for doing, or learning to do, certain kinds of things easily and quickly" (Johnson O'Connor Research Foundation, 1994). This should not be taken to indicate that these abilities cannot change, but rather that they are not likely to change in the normal course of experience--i.e., in the absence of specific change-inducing experiences. Research by Schaie (1996) and others has indicated that it is possible to modify abilities such as these to a limited degree by specific focused experiences. Schaie and Willis (1996, pp. 401-403) reviewed research on training programs and found that they can lead to gains in memory, reasoning, and spatial ability. Schaie (1996) has also reported research indicating that lack of activity can lead to score declines that exceed normative changes, especially in older adults. In view of these findings, we should qualify our conclusions by saying

[9] Although ideally we would be able to look at the data on a test-by-test basis, there were not enough 14- and 15-year-old examinees in our samples for this type of analyses. Instead, we pooled the data for all the primary samples, with the exception of Eye and Hand, even though we realize that it is not fully appropriate to combine data for various constructs that have distinct stability curves.

that we found stability for most persons in our sample, while we acknowledge that some individuals may have experiences that lead to change. We also cannot rule out the possibility that in other cultures or at other times in history, adults may have had ability-related experiences that altered their patterns of abilities. For the populations to which we address ourselves, however, it appears that stability of cognitive abilities is clearly the rule, and change in abilities is the exception.

REFERENCES

All technical reports, Statistical Bulletins, Test Information Bulletins, and test manuals listed in this reference section are published by the Johnson O'Connor Research Foundation. The Foundation's technical reports can be purchased by contacting our Order Department, Human Engineering Laboratory, 347 Beacon Street, Boston, MA 02116. All test manuals and Test Information Bulletins and some Statistical Bulletins contain confidential information and therefore cannot be made available to the public.

Alwin, D. F. (1994). Aging, personality, and social change: The stability of individual differences over the adult life span. In D. L. Featherman, R. M. Lerner, & M. Perlmutter (Eds.), *Life-span development and behavior* (Vol. 12, pp. 135-185). Hillsdale, NJ: Erlbaum.

Anastasi, A. (1958). *Differential psychology* (3rd ed.). New York: Macmillan.

Anastasi, A. (1976). *Psychological testing* (4th ed.). New York: Macmillan.

Bethscheider, J. K. (1986). *Report to test administrators on their score distributions for worksamples administered in 1984.* Unpublished document. Chicago: Johnson O'Connor Research Foundation.

Bethscheider, J. K. (1994). [Rules for converting scores on Foundation tests]. Unpublished document. Arlington, TX: Johnson O'Connor Research Foundation.

Bloom, B. (1964). *Stability and change in human characteristics.* New York: Wiley.

Botwinick, J. (1977). Intellectual abilities. In J. E. Birren and K. W. Schaie (Eds.), *Handbook of the psychology of aging* (pp. 580-605). New York: Van Nostrand Reinhold.

Broadley, M. E. (1986). *Your natural gifts* (3rd ed.). McLean, VA: EPM Publications.

Brody, N. (1992). *Intelligence* (2nd ed.). New York: Academic Press.

Campbell, D. P. (1965). A cross-sectional and longitudinal study of scholastic abilities over twenty-five years. *Journal of Counseling Psychology, 12,* 55-61.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cunningham, W. R., & Owens, W. A., Jr. (1983). The Iowa State study of the adult development of intellectual abilities. In K. W. Schaie (Ed.), *Longitudinal studies of adult psychological development* (pp. 20-39). New York: Guilford Press.

Daniel, M. H. (1975, July). *Tables of obtained percentile scores and grades on standard worksamples for examinees tested 1972 or earlier.* Unpublished document. Fort Worth, TX: Human Engineering Laboratory.

Dawis, R. V., Goldman, S. H., & Sung, Y. H. (1992). Stability and change in abilities for a sample of young adults. *Educational and Psychological Measurement, 52,* 457-465.

Hays, W. L. (1973). *Statistics for the social sciences* (2nd ed.). New York: Holt, Rinehart and Winston.

Hogan, R., Hogan, J., & Roberts, B. W. (1996). Personality measurement and employment decisions: Questions and answers. *American Psychologist, 51,* 469-477.

Horn, J. L., & Donaldson, G. (1980). Cognitive development in adulthood. In O. G. Brim, Jr., & J. Kagan (Eds.), *Constancy and change in human development* (pp. 445-529). Cambridge, MA: Harvard University Press.

Hoyer, W. J., & Rybash, J. M. (1994). Characterizing adult cognitive development. *Journal of Adult Development, 1,* 7-12.

Johnson O'Connor Research Foundation (1994). [Brochure for prospective clients]. Chicago: Author.

Lowman, R. L. (1991). *The clinical practice of career assessment: Interests, abilities, and personality.* Washington, DC: American Psychological Association.

McCrae, R. R., & Costa, P. T., Jr. (1990). *Personality in adulthood.* New York: Guilford.

McCrae, R. R., & Costa, P. T., Jr. (1994). The stability of personality: Observations and evaluations. *Current Directions in Psychological Science, 3,* 173-175.

Neisser, U., Boodoo, G., Bouchard, T. J., Jr., Boykin, A. W., Brody, N., Ceci, S. J., Halpern, D. F., Loehlin, J. C., Perloff, R., Sternberg, R. J., & Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51,* 77-101.

Nesselroade, J. R., & Baltes, P. B. (1974). Adolescent personality development and historical change: 1970-1972. *Monographs of the Society for Research in Child Development,* 39. (No. 1, Serial No. 154).

Norusis, M. J. (1990a). *SPSS/PC+ 4.0 Base manual.* Chicago: SPSS.

