Journal of Educational Psychology
1981, Vol. 73, No. 4, 573-586

0022-0663/81/7304-0573$00.75

Fluid Intelligence Performance in the Elderly: Intraindividual Variability and Conditions of Assessment

Brian F. Hofland, Sherry L. Willis, and Paul B. Baltes
College of Human Development
The Pennsylvania State University

The study of performance factors plays an increasingly salient role in understanding the range of intraindividual variability (plasticity) in intellectual aging. Among performance factors considered are aspects of the testing situation. Two studies are reported that examine intraindividual variability in performance on measures of fluid intelligence (figural relations, induction), varying either practice (retesting) or testing time (standard vs. power) conditions. Subjects are elderly community residents (mean age: 70.6 years; range: 60–84 years). Substantial improvement in level of correct performance (with no evidence for changes in test validity) is obtained for both retest and power conditions. Error patterns, however, differ for the two conditions, with a higher proportion of commission errors occurring under power conditions. Results are interpreted as contributing to the position (a) that older persons continue to show learning capacity and (b) that studying the range of performance under varying conditions is critical to an understanding of intellectual aging.

A major theme in gerontological intelligence research has been the issue of change and variability in intellectual functioning (Baltes & Labouvie, 1973; Botwinick, 1977; Horn, 1978). This research has focused on individual differences, primarily age differences in adult intellectual performance. However, much less attention has been given to the issue of intraindividual variability. Normative or average level of performance has been emphasized rather than the range of variability in the older adult's intellectual performance. The recently revived life-span approach to intellectual development (Baltes & Willis, 1979; Willis & Baltes, 1980) has emphasized variability in intellectual functioning across the life span and suggested that a comprehensive theory of adult intelligence requires information on both inter- and intraindividual variability.

Intraindividual variability can be defined in several ways (e.g., Fiske & Rice, 1955). Long-term intraindividual variability has been studied primarily in longitudinal studies, whereas short-term variability has been considered as change in individual functioning over a short time period. Short-term intraindividual variability has been approached with an emphasis on either uncontrolled fluctuation or the study of the determining factors. The latter is the focus of the present manuscript. Defining the range of short-term variability has typically involved examination of performance under a range of experimental conditions.

In the child development and mental retardation literature, considerable attention has been given to examining the range of short-term intraindividual variability in intellectual functioning. As discussed by Carlson and Wiedl (1978) in the mental re-

The first study reported was conducted as part of the first author's master's thesis under the direction of the second and third authors. The research was supported by National Institute of Aging Grant 5 R01 AG00403-03 awarded to Paul B. Baltes and Sherry L. Willis. The study is part of a research program entitled Adult Development and Enrichment Project (ADEPT), examining the effect of cognitive training on older adults' intellectual performance. Thanks are due to project research assistants Rosemary Blieszner, Steven Cornelius, Margie Lachman, and Ron Spiro, field and training staff Carolyn Nesselroade and Myrtle Williams, and John R. Nesselroade and Paul G. Games, statistical consultants. Brian Hofland was supported in part by National Institute of Aging Predoctoral Traineeship T32-A600048.

Paul B. Baltes is now at the Max-Planck Institute for Human Development and Education, 94 Lentzeallee, 1000 Berlin 33, West Germany.

Requests for reprints should be sent to Sherry L. Willis, S-110 College of Human Development, The Pennsylvania State University, University Park, Pennsylvania 16802.

tardation literature, a distinction can be made between two experimental approaches that have been employed to examine intraindividual variability. One approach has focused on direct instruction (training) by an external agent, such as teacher or parent. Recent training studies have frequently been associated with the concept of "learning potential" (Brown & French, 1979) or "learning diagnostics" (Guthke, 1976). It has been argued that diagnosis of future learning potential in disadvantaged and retarded populations may be more accurately assessed under training or prompting procedures than under standard testing conditions. The second experimental approach has employed less intensive intervention efforts, involving manipulation of variables in the assessment situation (e.g., practice without feedback, reinforcement). These two approaches have typically reflected somewhat different perspectives of the factors underlying deficiencies in intellectual performance. The cognitive training approach has focused on cognitive skills and processes assumed to be involved in certain cognitive tasks. The manipulation of testing conditions has emphasized factors that, although not intrinsic to the target ability per se, may influence intellectual performance. In the child literature, manipulation of variables such as the race and sex of the tester, testing materials format, reinforcement schedules, and amount of practice have been examined (Kinnie & Sternlof, 1971; Brody & Brody, 1976). Research involving the two classes of conditions has been useful not only in defining the range of intraindividual variability in intellectual functioning in childhood, but also the processes and factors influencing intellectual performance.

In the study of intellectual aging, short-term variability has been examined primarily through only one of the two types of strategies outlined, cognitive training research. Recent cognitive training studies (Hornblum & Overton, 1976; Labouvie-Vief & Gonda, 1976; Schultz & Hoyer, 1976; Willis, Blieszner, & Baltes, 1981) have demonstrated significant training effects for older adults' performance on a wide array of cognitive processes and intellectual abilities. Such training results suggest that the elderly can perform at higher levels than indicated by normative or average test scores.

The range of intraindividual variability in intellectual performance associated with manipulation of variables in the testing context has been less thoroughly explored in later adulthood. However, such research is particularly relevant in the study of intellectual aging. There is considerable debate regarding the nature of factors associated with lower intellectual performance in the aged (Botwinick, 1977; Baltes & Willis, 1979). Similar to discussions in prior decades in regard to children's intellectual functioning, some researchers attribute lower performance primarily to cognitive deficits while others emphasize the importance of non-ability-specific performance factors. For example, the elderly are relatively test naive, deficient in many of the test-taking skills (following instructions, using test answer forms) associated with "test-wiseness" (Sarnacki, 1979). Also, the slower response speed of the elderly has frequently been associated with lower intellectual performance on speeded ability tests (Birren, Woods, & Williams, 1980; Cunningham, 1980). Thus, assessment of the range of intraindividual variability associated with manipulation of assessment conditions would be useful in exploring the role of non-ability-specific performance factors.

