Neuropsychology
2007, Vol. 21, No. 4, 401–411

0894-4105/07/$12.00 DOI: 10.1037/0894-4105.21.4.401

Implications of Within-Person Variability in Cognitive and Neuropsychological Functioning for the Interpretation of Change

Timothy A. Salthouse
University of Virginia

Samples of adults across a wide age range performed a battery of 16 cognitive tests in 3 sessions within an interval of approximately 2 weeks. Estimates of within-person variability across the 3 assessments were relatively large and were equivalent in magnitude to the cross-sectional age differences expected over an interval of 15–25 years. These findings raise questions about the precision of assessments based on a single measurement and imply that it may be difficult to distinguish true change from short-term fluctuation. Because there were large individual differences in the magnitude of this variability, it is proposed that change might be most meaningfully expressed in units of each individual's own across-session variability.

Keywords: cognition, aging, longitudinal change, measurement burst

Several recent articles have reported the existence of substantial within-person variability in performance on the same cognitive and neuropsychological tests across multiple occasions. To illustrate, in studies by Salthouse and colleagues, the within-person (across-session) standard deviation averaged about 50% of the between-person standard deviation for a variety of different cognitive variables (e.g., Nesselroade & Salthouse, 2004; Salthouse & Berish, 2005; Salthouse, Nesselroade, & Berish, 2006).

This phenomenon of sizable within-person variability is interesting for at least three reasons. First, measures of within-person variability could have unique diagnostic significance compared with measures of central tendency. That is, how much a person's performance varies around his or her average level on a specific test could be an early predictor of impending change to a different level of functioning. Consistent with this interpretation are several reports of significant relations between measures of within-person variability and neurological status (e.g., Burton, Strauss, Hultsch, Moll, & Hunter, 2006; Hultsch, MacDonald, Hunter, Levy-Bencheton, & Strauss, 2000; Murtha, Cismaru, Waechter, & Chertkow, 2002; Strauss, MacDonald, Hunter, Moll, & Hultsch, 2002) and even risk of death (Shipley, Der, Taylor, & Deary, 2006).

A second reason why within-person variability is important is that it suggests that single assessments may not be sufficient for accurate evaluation of an individual if another assessment with the same test could lead to a different level of performance and, possibly, to a different diagnostic classification. Little is currently known about the potential impact of this type of problem, but it will likely depend on both the magnitude of the variability and the range of variables that exhibit within-person variability.

A third reason why the phenomenon of within-person variability is important is that the existence of moderate within-person variability could complicate the interpretation of within-person change, because it implies that each measurement can be viewed as only one score from a distribution of many possible scores that could have been observed for the individual (e.g., Nesselroade, 1991; Nesselroade & Salthouse, 2004; Salthouse et al., 2006). Some of what is interpreted as change, or lack of change, may therefore be attributable to short-term fluctuation and sampling variation rather than true change. The situation is even more complicated if people vary in the amount of within-person variability, because the same absolute amount of change could have different meanings in different individuals (e.g., Salthouse, Kausler, & Saults, 1986; Salthouse et al., 2006).

One possible solution to the concerns about imprecise assessment and ambiguity of change involves determining an individual's short-term fluctuation at each measurement occasion and then using the distribution of scores at each occasion to express that person's change in individually determined *t*-score units. This proposal is schematically illustrated in Figure 1, with each measurement occasion consisting of three separate assessments. The left panel indicates that with three assessments at each occasion, there are nine possible differences between the two time periods, with some of the differences representing potential increases in level of functioning and others representing potential decreases. The right panel in Figure 1 illustrates the proposed analytical method, in which the difference between the two distributions is evaluated in terms of the difference between the means scaled in standard deviation units. Expressing the across-time difference relative to the distribution of scores takes short-term fluctuation at each occasion into account when evaluating change, and individual differences in short-term variability are incorporated into the evaluations by calibrating change in terms of each individual's own level of short-term variability. Another advantage of the proposed method is that because change is expressed as a *t* score, the statistical significance of change can be evaluated within a single individual without reference to data from other individuals.