Norusis, M. J. (1990b). *SPSS/PC+ Statistics 4.0.* Chicago: SPSS.

Schaie, K. W. (1983). The Seattle Longitudinal Study: A twenty-one year exploration of psychometric intelligence in adulthood. In K. W. Schaie (Ed.), *Longitudinal studies of adult psychological development* (pp. 64-135). New York: Guilford Press.

Schaie, K. W. (1985). *Schaie-Thurstone Adult Mental Abilities Test* (test manual). Palo Alto, CA: Consulting Psychologists.

Schaie, K. W. (1993). The Seattle Longitudinal Studies of adult intelligence. *Current Directions in Psychological Science, 2,* 171-175.

Schaie, K. W. (1994). The course of adult intellectual development. *American Psychologist, 49,* 304-313.

Schaie, K. W. (1996). *Intellectual development in adulthood: The Seattle Longitudinal Study.* New York: Cambridge University Press.

Schaie, K. W., & Willis, S. L. (1996). *Adult development and aging* (4th ed.). New York: HarperCollins.

Schroeder, D. H., & Nakajima, M. (1997, February). *Age differences for a battery of aptitude tests.* Poster session presented at a conference titled "The New Rules of Measurement," University of Kansas, Lawrence, KS.

Siegler, I. C. (1983). Psychological aspects of the Duke longitudinal studies. In K. W. Schaie (Ed.), *Longitudinal studies of adult psychological development* (pp. 136-190). New York: Guilford Press.

Snow, R. E., Kyllonen, P. C., & Marshalek, B. (1984). The topography of ability and learning correlations. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 2, pp. 47-103). Hillsdale, NJ: Erlbaum.

Statistical Bulletin 1976-20. *Test-retest correlation (stability of scores) for Wks. 35A.* M. H. Daniel. Boston: Human Engineering Laboratory.

Statistical Bulletin 1977-25. *Test-retest stability of Wiggly Block scores.* M. H. Daniel. Boston: Human Engineering Laboratory.

Statistical Bulletin 1979-20. *Test-retest stability of the Number Checking Test, Worksample 380.* M. H. Daniel. Boston: Human Engineering Laboratory.

Statistical Bulletin 1989-5. *Long-term stability of four aptitudes: Graphoria, structural visualization, Inductive Reasoning, and Ideaphoria.* J. K. Bethscheider. Chicago: Johnson O'Connor Research Foundation.

Trembly, D. (1992). *Learning to use your aptitudes* (Rev. ed.). Chicago: Johnson O'Connor Research Foundation.

Willerman, L. (1979). *The psychology of individual and group differences.* San Francisco: W. H. Freeman.

APPENDIX A

Descriptive Statistics for Short-Term Samples[a]

Test	% females	Age at first testing (years)			Test-retest interval (mos.)		
		Median	*SD*	Range	Median	*SD*	Range
Number Checking	41.3	29.5	9.6	15-54	7	3.9	0-11
Ideaphoria	54.0	27.0	8.7	16-55	8	3.7	0-11
Inductive Reasoning	44.9	26.0	8.9	15-56	9	3.4	0-11
Analytical Reasoning	56.5	23.5	7.3	15-46	9	3.6	0-11
Wiggly Block	42.3	28.0	10.1	15-54	7	3.9	0-11
Memory for Design	52.4	28.0	9.5	16-50	9	3.5	0-11
Silograms	47.5	30.0	10.0	15-51	8	4.0	0-11
Number Memory	53.2	29.0	9.1	16-57	9	3.9	0-11
Observation	54.3	28.0	9.3	15-54	9	3.2	1-11
Word Association	58.5	28.5	9.7	15-56	9	3.6	0-11
Eye and Hand	55.2	27.0	9.7	15-57	9	3.9	0-11

[a] Samples consist of examinees with intervals of less than one year.

APPENDIX B

Descriptive Statistics for Test Scores for Short-Term Samples[a]

Test	*n*	Z-score at Time 1		Z-score at Time 2		*t*	Standardized practice effect[b]
		Mean	*SD*	Mean	*SD*		
Number Checking	92	.21	.95	.41	1.06	3.49*	.20
Ideaphoria	304	.30	.88	.48	.87	4.51*	.18
Inductive Reasoning	147	.16	.95	.67	.95	8.10*	.51
Analytical Reasoning	46	.36	.95	.69	.84	2.93*	.33
Wiggly Block	71	-.16	.92	.41	1.07	7.78*	.57
Memory for Design	63	.02	1.04	.51	.98	5.64*	.49
Silograms	59	.37	1.00	.90	1.02	6.57*	.53
Number Memory	62	-.18	.84	.28	.92	5.48*	.46
Observation	70	.17	.84	.66	.86	5.15*	.49
Word Association	94	-.15	.81	-.11	.98	.69	.04
Eye and Hand							
Eye	87	.64	.48	.67	.46	1.68	.06
Hand	87	.89	.26	.90	.25	1.77	.04

[a] Samples consist of examinees with intervals of less than one year. With the exception of Word Association and Eye and Hand, mean test scores are based on percentile scores that were converted to standard scores, or z-scores. For Word Association, mean test scores are based on certile, rather than percentile, scores that were converted to z-scores, with high scores indicating Objectivity and low scores indicating Subjectivity. For Eye and Hand, mean test scores are based on raw scores that were converted to ratios, so that a ratio of 1 indicates a completely right-eyed or right-handed person and a ratio of 0 a left-eyed or left-handed individual.

[b] Effect size was calculated using a population standard deviation of 1, thus making effect size the same as the difference between the z-score means, except for Eye and Hand.

*$p < .05$