Neuropsychology
2008, Vol. 22, No. 6, 800–811

0894-4105/08/$12.00 DOI:10.1037/a0013091

Implications of Short-Term Retest Effects for the Interpretation of Longitudinal Change

Timothy A. Salthouse and Elliot M. Tucker-Drob
University of Virginia

Although within-person comparisons allow direct assessments of change, some of the observed change may reflect effects associated with prior test experience rather than the processes of primary interest. One method that might allow retest effects to be distinguished from other influences of change involves comparing the pattern of results in a longitudinal study with those in a study with a very short retest interval. Three short-term retest studies with moderately large samples of adults are used to provide this type of reference information about the magnitude of change, test-retest correlations, reliabilities of change, and correlations of the change in different cognitive variables with each other, and with other types of variables.

Keywords: longitudinal change, practice effects, maturation, aging

It is widely recognized that a major advantage of longitudinal designs over cross-sectional designs is that within-person change can be measured directly instead of being inferred indirectly from comparisons of different people. Changes observed in longitudinal comparisons are usually attributed to influences operating during the interval between successive measurement occasions, with the nature of the influences varying according to the specific substantive focus of the research. For example, in developmental studies, most of the influences are assumed to reflect processes related to maturation, in intervention studies the influences are assumed to reflect processes related to the treatment, and in studies of disease progression, the influences are assumed to reflect factors associated with the course of the underlying pathology. Issues of interpreting longitudinal change are therefore quite general, but for the sake of simplicity, the following discussion will emphasize a developmental perspective in which processes of maturation are the primary change influences of interest.

Inferences about various aspects of change can be derived from different properties of longitudinal data. For example, the mean change from the first to the second measurement occasion is usually interpreted as a reflection of the magnitude of maturation influences operating across the retest interval. Second, the strength of the correlation between scores on successive occasions is sometimes used as an indirect indication of the amount of individual difference variation in change because these stability correlations can be expected to decrease with increases in the magnitude of individual differences in the size, and direction, of longitudinal change. Third, an inference that maturation affects something that is common to multiple variables might be reached when several variables are available from the same individuals, and the changes in different variables are found to be correlated. And finally, correlations of the measures of change with other variables are often used to identify possible moderators of cognitive aging. To illustrate, a finding that a higher level of education was associated with less negative change could lead to an inference that people with the greatest "cognitive reserve" (e.g., Stern, 2003) are more resistant to age-related cognitive decline.

Although the preceding inferences are often valid, longitudinal comparisons involve successive testing of the same individuals, and thus it is possible that at least some of the observed within-person change in performance is attributable to effects of prior test experience rather than to influences related to maturation. Retest effects are frequently ignored as an influence on longitudinal change, particularly in research on aging, because they are often assumed to be very small or short lasting. However, recent research indicates that retest gains can average .40 standard deviation (*SD*) units or more (for a recent meta-analysis see Hausknecht, Halpert, DiPaolo, & Gerrard, 2007), and can be detected up to five (Burke, 1997) and even 12 (Salthouse, Schroeder, & Ferrer, 2004) years after the initial test.

A number of methods have been developed to take retest effects into account when evaluating change. One such method within the field of neuropsychological assessment is the reliable change index (e.g., Chelune, Naugle, Luders, Sedlak, & Awad, 1993; Frerichs & Tuokko, 2005; Knight, McMahon, Skeaff, & Green, 2007). The primary rationale for our approach, however, is that methods to correct for the influences of retests effects can only be strongly justified, and eventually improved upon, after retest effects are fully characterized and understood. Moreover, in contrast to the reliable change index approach, we emphasize a multivariate perspective in which relations among short-term retest effects in different variables are of interest, and not just the magnitude of retest effects in a single variable.

A key assumption of the research described in this article is that maturation and retest influences might be distinguished with very short-term longitudinal studies, in which the intervals between tests are in the range of days instead of years. The rationale is that little or no influences associated with maturation are likely to be operating with short intervals, and therefore any changes evident

Timothy A. Salthouse and Elliot M. Tucker-Drob, Department of Psychology, University of Virginia.

This research was supported by National Institute on Aging Grant R37AG024270 to TAS.

Correspondence concerning this article should be addressed to Timothy A. Salthouse, Department of Psychology, University of Virginia, Charlottesville, VA 22904-4400. E-mail: salthouse@virginia.edu

under these conditions can be inferred to primarily reflect retest effects. Results from longitudinal studies with very short retest intervals might therefore provide a valuable baseline for interpreting results from conventional longitudinal studies in which the intervals between tests are 1 year or longer. Some allowance must be made for the possibility that retest effects are likely to decay over time, but as noted above, the interval until no effects are detectable could be as long as 12 years. In this article, we report analyses similar to those described above with data from longitudinal studies involving retest intervals averaging about 1 week to illustrate how conclusions from traditional longitudinal studies can be misleading if results from studies with very short retest intervals are not considered.

The data were obtained from three studies in which moderately large samples of adults ranging from 18 to over 80 years of age performed the same battery of 16 cognitive tests either two or three times, with intervals between the tests ranging from 1 day to a few weeks. The participants in Studies 1 and 2 performed different versions of the tests on each of three sessions, with the Study 1 participants tested in 2004, 2005, and 2006, and the Study 2 participants tested in 2007. Although Studies 1 and 2 were identical, they are reported separately to allow a comparison of change on tests with same and different items without a confounding of year of testing. That is, the participants in Study 3, who like those in Study 2 were also tested in 2007, performed exactly the same tests with identical items on the first and second sessions.

Change in two-occasion longitudinal comparisons is typically assessed in one of two ways. The simplest method is with a difference score computed by subtracting the score at the initial occasion (T1) from the score at a later occasion (T2). A second method involves computing a residual score by statistically partialing the influence of the score on the first assessment from the score on the second assessment. The two methods are related as both can be conceptualized in terms of a contrast of the T2 score with T,' where T' is equal to $a + b$(T1). However, in the difference score method the values of a and b are fixed at 0 and 1, respectively, whereas in the residual score method these two parameters are estimated from the data with a least-squares regression equation (cf., Cohen, Cohen, West, & Aiken, 2003, p. 570). Both measures of change are examined in the current report to illustrate potential differences in the patterns of results with the two methods of examining change.

Although estimates of the reliability of measures of change are seldom reported, this information is important for the interpretation of correlations because correlations of changes with other variables are limited by the reliabilities of the measures of change. The data in the current project were recorded at the level of individual items for every participant in each test, and thus separate scores could be computed for the odd-numbered items and for the even-numbered items on each session. This allowed differences and residuals to be computed for the odd and even items, which were then treated as units of analysis in estimating coefficient alpha reliability of the measures of change.

Two individual difference variables, age and general cognitive ability, were also examined with respect to their relations with the measures of short-term change. An estimate of general cognitive ability was created from the first principal component (1st PC) obtained in a principal components analysis based on all of the variables from the first session. An advantage of this method of assessing general cognitive ability is that the 1st PC represents the largest mathematical overlap of the variance among all variables, and involves minimal assumptions about what specific variables represent.

Method

Participants

Characteristics of the participants in the three studies are reported in Table 1. All participants were recruited from newspaper advertisements, flyers, and referrals from other participants, and were paid for their participation. The data in Study 1 were aggregated across several studies originally designed for another purpose, and some of the data were previously analyzed for a study of within-person variability (Salthouse, 2007). Studies 2 and 3 are new and no prior analyses of those data have been published. None of the participants had scores below 24 on the Mini Mental Status Exam (Folstein, Folstein, & McHugh, 1975) that is often used to screen for dementia. Inspection of the entries in Table 1 reveals that the average amount of education was greater than 15 years, and that the average rating of health was in the "very good" to "excellent" range. One method that can be used to evaluate the representativeness of a sample involves examining scores on standardized tests relative to the test norms. It is apparent in Table 1 that the means of the age-adjusted scaled scores for four standardized tests were about one half to one standard deviation above the averages of the nationally representative samples used to establish the norms for those tests. The participants in the current studies can therefore be inferred to have somewhat higher average levels of cognitive abilities than people in the general population, perhaps because they were self-selected volunteers. However, it is important to note that this is true to nearly the same extent at each age, and therefore there is no evidence that certain age groups had higher ability levels than others with respect to the population norms. It should also be noted that the standard deviations of the scaled scores were close to 3, the value in the normative samples, which indicates that these samples exhibited nearly the same amount of variability as the normative samples that were selected to be representative of the U.S. population.