Implementation of this proposal requires a measurement burst design (cf. Nesselroade, 1991), in which each individual is assessed multiple times at each occasion. There are obviously practical limits on the number of assessments that are feasible at each occasion, both because of potential burden on examinees and

This research was supported by National Institute on Aging Grants R37 AG24270 and R01 AG19627. I thank John Nesselroade and Elliot Tucker-Drob for their valuable suggestions and Cris Rabaglia for coordinating the data collection.

Correspondence concerning this article should be addressed to Timothy A. Salthouse, Department of Psychology, 102 Gilmer Hall, University of Virginia, Charlottesville, VA 22904-4400.

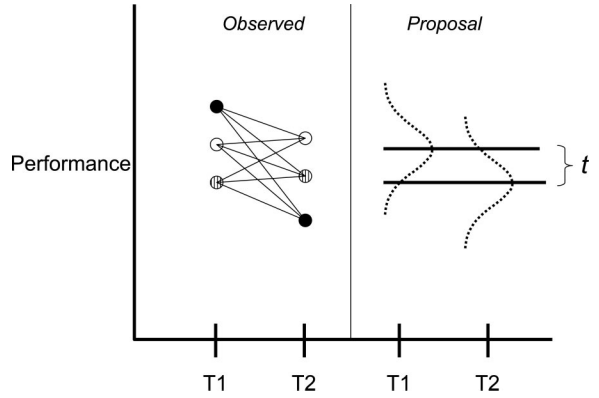

Figure 1. Schematic illustration of the ambiguity of change when there is variability in level of performance at each measurement occasion (i.e., T1 and T2). The right panel portrays the proposal to assess change in terms of the *t* value between the distributions of scores from the two occasions.

because of the difficulty of creating a large number of parallel versions of different cognitive or neuropsychological tests. A measurement burst study therefore inevitably involves a trade-off between the number of assessments necessary to obtain a reasonable estimate of variability and the pragmatic difficulties associated with extensive testing. The design implemented in a recent project in my laboratory involves administering three different versions of 16 tests in three sessions within a period of about 2 weeks (see, e.g., Salthouse et al., 2006).

The cognitive tests have been described in other reports (e.g., Salthouse, Atkinson, & Berish, 2003; Salthouse & Ferrer-Caja, 2003; Salthouse et al., 2006) and are briefly summarized in Table 1. Unlike many previous studies of within-person variability, our studies examined a broad variety of cognitive variables and not just various types of reaction time. For every variable, within-person variability for a given individual was represented by the standard deviation of his or her scores across the three sessions (versions).

To attribute any observed variability to fluctuations within the person rather than to differences in the difficulty of the tests, the researcher must ensure that the test versions are equivalent. The first study in the current project therefore involved the administration of the three versions (designated O for original and A and B for the two alternative versions) in a counterbalanced order across sessions to examine possible version differences, independently of the order in which they were presented. This is similar to the procedure used by Salthouse et al. (2006), but the sample in the current project consisted of adults across a wide age range rather than only young adults. The data from this calibration study were then used to equate scores on Version A (or B) to those of Version O by predicting the score from the A (or B) score in a linear regression equation, and then using the intercept and slope parameters of that equation to generate predicted A′ (or B′) values. The major advantage of this method over simple mean adjustment is that the degree of adjustment can vary according to the level of the predictor variable (i.e., the A or B score).

The second study in the current project involved administering the three test versions in the same order (i.e., O, A, B) to a large sample of adults across a wide range of ages. The primary purpose of this study was to conduct detailed analyses of the magnitude of within-person variability across a variety of cognitive and neuropsychological variables with a much larger sample than that in Salthouse et al. (2006). Secondary goals of the study were to assess the reliability of the measures of within-person variability and to determine the simple and unique relations of within-person variability to age.