In this article, two studies are reported, both examining the range of variability in the aged's intellectual performance on two fluid intelligence abilities as a function of two manipulations of the assessment conditions. The first study is the central piece of this article. In it, we examine the effects of multiple retest (practice) sessions on intellectual performance. Both the range of variability associated with practice effects and possible changes in correlational validity of the practice measures are considered. The second study was planned after the first was completed. In the second study, older adults' performance is assessed under standard time limits and a power condition. The range of variability in performance associated with practice (retesting) versus alternate testing conditions (standard vs. power) is compared. Both studies focus on intraindividual variability on measures of fluid

intelligence. Fluid intelligence was chosen as the construct of interest because it has been suggested as one of the most salient ability dimensions characterizing intellectual decline with aging (Horn, 1978).

Study 1: Retesting

Method

Subjects. The sample included 30 older adults (23 females and 7 males), aged 60–80 years (mean age = 69.2 years, SD = 5.18).[1] Mean educational level was 10.43 years (SD = 2.45; range: 7–16 years). Subjects were recruited from community organizations in rural, central Pennsylvania. Self-report health ratings were good, with no serious vision or hearing problems noted. Subjects were paid ($1.15/hr.) either individually or as a contribution to their organization.

Design and procedure. The study involved a pretest and eight 1-hour retest sessions. An extensive battery of fluid and crystallized intelligence measures was administered at pretest in two sessions (2.25 hr./session). Each pretest session involved multiple rest breaks. Pretesting was done by the first author.

At each of the eight retest sessions, two measures marking the fluid abilities of figural relations and induction were administered under standard timed conditions. No external feedback was given to participants. Order of presentation of the two retest measures was counterbalanced across sessions to neutralize possible order and fatigue effects. Because prior retest studies (Vernon, 1954) in younger age groups had usually considered effects across only three to four sessions, this study involved eight sessions, examining a broader range of retest effects. A state anxiety measure was administered at the end of the first and eighth sessions. All retest sessions were conducted by a middle-aged female tester. The eight retest sessions covered a 4-week period.

Measures

Pretest measurement battery. The pretest battery, presented in Table 1, involved 20 tests marking the intellectual dimensions of fluid and crystallized intelligence, and perceptual speed. The battery was a reduced (in test length) version of the fluid–crystallized test battery developed by the Adult Development and Enrichment Project (ADEPT) for use with older adults in other phases of a larger research program. Reduction of test length was based on alpha reliability coefficients computed from data on the full-length ADEPT battery. All reduced tests have estimated reliabilities greater than or equal to .65. Since item selection for the reduced version was stratified by item order, the substantive scope of the reduced version was representative of the full-length battery.

The battery included both published tests shown to represent fluid and crystallized intelligence as well as additional fluid measures (ADEPT Figural Relations, ADEPT Induction) developed in conjunction with the ADEPT project (Blieszner, Willis, & Baltes, in press; Plemons, Willis, & Baltes, 1978).

Retest measures. As mentioned before, the dimension of fluid intelligence has been suggested as evidencing a high level of sensitivity to intellectual decline with aging (Horn, 1978). Therefore, two primary abilities indexing fluid intelligence, figural relations and induction, were selected as retest domains (Cattell, 1971; Horn, 1978). The figural relations ability was represented by the Culture Fair Test (Scale 2, and Power Matrices from Scale 3; Cattell & Cattell, 1957, 1961, 1963). The induction ability was marked by the Induction Composite Test, involving the three tests of Letter Sets (Ekstrom, French, Harman, & Derman, 1976), Letter Series, and Number Series (Thurstone, 1962). Use of both figural relations and induction ability measures provided a broader base for examining retest variability in generalized fluid performance.

State anxiety measure. The Cattell-Nesselroade State Anxiety Scale–Form A (Cattell & Nesselroade, 1974), a 20-item state anxiety scale, was administered at the end of the first and eighth retest sessions. State anxiety was examined to obtain preliminary evidence on subjects' anxiety levels at the beginning and end of the retests.

Results

First, results regarding retest effects on the two retest measures are presented. Then, evidence on correlational relationships between retest scores and factor scores derived from the larger pretest battery is given.

Retest effects. On the left of Table 2, raw score means, standard deviations, score ranges, and the mean percentage correct for the Culture Fair Test (CFR) and Induction Composite Measures (I) are shown for each retest session. The Induction Composite Test score was the sum of subscores from the Letter Sets, Number Series, and Letter Series tests.

A one-factor analysis of variance (ANOVA) with repeated measures across retest trials was performed on raw scores for each of the retest measures. Significant performance gains ($p < .001$) were found for both mea-

[1] Thirty-seven persons began the study. Four subjects voluntarily dropped out before the third retest session because of ill health or scheduling conflicts. Data on three additional subjects who completed all sessions were omitted because they had serious vision and/or hearing problems that interfered with their participation. Study 1 was also conducted with 30 younger adults (mean age = 21.3 years). However, strong ceiling effects occurred early in retests and made definition of that group's performance trends and subsequent age comparisons impossible.

sures, CFR: $F(7, 203) = 16.81$; I: $F(1, 29) = 26.42$. For induction, the assumption of compound symmetry in the population variance–covariance matrix was violated, and the most conservative adjustment of degrees of freedom was used. Figure 1 presents graphically the mean percent of correct solutions for each measure across retest sessions.

Improvement in mean scores on both measures was approximately equivalent to three-fourths of a standard deviation. The

Table 1
Study 1: Pretest Battery of Fluid and Crystallized Intelligence Measures

General intellectual dimension	Primary mental ability	Name of test	Original source
Fluid	Induction	ADEPT Induction Test (Form A):	
		Letter Sets	Bliesznер et al., in press
		Number Series	
		Letter Series	
		Induction Composite Test:[a]	
		Letter Sets	Ekstrom et al., 1976
		Number Series	Thurstone, 1962
		Letter Series	Thurstone, 1962
	Figural Relations	Raven's Advanced Progressive Matrices (Set II)	Raven, 1962
		Culture Fair Test[a] (Scale 2, Form A) and Power Matrices (Scale 3, Form B [1961 ed.])	Cattell & Cattell, 1957, 1961, 1963
		ADEPT Figural Relations Test (Form A)	Plemons et al., 1978
	Memory Span	Visual Number Span	Ekstrom et al., 1976
		Auditory Number Span	After Ekstrom et al., 1976
		Auditory Number Span with Delayed Recall	After Ekstrom et al., 1976
Fluid and crystallized	Semantic Relations	Verbal Analogies	Guilford, 1969
Crystallized	Verbal Comprehension	Verbal Meaning	Thurstone, 1962
		Vocabulary (V-2, V-3, & V-4)	Ekstrom et al., 1976
	Experiential	Social Situations (EP03A)	Horn (Note 1)
	Evaluation	Social Translations (Form A)	Guilford, 1965; O'Sullivan, Guilford, & deMille, Note 2
Speed	Perceptual Speed	Finding A's	Ekstrom et al., 1976
		Number Comparison	Ekstrom et al., 1976
		Identical Pictures	Ekstrom et al., 1976