Tests

The cognitive tests are listed in the appendix, and have been described in detail in several other publications (e.g., Salthouse, 2004, 2005, 2007; Salthouse, Atkinson, & Berish, 2003; Salthouse, Berish, & Siedlecki, 2004; Salthouse & Ferrer-Caja, 2003; Salthouse, Siedlecki, & Krueger, 2007). The 16 tests were selected to represent five major cognitive abilities (i.e., reasoning, spatial visualization, episodic memory, perceptual speed, and vocabulary) that have been well established in the cognitive psychometric literature (e.g., Carroll, 1993; Salthouse, 2004). Although not all of these tests are frequently used in neuropsychology, earlier research has established that they have moderate to large correlations with common neuropsychological tests such as the Wisconsin Card Sorting Test, Tower of Hanoi, Stroop, Trail Making, and various fluency tests (e.g., Salthouse, 2005).

Different versions of the tests were performed on each of the three sessions in Studies 1 and 2. The scores on the versions administered on the second and third sessions were equated to the

Table 1
Descriptive Characteristics of the Samples

	Age group			
	18–39	40–59	60–97	All
		Study 1		
N	285	357	379	1,021
Age	25.5 (5.9)	50.8 (5.6)	71.2 (7.7)	51.3 (19.4)
Percent females	62	75	57	65
Self-rated health	2.2 (0.9)	2.3 (0.9)	2.5 (0.9)	2.4 (0.9)
Years of education	15.1 (2.2)	15.9 (2.5)	16.1 (2.8)	15.7 (2.6)
S1-S2 interval (days)	6.3 (8.0)	5.5 (5.5)	5.9 (6.5)	5.9 (6.7)
S2-S3 interval (days)	5.9 (8.6)	5.5 (6.6)	5.0 (6.7)	5.4 (7.3)
Scaled scores				
Vocabulary	13.6 (2.8)	12.5 (2.8)	13.3 (2.5)	13.1 (2.7)
Digit symbol	11.5 (2.4)	11.5 (2.8)	11.7 (2.8)	11.6 (2.7)
Word recall	12.3 (3.1)	12.5 (3.2)	12.8 (3.2)	12.5 (3.2)
Logical memory	11.9 (2.6)	11.8 (2.8)	12.5 (2.7)	12.1 (2.7)
		Study 2		
N	61	65	79	205
Age	25.3 (6.0)	52.0 (5.2)	71.8 (8.6)	51.7 (20.3)
Percent females	54	77	56	62
Self-rated health	2.2 (1.0)	2.2 (1.0)	2.2 (0.9)	2.2 (0.9)
Years of education	14.6 (2.4)	15.3 (3.4)	16.3 (3.1)	15.5 (3.1)
S1-S2 interval (days)	5.1 (5.3)	5.4 (4.6)	6.4 (6.0)	5.7 (5.4)
S2-S3 interval (days)	5.3 (7.8)	5.7 (7.9)	6.0 (7.2)	5.7 (7.5)
Scaled scores				
Vocabulary	12.2 (3.5)	11.6 (2.8)	12.8 (2.6)	12.3 (3.0)
Digit symbol	10.3 (2.9)	10.5 (2.8)	11.5 (3.0)	10.8 (3.0)
Word recall	11.2 (3.3)	12.0 (4.1)	12.3 (3.9)	11.9 (3.8)
Logical memory	11.2 (2.4)	11.7 (3.2)	11.7 (3.0)	11.5 (2.9)
		Study 3		
N	56	87	84	227
Age	25.8 (5.9)	51.6 (4.8)	70.0 (8.0)	52.0 (18.2)
Percent females	57	74	58	64
Self-rated health	1.9 (0.9)	2.1 (0.9)	2.2 (0.9)	2.1 (0.9)
Years of education	14.8 (2.1)	15.7 (2.4)	16.1 (3.4)	15.6 (2.8)
S1-S2 interval (days)	6.8 (8.5)	6.8 (8.3)	6.5 (8.0)	6.7 (8.2)
Scaled scores				
Vocabulary	12.2 (3.0)	11.2 (2.8)	11.7 (2.3)	11.6 (2.7)
Digit symbol	10.6 (2.9)	11.2 (3.0)	11.1 (2.6)	11.0 (2.8)
Word recall	11.7 (2.4)	11.3 (3.5)	10.9 (3.1)	11.3 (3.1)
Logical memory	12.0 (3.1)	11.5 (2.5)	11.2 (3.3)	11.5 (3.0)

Note. Health is a self rating on a scale ranging from 1 for excellent to 5 for poor. Scaled scores are age-adjusted and in the normative samples have means of 10.0 and standard deviations of 3.0 (Wechsler, 1997a, 1997b). S1, S2, and S3 refer to sessions 1, 2, and 3, respectively.

first session means with regression equations based on data from 90 individuals who received the three versions in a counterbalanced order (cf., Salthouse, 2007). Identical versions of the tests were presented in the first and second sessions in Study 3, with the third session containing different types of tests for a separate project. Because the sessions were scheduled according to the participants' availability, the intervals between sessions ranged from 1 day to over 30 days. Means and standard deviations of the retest intervals are presented in Table 1 where it can be seen that the average interval between test sessions was less than 7 days in each study.

Analysis Plan

Six sets of analyses were conducted to address the different aspects of change discussed in the introduction. The initial analyses were conducted to explore properties of the data sets and involved examining the effect of the length of the retest interval on the magnitude of change, and the structural relations among variables across sessions and across studies. The next analyses investigated the magnitude of the retest changes, and the magnitude of the correlations between the scores on successive sessions. The remaining analyses focused on change scores, with the first set examining reliability, and the second set examining intercorrelations among the changes in different cognitive variables. The final analyses examined correlations of age and general cognitive ability with the short-term changes.

Results

An initial set of analyses examined relations between the length of the interval between the first and second sessions and the magnitude of the changes in test performance. The analyses consisted of correlations between the length of the interval and the

Session 2 residual score for each variable. None of the correlations were very large, there were nearly as many positive as negative correlations, and the median correlation was −.01.[1] Therefore, it does not appear that there was much, if any, effect of the length of the interval between sessions on the retest gains in these studies, and thus the retest interval variable was ignored in subsequent analyses. However, it should be noted that the range of retest intervals was highly restricted, with the intervals for most of the participants ranging between 1 and 10 days, and influences of the length of the retest interval might be apparent with greater variation in the intervals.

A second set of analyses consisted of confirmatory factor analyses on the 16 variables from each session in each study. The results of these analyses closely resembled those from other samples (see Salthouse, Pink, & Tucker-Drob, in press). Of particular importance in the current context is that the patterns were also very similar across the sessions within each study as the congruence coefficients (cf., Jensen, 1998) were all greater than .95. The finding of nearly constant relations among the variables suggests that the variables have the same meaning at each session, and in each study.