Reliability of within-person variability is of interest because within-person variability is only meaningful as an individual difference variable if it is reliable. Very few studies have reported the reliability of measures of within-person variability,

Table 1
Brief Description of Cognitive Tests

Test	Description	Source
Vocabulary	Provide the definition of words read by the examiner	Wechsler (1997a)
Picture vocabulary	Name pictured objects	Woodcock & Johnson (1990)
Synonym vocabulary	Select the best synonym of the target word	Salthouse (1993)
Antonym vocabulary	Select the best antonym of the target word	Salthouse (1993)
Matrix reasoning	Select the best completion of the missing cell in a matrix of geometric patterns	Raven (1962)
Shipley abstraction	Select the best continuation of a series of items	Zachary (1986)
Letter sets	Select the set of letters that does not belong with the others	Ekstrom, French, Harman, & Dermen (1976)
Spatial relations	Determine which three-dimensional object matches the two-dimensional drawing	Bennett, Seashore, & Wesman (1997)
Paper folding	Determine which pattern of holes would result from the paper being folded and a hole being punched in the designated location	Ekstrom et al. (1976)
Form boards	Determine which pieces are needed to assemble a target shape	Ekstrom et al. (1976)
Word recall	Listen to 12 unrelated words, recall as many as possible in any order, and repeat for four trials	Wechsler (1997b)
Logical memory	Listen to a story and recall as much as possible	Wechsler (1997b)
Paired associates	Listen to 6 pairs of unrelated words and recall the second member of the pair when presented with the first	Salthouse, Fristoe, & Rhee (1996)
Digit symbol	Use a code table to substitute as many symbols for digits as possible within 120 s	Wechsler (1997a)
Letter comparison	Classify sets of letters as same or different as rapidly as possible	Salthouse & Babcock (1991)
Pattern comparison	Classify sets of line patterns as same or different as rapidly as possible	Salthouse & Babcock (1991)

and several (e.g., Allaire & Marsiske, 2005; Li, Aggen, Nesselroade, & Baltes, 2001) were based on 50 or more assessments, which is unlikely to be practical in research not explicitly focused on the issue of within-person variability. Salthouse et al. (2006) recently estimated the reliability of the within-person standard deviations across three sessions by using the standard deviations across the three possible pairs of sessions (i.e., 1 and 2, 1 and 3, 2 and 3) as items in coefficient alpha. Most of the values were in the moderate range, but the procedure may have led to overestimates of true reliability because items defined in this manner are not independent.

In the current studies, reliability was estimated by determining separate scores from the odd-numbered and even-numbered items in each test at each session, computing the standard deviations of the scores from the odd-numbered and even-numbered items across the three sessions for each individual, and then, with these two standard deviations as items in coefficient alpha, estimating the reliability of the within-person standard deviation for the entire test. Because 51 individuals had completed the same three-session cognitive test battery between 1 and 2 years earlier, stability coefficients were also computed in Study 3 to provide independent lower-bound estimates of reliability.

Although measures of within-person variability are often related to age, Salthouse and colleagues (e.g., Nesselroade & Salthouse, 2004; Salthouse & Berish, 2005; Salthouse et al., 2006) found little unique relation of within-person variability to age after controlling for the individual difference variation in the mean. Because measures of variability may not have any distinct diagnostic value if most of the individual differences in measures of variability are shared with individual differences in the individual's average level of performance, it is important that these results be replicated. That is, if all of the effects associated with measures of within-person variability are "carried" by effects associated with an individual's average level of performance, traditional measures of average performance may be sufficient for most purposes.

The third study in the current project involved a relatively small sample of 51 adults who performed the same measurement burst assessment after an interval of between 1 and 2 years. This retest interval was too short for much age-related cognitive change, but the data were nevertheless useful in examining the stability of within-person variability across the 1- to 2-year interval and in comparing reliability estimates for traditional change scores and for the proposed distribution-referenced change scores.

In summary, adults across a wide range of ages performed three different versions of each of 16 tests across three sessions within a period of about 2 weeks. In Study 1, the order of the versions was counterbalanced across participants to provide a basis for calibrating the difficulty of the test versions. In Study 2, a large sample of participants performed the versions in the same order, and in Study 3, a small sample of adults repeated the three-session assessment after an interval of approximately 1.5 years.