Note. This battery has been designed by Penn State's ADEPT (Adult Development and Enrichment Project) as a measurement framework for research on psychometric intelligence in gerontology. Extensive information on the assessment battery and its associated structural model is presented in Baltes et al. (1980).
[a] Full-length versions administered at eight retest sessions.

Table 2
Study 1: Descriptive Statistics Across Eight Retest Sessions for Two Fluid Intelligence Retest Measures

Retest session	Retest scores				Error pattern		
	M	Range	*SD*	Correct (%)	Commission (%)	Skipped-omission (%)	Unattempted omission (%)
				Figural relations[a]			
1	29.53	15–45	8.12	41.0	46.0	2.9	10.1
2	30.03	11–48	8.97	41.7	47.7	3.4	7.2
3	30.53	11–46	9.58	42.4	49.0	3.2	5.4
4	33.60	14–51	10.65	46.7	46.8	1.9	4.6
5	34.40	14–50	9.60	47.8	46.3	2.4	3.6
6	35.10	17–49	8.78	48.8	46.3	1.9	3.1
7	33.96	13–51	9.64	47.2	48.0	2.1	2.7
8	36.46	11–53	10.34	50.6	46.2	1.3	1.9
				Induction[b]			
1	21.07	7–45	10.14	30.0	37.9	6.4	25.6
2	21.40	6–49	10.15	30.6	38.1	8.2	23.1
3	24.03	2–50	9.28	34.3	39.2	7.0	19.4
4	25.03	12–50	11.66	35.8	39.7	6.7	17.9
5	26.33	8–54	11.70	37.6	39.1	8.5	14.8
6	28.54	12–60	11.70	40.8	38.1	6.8	14.3
7	28.63	8–58	11.42	40.9	38.5	8.3	12.3
8	29.86	10–64	13.18	42.7	39.2	7.0	11.1

Note. Mean percentage was computed as the mean number correct or of the specified error type divided by the maximum score for measure.
[a] Maximum score = 72. [b] Maximum score = 70.

performance pattern across retest sessions indicated subjects exhibited small, steady gains between consecutive trials. Separate trend analyses for the two measures indicated that only a linear component was significant, CFR: $t(203) = 10.14$; I: $t(29) = 9.17$, $p < .001$. No apparent performance asymptote was reached for either measure. Changes in standard deviations across retests appear minor. However, for the induction measure a trend toward increased variability with retest occurred. To examine retest-associated changes in variability for the induction measure, the test for compound symmetry in the population variance–covariance matrix was followed up with a priori selected contrasts between the variance of Trial 1 with the variances of Trials 7 and 8. The difference between the two dependent variances for Trials 1 and 8 (Glass & Stanley, 1970) was significant, $t(28) = 3.06$, $p < .05$. However, the dependent variances of Trials 1 and 7 did not significantly differ, $t(28) = 1.27$.

Rank ordering of intraindividual improvements across persons was investigated by computing the product-moment correlation between every possible pair of retest trials, separately by measure. These correlations correspond to multiple retest or stability coefficients. Retest stability correlations were high ($r \geq .8$ and $p < .001$ in all cases). These high intertrial correlations indicated that stability of individual rankings was high across retest trials. A subset of these correlations was more closely examined. For each measure, correlations of Trial 1 with the remaining seven trials were tested for significant differences (CFR: range of $r = .81$ to .88; I: range of $r = .87$ to .92). None of the correlations differed significantly. Therefore, the average performance trend depicted in Figure 1 can be taken as an accurate representation of intraindividual trends for this sample of older adults.

Error analysis. An error analysis of retest scores was conducted to examine the nature of retest gains. Incorrect item responses were categorized into three types: (a) commission errors (incorrect responses), (b) skipped-omission errors (items assumed to

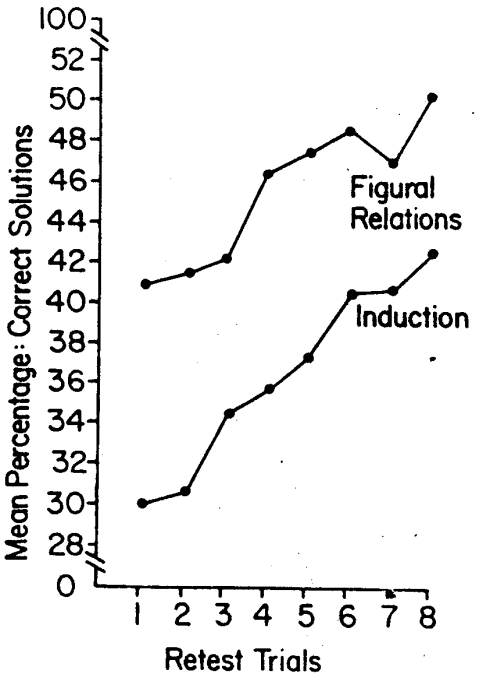

Figure 1. Study 1: Mean percentage of correct solutions across retest trials for measures of figural relations and induction.

have been attempted, but unanswered), and (c) unattempted-omission errors. Any unanswered item followed by a correctly or incorrectly answered item was considered to have been attempted and, therefore, was categorized as a skipped-omission error. Similarly, unanswered items at the end of a test or subtest that followed the last answered item (whether answered correctly or incorrectly) were considered unattempted-omission errors. The maximum score equals the sum of the three error scores plus the correct score. The mean percentages (mean divided by the total maximum correct solutions possible) for each of three error types were computed for both measures for each trial and are shown on the right side of Table 2.