Average Change

As noted above, each session in Study 1 involved different versions of the cognitive tests. Mean levels of performance for the cognitive variables on Sessions 2 and 3 in this study, expressed in standard deviation (*SD*) units from the scores on the first session, are portrayed in Figure 1. Because zero in this figure represents the average performance on the first session, the heights of the bars represent the size of the retest gains from the first session, scaled relative to the magnitude of individual differences on the task.[2] The magnitude of a given bar therefore corresponds to an estimated effect size for the retest gain, with the standard error bar indicating the precision of the estimate. Because a value that differs from zero by 2.33 standard errors is significant at the .01 significance level, means that are more than 2.33 standard errors from zero are statistically significant. Inspection of the figure reveals that for most variables the largest gains were from the first to the second assessment, with much smaller gains from the second to the third assessment.

There was some variation in the pattern of retest gains across cognitive abilities as the mean gains were small for reasoning variables, modest for memory variables, and large for the spatial visualization and speed variables. However, there was also variation in the magnitude of the retest effects within the same cognitive ability domain. For example, the average gain from the first to the second assessment was fairly small for the Form Boards variable, but relatively large for the Paper Folding and Spatial Relations variables.

Figure 2 uses the same format as Figure 1 to portray scores for Study 2 (with different test versions on the second test session), and for Study 3 (with identical test versions on the second test session). Note that the vertical axis for the episodic memory variables is in a different scale than the other variables to accommodate the large gains evident in some of these variables when successive tests contain identical items. Comparison of the black bars across Figures 1 and 2 reveals that the patterns of retest changes with different test versions on the first and second sessions were very similar in Studies 1 and 2. This finding is not surprising because the studies were exact replications of one another, differing only with respect to the years of testing. Examination of the black and gray bars in Figure 2 reveals that the pattern of changes for identical and different test versions varied across cognitive tests. To illustrate, the gains for identical versions (gray bars) were generally larger than the gains for different versions (black bars) in the reasoning and memory tests, but they were nearly the same magnitude for most of the speed and spatial visualization tests. Independent groups *t* tests revealed that the gains for identical versions were significantly ($p < .01$) greater than the gains for different versions for the Shipley, Form Boards, Word Recall, and Logical Memory tests, but surprisingly were significantly greater for the different version than for the identical version of the Spatial Relations test.

Scores for the vocabulary variables are portrayed in Figure 3, with values for Sessions 2 and 3 in Study 1 on the top, and values in Session 2 for Studies 2 (different versions) and 3 (same versions) in the bottom. It is apparent that the means of the vocabulary variables in Sessions 2 and 3 were relatively small when scaled in Session 1 *SD* units, indicating very little performance gain with retesting. Furthermore, the changes in the vocabulary tests were generally similar across tests with same and different items, with the exception of a significantly larger retest gain when identical items were repeated in the Picture Vocabulary test.

Test-Retest Correlations

Table 2 contains the correlations of the scores across the first two sessions for tests with different items in Studies 1 and 2, and for tests with identical items in Study 3. Medians of the correlations were .69 for Study 1, .75 for Study 2, and .82 for Study 3. Comparison of the correlations in Studies 2 and 3 with *t* tests on Fisher *r*-to-*z* transformed correlations revealed that the correlations with identical versions (Study 3) were significantly ($p < .01$) greater than those with different versions (Study 2) for the Shipley, Letter Sets, Letter Comparison, Spatial Relations tests, and for all four vocabulary tests. Because identical test versions were used on both sessions in Study 3, those values can be interpreted as test–retest reliability coefficients. The moderately high stability coefficients imply that individual differences in change were small

[1] Because of the moderately large sample size and the relatively large number of statistical comparisons, a significance level of .01 was used for all statistical comparisons.

[2] It is important to note that because the standard deviations used to scale the retest effects include variation associated with age, the reported retest gains are likely underestimates of what would be obtained in an age-homogeneous sample. That is, because retest effects correspond to the performance differences across the two sessions divided by the first session standard deviation, the effects in an age-restricted sample will be larger by an amount proportional to the ratio of the age-heterogeneous and age-homogeneous standard deviations. The median ratios of the standard deviations of the original scores and of the residuals after partialing the relations of age were computed in Study 1. The medians were 1.03 for the vocabulary variables, and 1.13, 1.15, 1.09, and 1.22, respectively, for the reasoning, spatial visualization, memory, and speed variables. Therefore, it can be inferred that the size of the retest estimates would likely be 10% to 20% larger in a sample with little or no age variation.

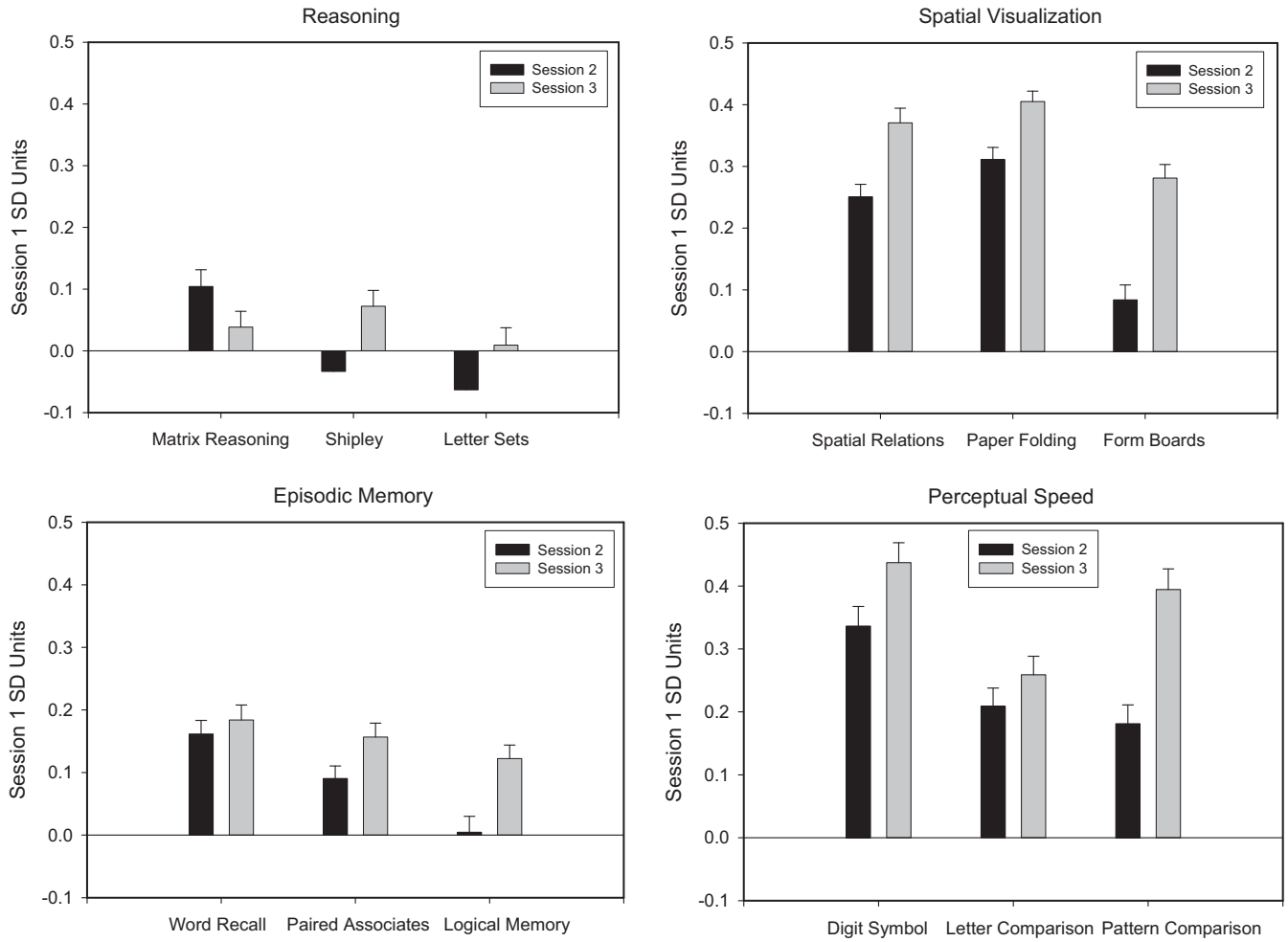

Figure 1. Scores on Sessions 2 and 3 in standard deviation units from the scores on Session 1, Study 1. Bars above each column are standard errors.

relative to the individual differences in the initial scores. Direct computations of the variances in the difference and residual measures of change confirmed this implication. That is, the median variance of the differences and residuals across the three studies were .45 for the difference scores and .28 for the residuals. Because both differences and residuals were assessed in z-score units scaled relative to the distribution of initial scores, and because z-scores have variances of 1.0, these values indicate that individual differences in the change scores were only about one fourth to one half the magnitude of the individual differences in the original scores.