General Method

The tasks administered to the participants are briefly described in Table 1, with additional details contained in other recent reports (e.g., Salthouse et al., 2003, 2006; Salthouse & Ferrer-Caja, 2003). Descriptive characteristics of those sampled in the three studies, who were all recruited by newspaper advertisements, flyers, and referrals from other participants, are reported in Table 2. It can be seen that the participants in each study were fairly similar, with an average of over 15 years of education and an average level of self-rated health in the very good range. Across all studies, the mean interval from the first to the third session was 10.9 days, with a median of 7 days.

The 16 cognitive tests were administered in the same order in each session, with a session requiring between 90 and 120 min. All participants were tested individually in the laboratory, and the sessions were usually scheduled at the same time each day.

Table 2
Descriptive Characteristics of the Participants

	Age range (years)			
Study and characteristic	18–39	40–59	60–97	All
Study 1				
N	30	30	30	90
Age	23.2 (4.2)	50.3 (5.5)	69.7 (8.3)	47.8 (20.1)
% women	47	80	60	62
Self-rated health	1.8 (0.8)	2.2 (0.9)	2.4 (0.8)	2.2 (0.8)
Years of education	14.6 (1.5)	15.6 (2.8)	15.5 (2.7)	15.2 (2.4)
Study 2				
N	373	593	634	1,600
Age	26.6 (6.3)	50.6 (5.3)	71.5 (7.9)	53.3 (18.6)
% women	63	73	58	65
Self-rated health	1.6 (0.8)	1.8 (0.9)	2.0 (0.9)	1.8 (0.9)
Years of education	15.1 (2.3)	15.8 (2.6)	16.2 (2.9)	15.8 (2.7)
Study 3				
N	10	19	22	51
Age (at T1)	29.8 (6.4)	53.5 (4.3)	70.9 (7.1)	56.4 (16.5)
% women	30	68	64	59
Self-rated health (at T1)	1.7 (0.9)	2.0 (0.8)	2.3 (0.9)	2.0 (0.9)
Years of education (at T1)	16.0 (2.3)	15.7 (1.5)	14.1 (4.1)	15.1 (3.1)
Retest interval (days)	597 (157)	563 (150)	557 (157)	567 (151)

Note. Numbers in parentheses are standard deviations. Health is a self-rating on a scale ranging from 1 for *excellent* to 5 for *poor*. T1 refers to the first measurement occasion.

Table 3
Mean Levels of Performance (and Standard Deviations) Across Sessions and Age Correlations When Versions Were Presented in Counterbalanced Order, Study 1 (N = 90)

	Session 1			Session 2			Session 3		
Variable	*M*	*SD*	*r*	*M*	*SD*	*r*	*M*	*SD*	*r*
Vocabulary	54.0	9.5	.08	54.5	9.9	.07	55.5	8.4	.15
Picture vocabulary	19.0	5.4	.25	19.6	9.9	.01	20.1	5.1	.12
Synonym vocabulary	6.9	2.2	.29	6.9	2.2	.18	7.0	2.4	.24
Antonym vocabulary	5.6	2.7	.03	6.0	2.4	.21	6.2	2.6	.27
Matrix reasoning	7.7	3.6	−.57	7.7	3.8	−.56	8.3	4.5	−.64
Letter sets	10.9	2.5	−.42	11.0	2.6	−.37	11.3	2.3	−.29
Shipley	12.3	3.7	−.55	13.0	3.8	−.55	13.3	4.0	−.48
Spatial relations	9.6	3.9	−.52	10.2	4.3	−.63	10.4	4.5	−.50
Paper folding	5.9	2.9	−.60	7.0	3.1	−.57	7.0	3.3	−.57
Form boards	7.4	4.9	−.58	8.4	4.9	−.70	8.1	5.3	−.48
Recall	34.1	7.0	−.48	35.7	7.1	−.60	35.1	7.2	−.53
Logical memory	47.5	11.4	−.39	48.3	12.1	−.44	50.1	11.2	−.33
Paired associates	3.0	1.9	−.44	3.2	1.8	−.49	3.2	1.9	−.51
Digit symbol	75.5	20.8	−.67	78.3	21.6	−.64	81.5	22.4	−.71
Letter comparison	10.4	2.6	−.51	10.1	2.7	−.49	10.5	2.7	−.52
Pattern comparison	15.4	4.3	−.67	16.4	4.6	−.59	16.6	5.1	−.68