A one-factor ANOVA with repeated measures was performed for each error type for the two retest measures. Since correct responses exhibited a trial main effect and correct responses and errors are algebraically dependent, some trial-related change in error would be expected statistically. However, the nature of the error pattern provides information beyond that relationship. For both tests, there was a significant ($p < .001$) decline in unattempted-omission errors across trials, CFR: $F(1, 29) = 19.15$; I: $F(1, 29) = 16.25$, indicating that more problems were attempted at each succeeding retest session. In contrast, the mean number of commission and skipped-omission errors remained relatively stable across trials. Improvement across trials appeared due to an increase in total items attempted rather than a reduction in commission or skipped-omission errors. In other words, if one takes percentage of commission errors as a measure of accuracy of performance, the level of accuracy remained the same as more items were attempted across trials and thus higher performance resulted.

An increase in the numbers of items attempted across trials and answered correctly takes on added significance if item difficulty increased for later test problems. A substantive task analysis of the two retest measures was performed to assess item difficulty. Item difficulty was evaluated on two dimensions: (a) the total number of relational rules involved in problem solution and (b) the number of distractors and/or irrelevant information. Comparison of the mean number of rules and/or distractors for the first and second halves of each test suggested an increase in item difficulty (CFR: 1st half $M = 13.0$, 2nd half $M = 18.3$; I: 1st half $M = 21.0$; 2nd half $M = 25.3$).

State anxiety. A main effect for trials, $F(1, 29) = 4.84$, $p < .05$, indicated a significant decrease in the Cattell-Nesselroade state anxiety scale from the first retest to the eighth retest occasion. State anxiety scores at Trials 1 and 8 were also compared with normative data on the Cattell-Nesselroade scale for a sample of 111 elderly subjects (mean age 71.5 years) from the same geographical area (Nesselroade, Mitteness, & Thompson, Note 3). Trial 1 scores for the present subjects were significantly higher, $t(139) = 2.07$, $p < .05$, than the normative scores, but the Trial 8 scores were nearly identical to the normative data.

Correlational validity. Possible changes in measurement validity of the retest measures as markers of fluid intelligence were examined by correlating retest scores with

pretest performance on a broad battery of fluid–crystallized measures. A similar strategy was used in an earlier study on learning–ability relationships (Labouvie, Frohring, Baltes, & Goulet, 1973).

Two types of validations were performed and are presented in Figure 2. The first validation involved correlations of retest trial scores with the same test at pretest (auto correlation), shown in the left part of Figure 2. Note, however, that since the pretest measures were reduced in length, the counterpart pretest measures had lower reliability than the full-length retest measures. Correlations of retests with the same test at pretest are high and stable across the eight trials, for both the Culture Fair Test (figural relations) and Induction Composite Measure. The level of these pretest–retest correlations approximates the test reliability estimates for the pretest measures.

In the second type of validation, retest scores at each trial were correlated with four ability factors derived from the pretest battery. A four-factor ability structure obtained in a previous study of this battery (Baltes, Cornelius, Spiro, Nesselroade, & Willis, 1980) was used to estimate four abil-

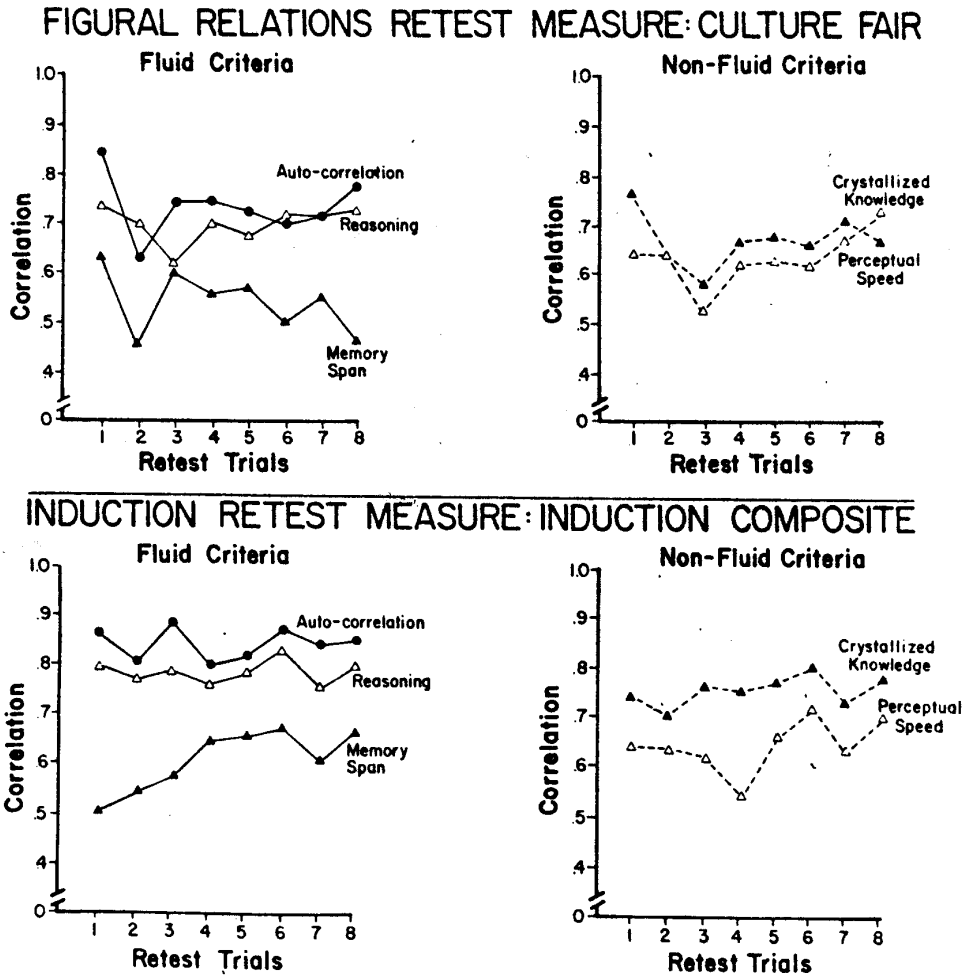

Figure 2. Study 1: Retest measurement validity: correlations between retest measures and pretest scores on same measures (auto-correlations), as well as four factor scores (Reasoning, Memory Span, Crystallized Knowledge, and Perceptual Speed) derived from pretest ability battery.

ity factor scores per subject for pretest scores. To avoid artificially high reliability due to item overlap, computation of these factor score estimates did not include the retest measures. The four ability factors were Reasoning, Memory Span, Crystallized Knowledge, and Perceptual Speed. The reasoning and memory span factors were interpreted to index fluid abilities, and the crystallized knowledge and perceptual speed factors to index nonfluid abilities. In the Baltes et al. study, the intercorrelations among the reasoning, crystallized knowledge, and perceptual speed factors were each approximately .7. In contrast, memory span had correlations of .6 with reasoning and crystallized knowledge, and .4 with perceptual speed.