Measurement error can be minimized by forming latent constructs at each occasion, and then examining the across-session correlations at the level of latent constructs. These latent construct analyses were carried out using the AMOS (Arbuckle, 2007) structural equation-modeling program, with separate analyses for each construct, and correlations allowed between the residuals for each variable to account for variable-specific relations across sessions. The latent construct correlations are presented in the bottom of Table 2, where it can be seen that they are all close to 1.0. When examined at the level of latent constructs, therefore, individual differences in short-term change can be inferred to be either extremely small, or possibly even nonexistent.

Reliability of Short-Term Change

Because accuracy was recorded for every item in each test, scores could be computed for the odd-numbered and even-numbered items on each session, as well as for the differences and residuals across sessions. These "odd" and "even" scores were then used to compute coefficient alpha reliability for the Session 1 scores, and for the differences and residuals across Sessions 1 and 2. The estimated reliabilities computed in this manner are summarized in Table 3. The values in the first three columns are reliability estimates for the Session 1 scores, values in the next three columns are estimates of the reliability for the differences, and those in the last three columns are estimates of the reliability for the Session 2 residual scores. Across the three studies, the median estimated reliability for the Session 1 scores was .85, and the corresponding medians for the reliabilities of the differences and residuals were .32 and .42, respectively. It is clear from the

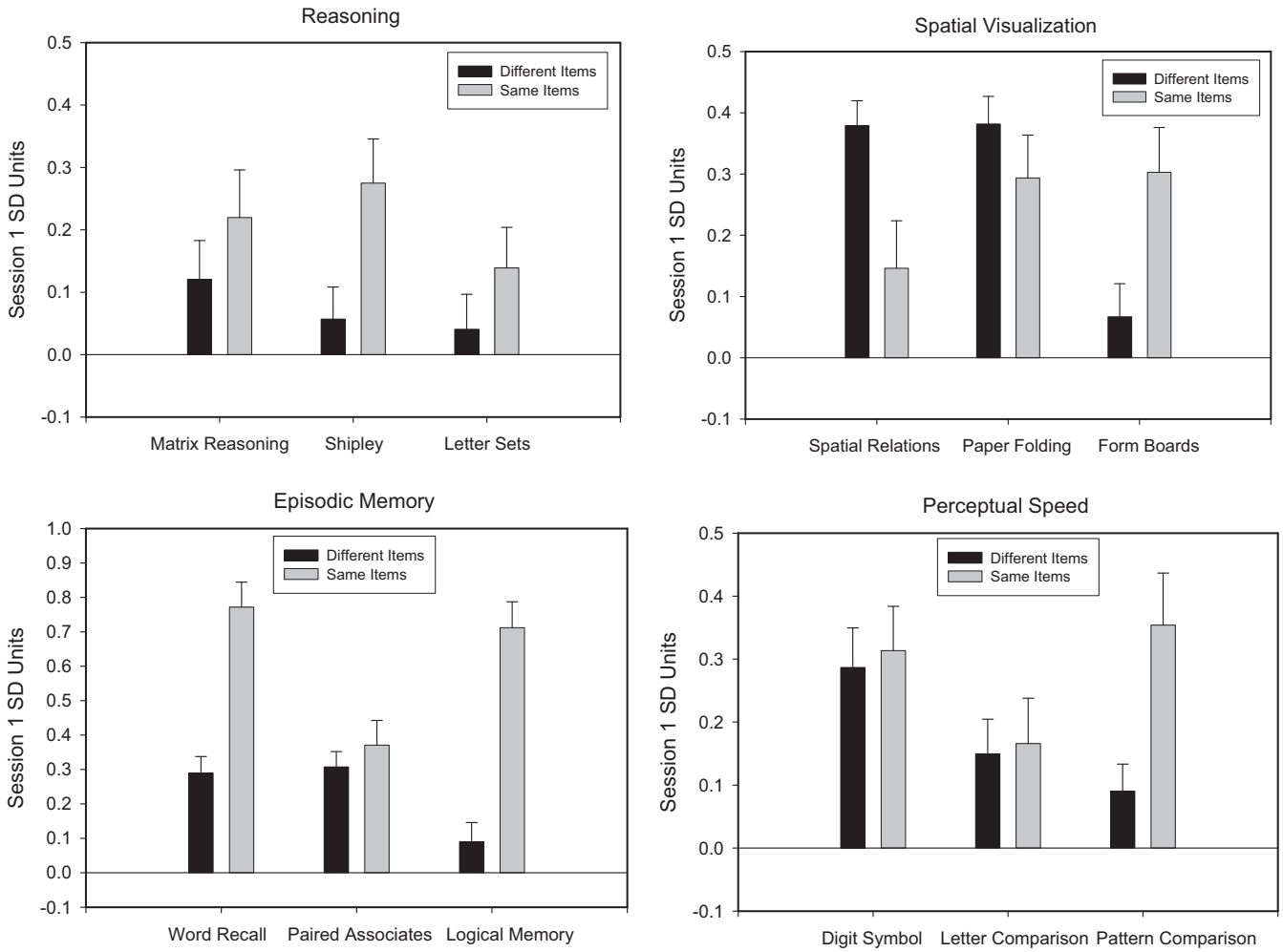

Figure 2. Scores on Session 2 in standard deviation units from the scores on Session 1. Study 2 involved different items on the two sessions, and Study 3 involved the same items on both sessions. Bars above each column are standard errors.

results that reliability is much lower for the measures of change than for the scores on the initial session.

Correlations Among the Short-Term Changes

One of the simplest methods of examining the pattern of interrelations among variables is with exploratory factor analyses. Because the structural relations among the variables are not necessarily similar for Session 1 scores and for differences or residuals, separate exploratory factor analyses were conducted on each type of variable. Across the three studies the first factor was associated with between 40.3% and 42.7% of the variance for the Session 1 scores, but with only from 10.5% to 17.3% of the variance for the differences or for the residuals. Five factors accounted for between 75.9% and 78.1% of the variance in the Session 1 scores, but for only between 40.3% and 54.5% for the differences and for the residuals. Furthermore, the pattern of weaker relations among the measures of change was still evident when the analyses were repeated after adjusting each correlation for unreliability.

Correlations of the changes were also examined among the variables within each domain of cognitive ability. Because residuals are independent of the initial scores, they are the most meaningful measures of change for these analyses. The pattern was very similar in each study, and thus only medians across studies are reported. These medians were .15 for Reasoning, .11 for Space, .18 for Memory, .10 for Speed, and .06 for Vocabulary. For purpose of comparison, the corresponding correlations among the Session 1 scores were .64 for Reasoning, .62 for Space, .51 for Memory, .66 for Speed, and .69 for Vocabulary. As was the case with the reliabilities, therefore, the values for the change measures were markedly weaker than those for the scores on the initial session.

Predictors of Individual Differences in Short-Term Changes

An estimate of general cognitive ability was created from the first principal component (1st PC) obtained in a principal components analysis based on all of the variables from the first session. As noted above, this component was associated with

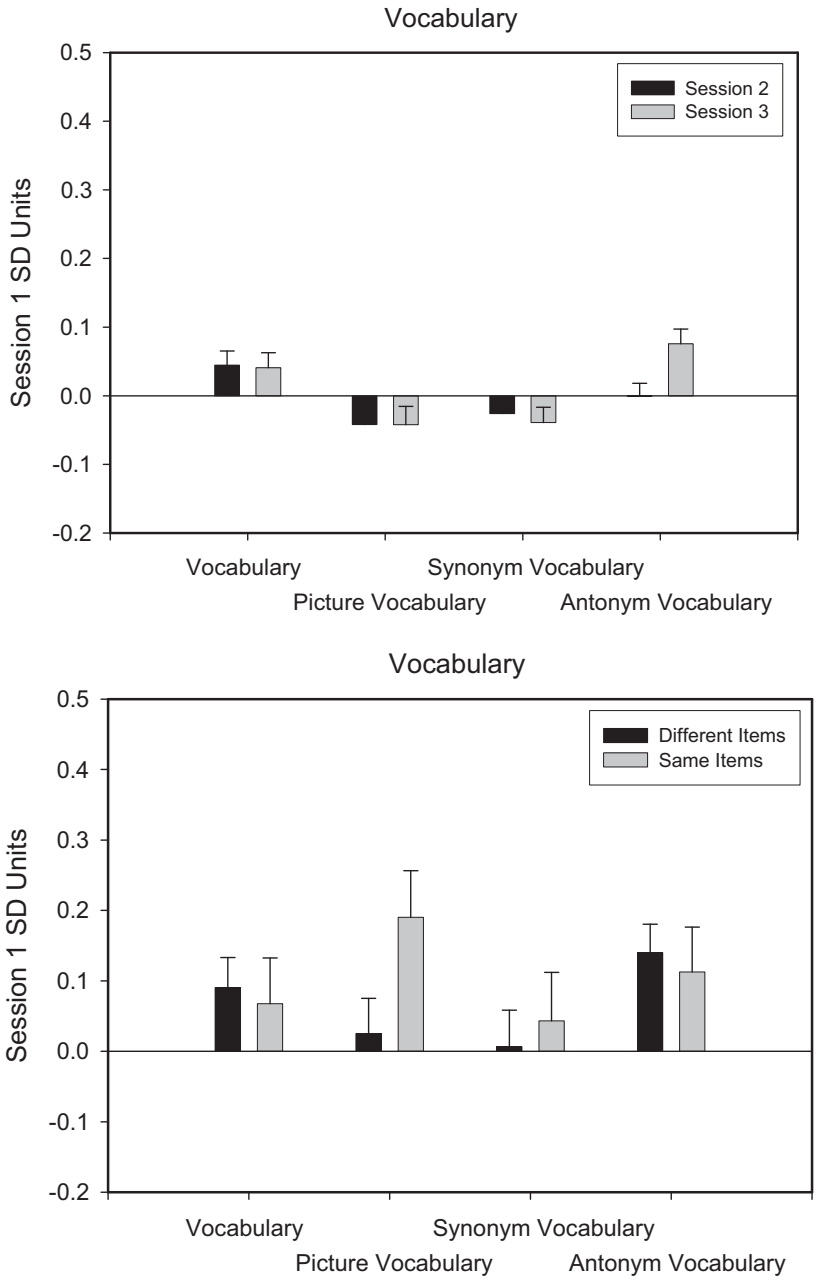

Figure 3. Vocabulary scores on Sessions 2 and 3 in Study 1 (top) and on Session 2 in Studies 2 and 3 (bottom) in standard deviation units from the scores on Session 1. Bars above each column are standard errors.

between 40.3% and 42.7% of the variance in the three studies, and its correlations with age in Studies 1, 2, and 3 were, respectively, −.55, −.40, and −.47.

Table 4 contains simple correlations of age and of the 1st PC for the difference scores and residuals in each study.[3] Because of the negative relation between age and general cognitive ability, some of the relations of the change measures with age are probably mediated through effects on cognitive ability. Indeed, the unique age-related effects, obtained from analyses in which age and the 1st PC were simultaneous predictors of the target variable, were consistently smaller than those reported in Table 4.

Many of the difference scores were positively related to age, and negatively related to general cognitive ability. However, a reverse pattern was evident for the residuals as many of them were

[3] Because the 1st PC is based on the initial values of all of the variables, it is not necessarily independent of the change score for a particular variable. The analyses for each cognitive variable were therefore repeated after deleting that variable from the principal components analyses. Perhaps because each variable was only one of 16 variables contributing to the 1st PC, the results of these analyses were nearly identical to those reported in Table 4.

Table 2
Correlations Between Scores on the First and Second Measurement Occasions

	Study 1	Study 2	Study 3
Reasoning			
Matrix reasoning	.76	.81	.80
Shipley abstraction	.66	.78	.86
Letter sets	.54	.65	.78
Spatial visualization			
Spatial relations	.64	.63	.81
Paper folding	.67	.69	.77
Form boards	.73	.75	.80
Memory			
Word recall	.69	.74	.82
Paired associates	.65	.72	.78
Logical memory	.69	.77	.81
Speed			
Digit symbol	.88	.90	.91
Letter comparison	.72	.73	.86
Pattern comparison	.77	.84	.87
Vocabulary			
WAIS vocabulary	.75	.82	.89
Picture vocabulary	.74	.77	.93
Synonym vocabulary	.58	.55	.85
Antonym vocabulary	.53	.62	.81
Latent constructs			
Reasoning	.99	.98	.98
Spatial visualization	.97	.98	.99
Memory	.92	1.00	.97
Speed	.99	.98	.99
Vocabulary	.94	.99	.99

Note. All values were significantly different from zero at $p < .01$.

negatively related to age, but positively related to general cognitive ability. This reversal is likely attributable to the negative relations between the differences and the original scores, as the median correlations between the T1 score and the T2–T1 difference were −.66 in Study 1, −.68 in Study 2, and −.17 in Study 3.

Discussion

It is apparent in Figures 1 and 2 that there was considerable variation across cognitive variables in the magnitude of the average change from the first to the second session, from the second to the third session, and according to whether successive tests contained identical items or different items. The spatial tests tended to have large average gains, possibly because the items in these tests are unfamiliar to most people. Relatively large gains were also evident on the perceptual speed tests, perhaps because they involve a somewhat novel mode of behavior. The gains were small on reasoning and memory tests when the test versions involved different items, but the increase in performance for some memory tests was as much as .75 *SD* units when the tests on both sessions consisted of identical items.

These results are consistent with earlier reports in several respects. For example, significant short-term retest gains have been reported in a variety of different cognitive tests (e.g., Basso, Bornstein, & Lang, 1999; Benedict, 2005; Benedict & Zgaljardic, 1998; Dikmen, Heaton, Grant, & Temkin, 1999; Duff, Beglinger, Schoenberg, Patton, Mold, Scott, & Adams, 2005; Knight, McHahon, Skeaff, & Green, 2007; Lemay, Bedard, Roulea, & Tremblay, 2004; Levine, Miller, Becker, Selnes, & Cohen, 2004; Lowe & Rabbitt, 1998; Reeve & Lam, 2005; Salinsky, Storzbach, Dodrill, & Binder, 2001; Theisen, Rapport, Axelrod, & Brines, 1998; Wilson, Watson, Baddeley, Emslie, & Evans, 2000; Woods, Delis, Scott, Kramer, & Holdnack, 2006). Furthermore, several studies with three or more test sessions have found that the greatest gain occurs from the first to the second assessment (e.g., Begliner, Gaydos, Tangphao-Daniels, Duff, Kareken, Crawford, Fastenau, & Siemers, 2005; Benedict & Zgaljardic, 1998; Collie, Maruff, Darby, & McStephen, 2003; DeMonte, Geffen, & Kwapil, 2005; Falleti, Maruff, Collie, & Darby, 2006; Hausknecht, Trevor, & Farr, 2002; Hausknecht et al., 2007; Ivnik, Smith, Lucas, Petersen, Boeve, Kokmen, & Tangalos, 1999; Lemay et al., 2004; Rapport, Brines, Axelrod, & Theisen, 1997; Reeve & Lam, 2005; Theisen et al., 1998).