Figure 2 indicates an overall picture of high correlational stability of the two retest measures. This applies both to their correlations with fluid dimensions (shown on left) and nonfluid dimensions (shown on right). The average correlations of retest trials (figural relations & induction measures combined) with the four factor-score estimates were: reasoning, .75; memory span, .59; crystallized knowledge, .72; and perceptual speed, .64.

Possible deviations from the retest–invariant correlation pattern were examined more closely by testing all possible pairs of dependent correlations for significant differences. Fifteen significant ($p < .05$) differences were found. If these correlational differences were based on independent samples, 14 out of the possible 280 differences would be expected to be significant ($p < .05$). The present finding of 15 significant differences, then, appears to be a chance outcome. One possible exception deals with the memory factor. The memory pretest appears to show a trial-related difference. There were four significant differences between dependent correlations (Trials 1 and 5, 1 and 6, 1 and 8, 3 and 6). A positive and significant increase in the magnitude of the correlation occurs, with memory contributing more variance on induction retests over trials.

Study 2: Speed Versus Power Assessment

The findings of Study 1 indicated considerable improvement in older adults' intellectual performance on two fluid measures related to a retest (practice) condition. It also showed that the correlational validity of the retest measures remained fairly invariant with retesting.

Study 1 did not allow an examination of possible factors underlying retest improvement. Such clarification of the mechanisms underlying retest improvement will require much additional process-oriented research. Of particular interest in this study were non-ability-specific performance factors. Given the salience of the speed dimension in aging research, one interpretation of Study 1 findings would focus on response speed. Improvement across retests may have been partially a function of the older adults' spending less time on items solved in prior sessions and thus having more time to answer previously unattempted items.[2]

Study 2 examined length of testing time as a non-ability-specific variable affecting performance on fluid measures. In Study 2 older adults' performance was examined under two time conditions. In the first time condition the testing procedure (including standard test time limits) was the same as for Study 1. In the second time condition (power administration), the time limits were extended. Comparison was then made of performance gains made from the standard to the more generous time condition. In addition, retest performance improvement observed in Study 1 was compared with performance under the generous time condition of Study 2.

Method

Subjects. The sample included 40 older adults (32 females and 8 males), aged 61–84 years (mean age = 71.8 years, SD = 5.90). Mean educational level was 10.9 years (SD = 2.54; range: 6–16 years). Subjects were similar to those in Study 1 in mean age and educational level and were recruited from community organizations in the same geographical area. Self-report health ratings were good, with no serious vision or hearing problems noted. Subjects were paid ($2.00/hr.) either individually or as a contribution to their organization.

Design and procedure. The study involved two 2-hour sessions. Each session involved administration of one of the two retest measures (Culture Fair Test,

[2] We would like to acknowledge the contribution of an anonymous reviewer. Study 2 was conducted to clarify one of the questions raised in the initial review process dealing with Study 1.

Induction Composit Test) used in Study 1, under a timed and a power condition. Subjects were randomly assigned to one of the two possible orders of presentation of the two measures to neutralize order effects. Each session included multiple rest breaks. The two sessions each dealing with one of the retest measures occurred within a 1-week period. The first author did all testing, assisted by a proctor.

At each session one measure was administered under two conditions to the same subjects. First, the test was given under the same standard time condition as followed in the retest sessions in Study 1. On completion of test administration under standard time conditions, testing was continued under a power condition[3] that allowed twice the standard time limit for that test. Specifically, after administering the test under standard time limits, the subjects were given colored pencils without erasers and told that they would be given two times the previous time limits (per subtest) to continue working on the test. No specific instructions were given on how this additional time was to be used. Subjects were told that they could change any answer marked during the prior administration and/or could answer additional items. Subjects, therefore, were not required to rework items previously answered, but could do so if they wished.

Measures

The two retest measures, the Culture Fair Test (CFR) and Induction Composite Test (I), were the same as in Study 1. However, because in Study 1 there was no evidence for changes in correlational validity, no pretest battery was given.

Results

Scoring procedure. Two scores per subject for each measure were derived: a timed score representing the subject's total correct responses under the standard time condition and a power score representing the subject's correct responses under the timed condition (correct items unchanged under the power condition) plus additional correct responses made under the power condition. The power score, therefore, represented assessment under a condition three times the standard time limit.

Order differences. There were two orders of presentation of the two measures across the two sessions: Order 1 (CFR–I) and Order 2 (I–CFR), with half of the subjects randomly assigned to each order. There was no significant difference in subjects' mean age for Order 1 ($n = 21$) and Order 2 ($n = 19$). The two groups did differ significantly in mean educational level (Order 1: $M = 11.8$; Order 2: $M = 9.9$), $t(38) = 2.39$, $p < .05$. However, neither order differed significantly from the Study 1 subjects in mean age or education.

The two orders did not significantly differ on standard time assessments of Induction Composite Test and Culture Fair Test or on the induction power assessment. However, on the Culture Fair Test power assessment, Order 1 performed significantly ($p < .05$) higher than Order 2, $t(38) = 2.69$. Since Order 1 received the Culture Fair Test at the first session, the higher performance at the power assessment cannot be interpreted as a practice effect. Higher performance is likely a function of higher level of education for the Order 1 group. In all subsequent analyses, both levels of order were combined.

Timed and power scores. On the left of Table 3, raw score means for the Culture Fair Test (CFR) and Induction Composite Test (I), their standard deviations, score ranges, and the mean percent correct are shown for both the standard time and power conditions. Also shown are data for Retest Sessions 1 and 8 of Study 1.

For Study 2 data, dependent t tests were performed on the standard time and power raw mean scores for each measure. The mean scores of standard time and power assessments differed significantly ($p < .001$) for each measure, CFR: $t(39) = 11.77$; I: $t(39) = 11.28$. The difference in mean scores (from standard time to power condition) for figural relation was approximately equivalent to one standard deviation; the difference on induction approximated three-quarters of a standard deviation. Variability of both measures increased from standard time to power assessment. The difference between the two dependent variances (Glass &

[3] Theoretically, a pure power condition involves an untimed assessment procedure. However, as Nunnally (1978) and Anastasi (1976) have pointed out, it is extremely rare for a pure power condition to be employed because of the practical considerations involved. Operationally, a comfortable time limit has been used, defined "as the amount of time required for 90% of the persons to complete a test under power conditions" (Nunnally, 1978, p. 632). The power condition in this study, involving three times the standard time limit, successfully fulfills this operational definition. Note that subjects in the power condition were not required to rework problems answered under the timed condition. Standard time limits for the test were: Culture Fair Test—22 min.; Induction Composite Measure—26 min.

Table 3
Descriptive Statistics for Two Fluid Intelligence Measures: Study 1 and Study 2

Assessment measure/ condition	Performance scores				Error pattern		
	M	Range	*SD*	Correct (%)	Commission (%)	Skipped omission (%)	Unattempted omission (%)
				Figural relations			
Standard time	23.25	7–47	10.92	32.3	45.9	3.0	18.9
Power[a]	32.25	12–56	12.02	44.8	52.3	.6	2.3
Retest 1[b]	29.53	15–45	8.12	41.0	46.0	2.9	10.1
Retest 8[b]	36.46	11–53	10.34	50.6	46.2	1.3	1.9
				Induction			
Standard time	20.83	8–54	11.90	29.8	27.0	12.7	30.5
Power[a]	31.23	15–65	14.72	44.6	44.6	6.0	4.9
Retest 1[b]	21.07	7–45	10.14	30.0	37.9	6.4	25.6
Retest 8[b]	29.86	10–64	13.18	42.7	39.2	7.0	11.1

[a] Tests administered under two times the standard time conditions. Score reflects subject's responses under standard timed condition, and changes and/or additional responses given under power condition. [b] Study 1(n = 30); tests administered under standard time conditions.

Stanley, 1970) for induction was significant, $t(38) = 3.41$, $p < .01$, but the dependent variances for figural relations did not significantly differ, $t(38) = 1.58$. The stability of individual rankings across timed and power conditions was examined by computing product-moment correlations for the standard time scores with the difference scores between the power and the standard scores (CFR: $r = .02$; I: $r = .30$, with $p < .05$ for the latter).

Error analysis. An error analysis of standard time and power scores was conducted to examine the nature of improvement under the power condition. As in Study 1, incorrect responses were categorized into three types: (a) commission errors, (b) skipped-omission errors, and (c) unattempted-omission errors. The mean percentages for each of the three error types were computed for the two scores for each measure and are shown on the right side of Table 3. When interpreting these analyses note that the scores are ipsatively dependent and need to be seen in conjunction.

A dependent t test was performed for each error type for each measure. For both tests, there was a significant ($p < .001$) difference between standard time and power assessment in unattempted-omission errors, CFR: $t(39) = 8.86$; I: $t(39) = 12.33$, and a significant ($p < .01$) difference in skipped-omission errors, CFR: $t(39) = 4.07$; I: $t(39) = 3.63$. This finding indicates that more items were attempted and answered (correctly or incorrectly) under the power condition. There was, however, also a significant difference ($p < .01$) between the standard time and power conditions in the mean number of commission errors for both measures, CFR: $t(39) = 3.00$; I: $t(39) = 9.90$. This outcome indicates that the additional items answered under the power condition resulted in a significant increase in incorrect (as well as correct) responses.

Comparison of performance under retest and power conditions. Performances under varying conditions in Study 1 and Study 2 were compared. The samples, of course, are not random samples. Thus, the comparative assessment is purely descriptive. First, to examine the initial level of similarity of the two groups, the independent t test for each measure was computed between the Study 1 Retest 1 mean and the Study 2 standard time score mean. For figural relations, Study 1 Retest 1 performance was significantly higher, $t(68) = 2.61$, $p < .05$. Another descriptive comparison of interest is that between Study 2 power condition and Study 1 Retest 8. The conditions did not differ from each other in mean number of correct responses.

Error patterns for various conditions in the two studies were also compared. The commission and skipped error patterns

across retests for Study 1 were found to differ from error patterns between standard time and power conditions in Study 2. That is, in Study 1 the mean number of commission and skipped-omission errors remained stable across the eight retest trials with only a significant difference in unattempted-omission errors across retests. However, in Study 2 significant differences between the standard time and power conditions were found in all three error types. Significantly more items were attempted under the power condition, indicated by the skipped- and unattempted-omission error patterns. However, the power condition resulted not only in more correct solutions but also in a higher percentage of commission errors. This was not true for Study 1; as seen in comparing Retest 1 with Retest 8 (Table 3), the percentage of commission errors remained the same. For both measures, the mean number of commission errors for the Study 2 power condition was higher than the mean number of commission errors for Study 1 Retest 8, although in neither case was this difference statistically significant. Thus, whereas in Study 1 the increase in items attempted across retests is directly reflected in the increase in correct solutions with no difference in commission errors, in Study 2, the increase in items attempted in the power condition resulted in more incorrect as well as more correct solutions.

Discussion and Conclusions

In these two studies, the range of variability in older adults' fluid intellectual performance was examined as a function of manipulation of two variables (practice, speed) in the assessment context. It was predicted that performance gains associated with these two factors would occur. These predictions relate to the assumption that the lower performance level of the elderly is partially related to performance factors in the testing context, such as lack of test sophistication, slower response speed, lack of motivation, and test anxiety.