The median short-term change for nonvocabulary variables in successive tests with identical items was .30 *SD* units. The median cross-sectional age slope for these 12 variables was −.024 *SD* per year, and thus the effects of a single prior test are larger than what would be expected across more than 10 years of cross-sectional aging. If these effects were ignored, inferences about the magnitude, and even the direction, of maturational change could be very misleading. This basic point has been recognized for many years, but it has not always been appreciated that the retest influence varies considerably across different cognitive variables. For example, the short-term changes with identical test versions were much larger for certain memory tests than for some tests of reasoning and perceptual speed. In a conventional longitudinal study, results such as these might be interpreted as evidence that cognitive variables differ in their rates of aging, but because the interval between sessions in the current project averaged less than 1 week, all of the differences are attributable to variations in the magnitude of short-term retest effects.

The results of the current studies, and of several earlier studies (e.g., Beglinger et al., 2005; Benedict, 2005; Benedict & Zgaljardic, 1998; Dikmen et al., 1999; Hausknecht et al., 2007; Woods et al., 2006), indicate that for some variables the average retest influences can be minimized, or possibly even eliminated, by the use of alternate forms on successive occasions. However, it is important to note that this is not the case for all variables, because substantial retest gains were apparent in spatial visualization and perceptual speed tests even when the successive tests contained different items.

As noted in the introduction, the magnitude of stability coefficients can be used as an indirect reflection of the amount of between-person variability in change. However, because the test-retest correlations are not 1.0 at intervals ranging from 1 day to a few weeks, correlations with very short-term retest intervals need to be considered when interpreting test–retest correlations with longer intervals. To illustrate, the short-term stability coefficient for the Matrix Reasoning variable in these studies was about .8, and thus the corresponding value in a conventional longitudinal study would have to be appreciably lower than this to justify a conclusion that people differed in their rates of age-related change on this variable.

The across-session correlations between latent constructs formed from three or more variables at each occasion were very close to 1.0. Stability coefficients for latent constructs in conventional longitudinal studies are also often quite high, but there is

Table 3
Estimates of Reliability for Scores on the First Session and for Differences and Residuals

	Session 1 scores			Differences			Residuals		
Study	1	2	3	1	2	3	1	2	3
Reasoning									
Matrix reasoning	.80	.80	.79	.09	.15	.23	.32	.38	.37
Shipley abstraction	.85	.89	.86	.41	.35	.24	.52	.42	.36
Letter sets	.73	.79	.78	.17	.19	.05	.35	.39	.23
Spatial visualization									
Spatial relations	.88	.89	.86	.56	.59	.19	.54	.35	.31
Paper folding	.72	.70	.73	.12	.20	.13	.33	.32	.32
Form boards	.89	.88	.88	.54	.50	.58	.63	.60	.63
Memory									
Word recall	.91	.90	.90	.74	.74	.61	.79	.80	.69
Paired associates	.83	.82	.82	.51	.33	.44	.62	.49	.52
Logical memory	.69	.75	.74	.38	.35	.35	.48	.46	.45
Perceptual speed									
Letter comparison	.86	.86	.88	.47	.44	.26	.54	.50	.36
Pattern comparison	.85	.88	.83	.54	.34	.22	.61	.44	.31
Vocabulary									
WAIS vocabulary	.90	.91	.90	.58	.54	.43	.51	.30	.49
Picture vocabulary	.86	.89	.87	.34	.32	.24	.45	.37	.30
Synonym vocabulary	.79	.81	.80	.33	.36	.13	.28	.33	.27
Antonym vocabulary	.79	.81	.82	.43	.31	.16	.45	.29	.27

Note. Reliability estimates were computed by using the scores for odd-numbered items and for even-numbered items as the "items" in coefficient alpha. No estimates are available for the Digit Symbol variable because it is based on a single score at each session.

seldom any information about the values of the correlations with very short retest intervals. For example, Schaie (2005, Table 8.10) reported correlations across a 7-year interval of .8 or greater for several factor scores, but there was no mention of the correlations across very short intervals that would allow these values to be interpreted as reflections of the magnitude of individual differences in maturational influences.

Most of the estimates of the reliability of the changes were fairly low, which set limits on the relations the measures of change can

Table 4
Correlations of Age and an Estimate of General Cognitive Ability (1st PC) on Difference and Residual Estimates of Change

	Age						1st PC					
	Difference (T2 – T1)			Residual (T2.T1)			Difference (T2 – T1)			Residual (T2.T1)		
Study	1	2	3	1	2	3	1	2	3	1	2	3
Reasoning												
Matrix reasoning	.18*	.16	−.13	−.16*	−.11	−.19*	−.24*	−.20*	.10	.22*	.23*	.21*
Shipley abstraction	.10*	.13	−.03	−.23*	−.20*	−.09	−.29*	−.40*	.13	.25*	.19*	.25*
Letter sets	.15*	.15	−.08	−.06	.02	−.15	−.23*	−.27*	−.01	.18*	.21*	.25*
Spatial visualization												
Spatial relations	.26*	.10	−.04	−.08	−.17	−.06	−.53*	−.49*	.08	.08	.17	.15
Paper folding	.31*	.18*	.04	−.13*	−.23*	−.08	−.48*	−.39*	.04	.18*	.26*	.24*
Form boards	.20*	.07	−.14	−.22*	−.32*	−.22*	−.35*	−.31*	.12	.16*	.17	.24*
Memory												
Word recall	.11*	.04	−.06	−.28*	−.37*	−.14	−.18*	.03	.05	.32*	.55*	.14
Paired associates	.20*	.24*	−.13	−.18*	−.16	−.22*	−.36*	−.37*	.09	.23*	.16	.25*
Logical memory	−.05	−.04	−.06	−.27*	−.26*	−.12	−.13*	−.13	.17	.28*	.28*	.28*
Speed												
Digit symbol	.01	−.02	−.17	−.16*	−.16	−.21*	−.06	−.11	.08	.10*	.05	.14
Letter comparison	.11*	.22*	−.11	−.15*	−.10	−.17*	−.10*	−.21*	.10	.16*	.23*	.17
Pattern comparison	−.00	.26*	−.28*	−.26*	−.01	−.22*	−.02	−.04	.24	.23*	.24*	.18
Vocabulary												
WAIS vocabulary	−.03	−.14	−.03	−.01	−.02	−.02	−.33*	−.39*	−.09	.10*	.12	.10
Picture vocabulary	−.37*	−.42*	−.14	−.24*	−.25*	−.07	−.02	−.16	.12	.27*	.21*	.19*
Synonym vocabulary	−.18*	−.26*	.15	.03	−.00	.19*	−.12*	−.18	−.07	.15*	.19*	.03
Antonym vocabulary	−.12*	−.16	−.06	.03	.10	−.03	−.26*	−.31*	−.01	.13*	.15	.19*

* $p < .01$.

have with other variables. However, it is noteworthy that there was considerable variation in the reliabilities of the change measures across different cognitive variables. As an example, the estimated reliabilities of the measures of change in the Word Recall variable were in the .6 to .8 range, but the estimated reliabilities of the changes in other variables, such as Matrix Reasoning and Paper Folding, were very low. In a conventional longitudinal study, reliability differences such as these could lead to conclusions that some variable, such as physical exercise, cognitive stimulation, type of personality, and so forth, has greater effects on the age-related changes in memory than on the age-related changes in reasoning, when the differential relations could simply reflect differential reliability of the measures of change. More reliable measures of change might be possible by examining change among composite scores or latent constructs, which will tend to have higher reliability at each occasion than the individual scores contributing to the composite, or by using latent difference score analyses (e.g., McArdle & Nesselroade, 1994). However, such approaches do not, by themselves, distinguish between reliable retest-related change and reliable maturation-related change. For such purposes, more sophisticated modeling procedures (e.g., McArdle & Woodcock, 1997) should be considered.