The first study examined practice effects across eight retest sessions. As predicted, test performance on two fluid intelligence measures (figural relations, induction) increased significantly across the eight retest trials on the order of approximately three-fourths of a standard deviation. The pattern of performance gains across the eight retest trials indicated a linear trend of steady increments between consecutive trials with no evidence for a performance asymptote. Retest stability coefficients were consistently high, suggesting similar curves of improvement for all subjects. Moreover, error analyses indicated that their profile remained invariant over retesting. For example, accuracy of performance, if measured by percentage of commission errors, was similar at early and late stages of retesting.

Structural change in correlational validity was also examined. Despite quantitative changes in performance level due to practice, the two fluid intelligence retest measures evidenced traitlike features at the correlational validity level, thereby suggesting that the process (mechanisms) of performance is similar as retesting unfolds. This was evidenced both in correlations of retest measures with their pretest equivalent measures and in correlations of retest measures with estimated factor scores for four broad ability factors (reasoning, memory span, crystallized knowledge, and perceptual speed) obtained at pretest. Thus, in general, the retest measures maintained their validity over time, thereby indexing improvement on the same ability dimensions, figural relations and induction. Contrary to expectations from practice research with younger age groups (Fleishman & Hempel, 1955), there is no evidence of increasing specificity of the retest measures as a function of practice. Increased specificity would have been shown in lower correlations with the four ability factor scores.

In Study 2, standard time limits were manipulated in order to examine the range of intraindividual variability within one session. Performance of a comparable sample of older adults was examined on the two retest measures both under the standard time limits of Study 1 and under a power condition involving a total of three times the standard time limit. This study permitted examination of the magnitude of improvement between standard time and power assessment conditions, and also a descriptive comparison of practice (retest) effects with performance in the power condition. Sig-

nificant improvement from standard time to power conditions was found for both measures with improvement equivalent to three-fourths to one standard deviation. However, in addition to higher performance under the power condition, there was also a different error pattern. Of particular interest is the larger percentage of commission errors for the power condition. In Study 2, then, higher performance associated with more time available for item responding resulted also in more commission errors.

Comparisons of the findings for Study 1 and Study 2 are limited due to the studies' not being based on directly matched samples and, thus, differences in level are somewhat difficult to interpret. However, comparison of within-study effects (e.g., comparing the magnitude of performance improvement for retesting versus variation in time limits) is more defensible. We offer the following observations.

The first descriptive comparison involves level of correct performance. Mean number of correct solutions at Retest 8 (Study 1) and under the power condition (Study 2) were compared. In both instances, performance increments were sizeable, approximating one standard deviation. No significant differences were found between mean correct scores for Retest 8 in Study 1 and mean correct scores for the power condition in Study 2. That is, older adults' performance under generous time limits approximated the performance level attained after multiple retest sessions, when retests were administered under timed conditions.

This finding, however, cannot be taken as indicating that the level of ability following eight retests was similar to that existing at assessment under power conditions. This is so because improvement due to retesting was assessed under standard time conditions only. This is speeded conditions only. Further gain due to retest experience probably would have been shown in Study 1 if Retest 8 had been administered under the power condition of Study 2. The time limits imposed on retest sessions may have restricted the range of variability associated with practice effects which could be demonstrated. In this regard, our current comparisons of Studies 1 and 2 results are limited. The two studies show only that in terms of level of performance, each of the assessments results in substantial evidence for intraindividual variability.

The second descriptive comparison involved examination of error patterns under various conditions in the two studies. Although the mean percentages of items correct are roughly equivalent at Retest 8 and under the power condition, the changes in error patterns between conditions within each study differ. That is, in Study 1, only the percentage of unattempted items decreased significantly across retests, suggesting that improvement across retests was primarily due to subjects' correctly answering more previously unattempted items. There was no evidence of retesting-related changes in error behavior. In contrast, there were significant changes in all three error types between the standard time and power conditions of Study 2. Of particular significance is that subjects under the power condition had a higher proportion of commission errors. Generous time limits, then, while promoting a higher level of correct solutions, also resulted in more errors. Different conditions of the testing context, then, may not simply result in differences in level of correct performance. In addition, they may reflect different processes associated with performance. If magnitude of commission errors is taken as an index of guessing, for example, guessing is higher under power than under standard time or retest conditions.

Together, the studies suggest the importance of examining performance factors in assessment of intraindividual variability in fluid intellectual functioning in later adulthood. Significant improvement was found under both practice and time-relaxed power conditions. It appears that given favorable assessment conditions, the elderly are able, with no direct instruction, to activate cognitive skills already within their repertoire and to significantly improve their performance.

Recent reviews of the literature on intellectual performance in old age (Baltes & Willis, in press) have suggested that systematic information on the range of intraindividual functioning is important in order to achieve an assessment of the potential (Willis & Baltes, 1980) of the elderly.

The present findings contribute to such a posture. Assessment-related conditions are among the performance factors that moderate the level of performance seen in elderly persons. Another major strategy for intervention is, of course, the direct manipulation of cognitive skills as implemented in training research mentioned in the introductory section.

The present studies have focused on within-age comparisons in order to highlight the range of intellectual plasticity within elderly individuals. The results do not speak directly to the question of whether these are life-span differences or changes in the range of intraindividual variability associated with such performance factors. The differential saliency of certain performance factors across the life span appears likely, however, and we tend to believe that the elderly are among those showing marked benefits from enhancing the context of assessment. Examination of the relevant child literature indicates that emphasis often has been placed on different performance factors than those now most salient in the aging literature. Moreover, the salience of a given performance factor may vary with the mental ability or cognitive process examined, as well as the developmental level of the individual. Research on such issues would appear important in development of learning theories and instructional technologies appropriate for educational efforts across the life span.

Reference Notes

1. Horn, J. L. *Social situations—EP03A.* Unpublished test, Department of Psychology, University of Denver, 1967.
2. O'Sullivan, M., Guilford, J. P., & deMille, R. *Measurement of social intelligence* (Report No. 34 from psychology laboratory). Los Angeles: University of Southern California, 1965.
3. Nesselroade, J. R., Mitteness, L. S., & Thompson, L. K. *Structure and stability of anxiety, fatigue, and other pyschological states in the behavior of older adults.* Unpublished manuscript, College of Human Development, The Pennsylvania State University, November 1979.

References

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

Baltes, P. B., Cornelius, S. W., Spiro, A., III, Nesselroade, J. R., & Willis, S. L. Integration vs. differentiation of fluid–crystallized intelligence in old age. *Developmental Psychology*, 1980, *16*, 625–635.