The discovery of weak structure among the measures of change should not be surprising in light of the low reliabilities. The small correlations among the changes are inconsistent with the idea that different variables, even those representing the same type of cognitive ability, change together across short intervals. Stronger evidence for correlated change might be found in a conventional longitudinal study with longer retest intervals, but it would still be informative to compare the correlations with those from a short-term retest study to distinguish the contribution of correlated retest effects from correlated maturation effects (cf., Ferrer, Salthouse, McArdle, Stewart, & Schwartz, 2005).

Another noteworthy finding in the current project is that the direction of the relations of the change measures with other variables depends on how change is assessed. The results in Table 4 reveal that completely opposite conclusions could be reached about the influence of cognitive ability or of age on short-term changes according to whether change was evaluated with difference scores or with residuals. These patterns are likely because of the fact that some of the relations apparent with difference scores reflect relations with the original scores, whereas influences of the original scores are statistically removed with residuals. That is, if age is negatively related to the T1 score then it will tend to be positively related to a difference created by subtraction of the T1 score from the T2 score. Residual measures of change may therefore be more meaningful if one is interested in relations of change measures that are independent of relations among the initial scores.

Many of the residual change measures had negative relations with age, and positive relations with a measure of general cognitive ability. In a conventional longitudinal study, correlations such as these might be interpreted as reflecting influences on rates of aging. For example, the negative age correlations might be interpreted as reflecting more rapid decline at older ages, but because the same pattern is apparent with a very short interval between successive tests, the results could actually reflect smaller benefits of prior testing experience with increased age. Furthermore, the finding of a larger increase (or smaller decline) among individuals with higher levels of general cognitive ability is consistent with the pattern sometimes interpreted as evidence for the notion of cognitive reserve (Stern, 2003), but the results cannot reflect effects on the rate of aging when, as in these studies, the interval between measurement occasions is in the range of days instead of years. The positive relation between initial level of cognitive ability and the magnitude of the retest gain is also consistent with the "rich get richer" suggestion by Rapport et al., (1997), but is inconsistent with recent results by Coyle (2006).

Lower mean levels of performance might be expected on the second assessment when the intervals between tests are longer because of maturation-related declines in ability combined with decay of the retest gains over time. Moreover, if people age at different rates, one might expect relatively low test-retest correlations (i.e., less stability), moderately high reliability of the measures of change, and possibly larger correlations of the measures of change with one another and with other variables. However, the current results with very short-term retest intervals indicate that the values are not 0 (for reliabilities and intercorrelations) or 1.0 (for test–retest correlations) when no maturational influences are operating, and thus the absolute magnitudes of these parameters can only be meaningfully interpreted by considering the corresponding values with very short retest intervals.

The major implication of the current analyses for neuropsychological research is that merely because changes are observed does not mean that neurodegenerative processes related to disease, pathology, trauma, or aging are being evaluated (or at least solely evaluated). Retest effects were found to not only influence mean levels of performance, but also to differentially impact individuals of different ages and ability levels. Conventional examinations rely on the use of predictors of longitudinal changes to make inferences about risk or protective factors associated with cognitive/neuropsychological deficits, but the current results suggest that some of the relations may be attributable to individual differences in the magnitude of retest effects. However, one can have greater confidence that such patterns reflect only the processes of interest when patterns from long retest intervals (or from patient groups) are substantially different from the patterns with very short retest intervals (or from healthy control groups). Although it will likely add to the time and expense of the research, including such "control" observations could greatly increase the interpretability of longitudinal research.

References

Arbuckle, J. L. (2007). *AMOS 7.0 User's Guide.* SPSS, Inc., Chicago.

Basso, M. R., Bornstein, R. A., & Lang, J. M. (1999). Practice effects on commonly used measures of executive function across twelve months. *The Clinical Neuropsychologist, 10,* 283–292.

Beglinger, L. J., Gaydos, B., Tangphao-Daniels, O., Duff, K., Kareken, D. A., Crawford, J., et al. (2005). Practice effects and the use of alternate forms in serial neuropsychological testing. *Archives of Clinical Neuropsychology, 20,* 517–529.

Benedict, R. H. (2005). Effects of using same- versus alternate-form memory tests during short-interval repeated assessments in multiple sclerosis. *Journal of the International Neuropsychological Society, 11,* 727–736.

Benedict, R. H. B., & Zgaljardic, D. L. (1998). Practice effects during repeated administrations of memory tests with and without alternate forms. *Journal of Clinical and Experimental Neuropsychology, 20,* 339–352.

Bennett, G. K., Seashore, H. G., & Wesman, A. G. (1997). *Differential Aptitude Test.* San Antonio, TX: The Psychological Corporation.

Burke, E. F. (1997). A short note on the persistence of retest effects on aptitude scores. *Journal of Occupational and Organizational Psychology, 70,* 295–301.

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

Chelune, G. J., Naugle, R. I., Luders, H., Sedlak, J., & Awad, I. A. (1993). Individual change after epilepsy surgery: Practice effects and base-rate information. *Neuropsychology, 7,* 41–52.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Collie, A., Maruff, P., Darby, D. G., & McStephen, M. (2003). The effects of practice on the cognitive test performance of neurologically normal individuals assessed at brief test-retest intervals. *Journal of the International Neuropsychological Society, 9,* 419–428.

Coyle, T. R. (2006). Test-retest changes on scholastic aptitude tests are not related to g. *Intelligence, 34,* 15–27.

DeMonte, V. E., Geffen, G. M., & Kwapil, K. (2005). Test-retest reliability and practice effects of a rapid screen of mild traumatic brain injury. *Journal of Experimental and Clinical Neuropsychology, 27,* 624–632.

Dikmen, S. S., Heaton, R. K., Grant, I., & Temkin, N. R. (1999). Test-retest reliability and practice effects of expanded Halstead-Reitan Neuropsychological Test Battery. *Journal of the International Psychological Society, 5,* 346–356.

Duff, K., Beglinger, L. J., Schoenberg, M. R., Patton, D. E., Mold, J., Scott, J. G., et al. (2005). Test-retest stability and practice effects of the RBANS in a community dwelling elderly sample. *Journal of Clinical and Experimental Neuropsychology, 27,* 565–575.

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

Falleti, M. G., Maruff, P., Collie, A., & Darby, D. G. (2006). Practice effects associated with the repeated assessment of cognitive function using the CogState Battery at 10-minute, one week and one month test-retest intervals. *Journal of Clinical and Experimental Neuropsychology, 28,* 1095–1112.

Ferrer, E., Salthouse, T. A., McArdle, J. J., Stewart, W. F., & Schwartz, B. S. (2005). Multivariate modeling of age and retest in longitudinal studies of cognitive abilities. *Psychology and Aging, 20,* 412–422.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). Mini-mental state: A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research, 12,* 189–198.

Frerichs, R. J., & Tuokko, H. A. (2005). A comparison of methods for measuring cognitive change in older adults. *Archives of Clinical Neuropsychology, 20,* 321–333.

Hausknecht, J. P., Halpert, J. A., Di Paolo, N. T., & Gerrard, M. O. M. (2007). Retesting in selection: A meta-analysis of coaching and practice effects for tests of cognitive ability. *Journal of Applied Psychology, 92,* 373–385.

Hausknecht, J. P., Trevor, C. O., & Farr, J. L. (2002). Retaking ability tests in a selection setting: Implications for practice effects, training performance, and turnover. *Journal of Applied Psychology, 87,* 243–254.