Baltes, P. B., & Labouvie, G. V. Adult development of intellectual performance: Description, explanation, modification. In C. Eisdorfer & M. P. Lawton (Eds.), *The psychology of adult development and aging.* Washington, D.C.: American Psychological Association, 1973.

Baltes, P. B., & Willis, S. L. The critical importance of appropriate methodology in the study of aging: The sample case of psychometric intelligence. In F. Hoffmeister & C. Müller (Eds.), *Brain function in old age.* Heidelberg, New York: Springer, 1979.

Baltes, P. B., & Willis, S. L. Plasticity and enhancement of intellectual functioning in old age: Penn State's Adult Development and Enrichment Project (ADEPT). In F. I. M. Craik & S. E. Trehub (Eds.), *Aging and cognitive processes.* New York: Plenum Press, in press.

Birren, J. E., Woods, A. M., & Williams, M. V. Behavioral slowing with age: Causes, organization, and consequences. In L. W. Poon (Ed.), *Aging in the 1980s: Psychological issues.* Washington, D.C.: American Psychological Association, 1980.

Blieszner, R., Willis, S. L., & Baltes, P. B. Training research in aging on the fluid ability of inductive reasoning. *Journal of Applied Developmental Psychology*, in press.

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

Brody, E. B., & Brody, N. *Intelligence: Nature determinants and consequences.* New York: Academic Press, 1976.

Brown, A., & French, L. The zone of potential development: Implication for intelligence testing in the year 2000. *Intelligence*, 1979, *3*, 255–277.

Carlson, J. S., & Wiedl, K. H. Use of testing-the-limits procedures in the assessment of intellectual capabilities in children with learning difficulties. *American Journal of Mental Deficiency*, 1978, *82*, 559–564.

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

Cattell, R. B., & Cattell, A. K. S. *Test of "g": Culture Fair* (Scale 2, Form A). Champaign, Ill.: Institute for Personality and Ability Testing, 1957.

Cattell, R. B., & Cattell, A. K. S. *Test of "g": Culture Fair* (Scale 3, Form B). Champaign, Ill.: Institute for Personality and Ability Testing, 1961.

Cattell, R. B., & Cattell, A. K. S. *Test of "g": Culture Fair* (Scale 3, Form A). Champaigne, Ill.: Institute for Personality and Ability Testing, 1963.

Cattell, R. B., & Nesselroade, J. R. *The state-trait anxiety battery (STAB).* Champaign, Ill.: Institute for Personality and Ability Testing, 1974.

Cunningham, W. R. Speed, age, and qualitative differences in cognitive functioning. In L. W. Poon (Ed.), *Aging in the 1980s: Psychological issues.* Washington, D.C.: American Psychological Association, 1980.

Ekstrom, R. B., French, J. W., Harman, H., & Derman, D. *Manual for kit of factor-referenced tests.* Princeton, N.J.: Educational Testing Service, 1976.

Fiske, D. W., & Rice, L. Intra-individual response variability. *Psychological Bulletin*, 1955, *52*, 217–250.

Fleishman, E. A., & Hempel, W. E., Jr. The relation between abilities and improvement with practice in a visual discrimination task. *Journal of Experimental Psychology*, 1955, *49*, 301–312.

Glass, G. V., & Stanley, J. C. *Statistical methods in education and psychology*. Englewood Cliffs, N.J.: Prentice-Hall, 1970.

Guilford, J. P. *Social translations test*. Beverly Hills, Calif.: Sheridan Psychological Services, 1965.

Guilford, J. P. *Verbal analogies test, I*. Beverly Hills, Calif.: Sheridan Psychological Services, 1969.

Guthke, V. J. Entwicklungsstand und Probleme der Lernfähigkeitsdiagnostik. *Zeitschrift für Psychologie*, 1976, *184*, 215–239.

Horn, J. L. Human ability systems. In P. B. Baltes (Ed.), *Life-span development and behavior* (Vol. 1). New York: Academic Press, 1978.

Hornblum, J. N., & Overton, W. F. Area and volume conservation among the elderly: Assessment and training. *Developmental Psychology*, 1976, *12*, 68–74.

Kinnie, E. J., & Sternlof, R. E. The influences of non-intellective factors on the IQ scores of middle and lower class children. *Child Development*, 1971, *42*, 1989–1995.

Labouvie, G. V., Frohring, W. R., Baltes, P. B., & Goulet, L. R. Changing relationship between recall performance and abilities as a function of stage of learning and timing of recall. *Journal of Educational Psychology*, 1973, *64*, 191–198.

Labouvie-Vief, G., & Gonda, J. N. Cognitive strategy training and intellectual performance in the elderly. *Journal of Gerontology*, 1976, *31*, 327–332.

Nunnally, J. C. *Psychometric theory* (2nd ed.). New York: McGraw-Hill, 1978.

Plemons, J. K., Willis, S. L., & Baltes, P. B. Modifiability of fluid intelligence in aging: A short-term longitudinal training approach. *Journal of Gerontology*, 1978, *33*, 224–231.

Raven, J. C. *Advanced progressive matrices, Set II* (1962 revision). London: Lewis, 1962.

Sarnacki, R. E. An examination of test-wiseness in the cognitive test domain. *Review of Educational Research*, 1979, *49*, 252–279.

Schultz, N. R., & Hoyer, W. J. Feedback effects on spatial egocentrism in old age. *Journal of Gerontology*, 1976, *31*, 72–75.

Thurstone, T. G. *Primary mental abilities, grades 9–12*, (1962 revision). Chicago: Science Research Associates, 1962.

Vernon, P. E. Practice and coaching effects in intelligence tests. *Educational Forum*, 1954, *18*, 269–280.

Willis, S. L., & Baltes, P. B. Intelligence in adulthood and aging: Contemporary issues. In L. W. Poon (Ed.), *Aging in the 1980s: Psychological issues*. Washington, D.C.: American Psychological Association, 1980.

Willis, S. L., Blieszner, R., & Baltes, P. B. Intellectual training research in aging: Modification of performance on the fluid ability of figural relations. *Journal of Educational Psychology*, 1981, *73*, 41–50.

Received November 7, 1980 ■