Ivnik, R. J., Smith, G. E., Lucas, J. A., Petersen, R. C., Boeve, B. F., Kokmen, E., et al. (1999). Testing normal older people three or four times at 1- to 2-year intervals: Defining normal variance. *Neuropsychology, 13,* 121–127.

Jensen, A. (1998). *The g factor: The science of mental ability.* Westport, CT: Praeger.

Knight, R. G., McMahon, J., Skeaff, C. M., & Green, T. J. (2007). Reliable change index scores for persons over the age of 65 tested on alternate forms of the Rey AVLT. *Archives of Clinical Neuropsychology, 22,* 513–518.

Lemay, S., Bedard, M. A., Roulea, I., & Tremblay, P. L. G. (2004). Practice effect and test-retest reliability of attentional and executive tests in middle-aged to elderly subjects. *The Clinical Neuropsychologist, 18,* 284–302.

Levine, A. J., Miller, E. N., Becker, J. T., Selnes, O. A., & Cohen, B. A. (2004). Normative data for determining significance of test-retest differences on eight common neuropsychological instruments. *The Clinical Neuropsychologist, 18,* 373–384.

Lowe, C., & Rabbitt, P. (1998). Test/re-test reliability of the CANTAB and ISPOCD neuropsychological batteries: Theoretical and practical issues. *Neuropsychologia, 36,* 915–923.

McArdle, J. J., & Nesselroade, J. R. (1994). Using multivariate data to structure developmental change. In S. H. Coren & H. W. Reese (Eds.), *Lifespan developmental psychology: Methodological contributions* (pp. 223–267). Hillsdale, NJ: Erlbaum.

McArdle, J. J., & Woodcock, J. R. (1997). Expanding test-rest designs to include developmental time-lag components. *Psychological Methods, 2,* 403–435.

Rapport, L. J., Brines, D. B., Axelrod, B. N., & Theisen, M. E. (1997). Full scale IQ as mediator of practice effects: The rich get richer. *The Clinical Neuropsychologist, 11,* 375–380.

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

Reeve, C. L., & Lam, H. (2005). The psychometric paradox of practice effects due to retesting: Measurement invariance and stable ability estimates in the face of observed score changes. *Intelligence, 33,* 535–549.

Salinsky, M. C., Storzbach, D., Dodrill, C. B., & Binder, L. M. (2001). Test-retest bias, reliability, and regression equations for neuropsychological measures repeated over a 12–16-week period. *Journal of International Neuropsychological Society, 7,* 597–605.

Salthouse, T. A. (1993). Speed and knowledge as determinants of adult age differences in verbal tasks. *Journal of Gerontology: Psychological Sciences, 48,* P29–P36.

Salthouse, T. A. (2004). Localizing age-related individual differences in a hierarchical structure. *Intelligence, 32,* 541–561.

Salthouse, T. A. (2005). Relations between cognitive abilities and measures of executive functioning. *Neuropsychology, 19,* 532–545.

Salthouse, T. A. (2007). Implications of within-person variability in cognitive and neuropsychological functioning on the interpretation of change. *Neuropsychology, 21,* 401–411.

Salthouse, T. A., Atkinson, T. M., & Berish, D. E. (2003). Executive functioning as a potential mediator of age-related cognitive decline in normal adults. *Journal of Experimental Psychology: General, 132,* 566–594.

Salthouse, T. A., & Babcock, R. L. (1991). Decomposing adult age differences in working memory. *Developmental Psychology, 27,* 763–776.

Salthouse, T. A., Berish, D. E., & Siedlecki, K. L. (2004). Construct validity and age sensitivity of prospective memory. *Memory & Cognition, 32,* 1133–1148.

Salthouse, T. A., & Ferrer-Caja, E. (2003). What needs to be explained to account for age-related effects on multiple cognitive variables? *Psychology and Aging, 18,* 91–110.

Salthouse, T. A., Fristoe, N., & Rhee, S. H. (1996). How localized are age-related effects on neuropsychological measures? *Neuropsychology, 10,* 272–285.

Salthouse, T. A., Pink, J. E., & Tucker-Drob, E. M. (in press). Contextual analysis of fluid intelligence. *Intelligence.*

Salthouse, T. A., Schroeder, D. H., & Ferrer, E. (2004). Estimating retest effects in longitudinal assessments of cognitive functioning in adults between 18 and 60 years of age. *Developmental Psychology, 40,* 813–822.

Salthouse, T. A., Siedlecki, K. L., & Krueger, L. E. (2006). An individual differences analysis of memory control. *Journal of Memory and Language, 55,* 102–125.

Schaie, K. W. (2005). *Developmental influences on adult intelligence: The Seattle Longitudinal Study.* New York: Oxford University Press.

Stern, Y. (2003). The concept of cognitive reserve: A catalyst for research. *Journal of Clinical and Experimental Neuropsychology, 25,* 589–593.

Theisen, M. E., Rapport, L. J., Axelrod, B. N., & Brines, D. B. (1998). Effects of practice in repeated administrations of the Wechsler Memory Scale-Revised in normal adults. *Psychological Assessment, 5,* 85–92.

Wechsler, D. (1997a). *Wechsler Adult Intelligence Scale*–Third Edition. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1997b). *Wechsler Memory Scale*–Third Edition. San Antonio, TX: The Psychological Corporation.

Wilson, B. A., Watson, P. C., Baddeley, A. D., Emslie, H., & Evans, J. J. (2000). Improvement or simply practice? The effects of twenty repeated assessments on people with and without brain injury. *Journal of the International Neuropsychological Society, 6,* 469–479.

Woodcock, R. W., & Johnson, M. B. (1990). *Woodcock-Johnson Psycho-Educational Battery*–Revised. Allen, TX: DLM.

Woods, S. P., Delis, D. C., Scott, J. C., Kramer, J. H., & Holdnack, J. A. (2006). The California Verbal Learning Test–2nd ed. Test-retest reliability, practice effects, and reliable change indices for the standard and alternate forms. *Archives of Clinical Neuropsychology, 21,* 413–420.

Zachary, R. A. (1986). *Shipley Institute of Living Scale*–Revised. Los Angeles: Western Psychological Services.

Appendix

Description of Reference Variables and Sources of Tasks

Variable	Description	Source
Matrix reasoning	Determine which pattern best completes the missing cell in a matrix	Raven (1962)
Shipley abstraction	Determine the words or numbers that are the best continuation of a sequence	Zachary (1986)
Letter sets	Identify which of five groups of letters is different from the others	Ekstrom et al. (1976)
Spatial relations	Determine the correspondence between a 3-D figure and alternative 2−D figures	Bennett et al. (1997)
Paper folding	Determine the pattern of holes that would result from a sequence of folds and a punch through folded paper	Ekstrom et al. (1976)
Form boards	Determine which combinations of shapes are needed to fill a larger shape	Ekstrom et al. (1976)
Logical memory	Number of idea units recalled across three stories	Wechsler (1997b)
Free recall	Number of words recalled across trials 1 to 4 of a word list	Wechsler (1997b)
Paired associates	Number of response terms recalled when presented with a stimulus term	Salthouse et al. (1996)
Digit symbol	Use a code table to write the correct symbol below each digit	Wechsler (1997a)
Letter comparison	Same/different comparison of pairs of letter strings	Salthouse & Babcock (1991)
Pattern comparison	Same/different comparison of pairs of line patterns	Salthouse & Babcock (1991)
WAIS vocabulary	Provide definitions of words	Wechsler (1997a)
Picture vocabulary	Name the pictured object	Woodcock & Johnson (1990)
Antonym vocabulary	Select the best antonym of the target word	Salthouse (1993)
Synonym vocabulary	Select the best synonym of the target word	Salthouse (1993)

Received January 29, 2008
Revision received May 29, 2008
Accepted June 11, 2008 ■