Study 1

In this study, the test versions were administered in counterbalanced order across the three sessions. Six groups of 15 participants (5 from each of the three age groups in Table 1) received the versions in each of the six possible orders (i.e., OAB, OBA, AOB, ABO, BOA, and BAO, where O refers to the original version and A and B refer to the alternate versions described in Salthouse et al., 2006).

Table 3 contains means, between-person standard deviations, and correlations between age and performance for each variable at each session.[1] It can be seen that there was an increase across sessions in the means of many of the variables, indicating better performance with more prior experience. However, the age correlations remained fairly constant across sessions, as the medians for the vocabulary variables were .17, .13, and .20, and the medians for the other variables were −.54, −.57, and −.52, for Sessions 1, 2, and 3, respectively.

Means and standard deviations of the within-person means and standard deviations across the three sessions are reported in Table 4. These values were generated by first computing a mean and a standard deviation for each participant's scores across the three sessions for each variable. The between-person means and standard deviations of these 90 within-person values were then computed and reported in columns 2 through 5 in the table. The between-person standard deviations of the within-person standard deviations, contained in the 5th column of Table 4, indicate that there were substantial individual differences in the magnitude of the estimates of within-person variability. In fact, the between-person variability in the measures of within-person variability was greater than that in the means, as the median coefficient of variation (i.e., standard deviation/mean) was .58 for the measure of within-person variability (i.e., column 5 divided by column 4) and only .27 for the mean measure (i.e., column 3 divided by column 2).

The 6th column in Table 4 contains the ratio of the mean within-person variability (i.e., column 4) to the between-person variability of the means (i.e., column 3) for that variable. These ratios ranged from .28 to .78, with a median of .54. Within-person variability is therefore relatively large for each of the variables, as it corresponds to about half the magnitude of the variability, expressed in standard deviations, apparent across people in their mean levels of performance.

The last column in Table 4 reports within-person variability in years of cross-sectional age differences (cf. Nesselroade & Salthouse, 2004; Salthouse & Berish, 2005; Salthouse et al., 2006). These values were computed by dividing the mean within-person standard deviation (i.e., column 4) by the slope derived from a linear regression equation relating score to age in the total sample of 1,600 individuals from Study 2. No values for vocabulary variables are reported, because the age relations for these variables were small and were positive rather than negative. Values for the other variables range from about 10 to 26 years, with a median of 23.1. These results therefore suggest that the average short-term fluctuation corresponds to the amount of variation associated with about 23 years of cross-sectional aging.

Correlations of age with each participant's mean and standard deviation computed across the three sessions, before and after partialling the variation in the other variable, are reported in Table 5. The second column in the table contains correlations between the within-person means and the within-person standard deviations. Of the variables, 10 had negative correlations between the mean and the standard deviation, indicating that within-person variability was smaller when the mean level of performance was

[1] Because of the large number of statistical comparisons, a significance level of .01 was used in all statistical tests.

Table 4
Means and Standard Deviations of Means and Within-Person Standard Deviations Across the Three Sessions and Two Estimates of the Magnitude of the Within-Person Variability, Study 1 (N = 90)

Variable	Mean		*SD*		*M*(*SD*)/*SD*(*M*)	Years cross-sect. age diff.
	M	*SD*	*M*	*SD*		
Vocabulary	54.7	8.6	3.6	2.5	0.42	NA
Picture vocabulary	19.6	4.9	2.1	1.2	0.43	NA
Synonym vocabulary	6.9	2.0	1.1	0.6	0.55	NA
Antonym vocabulary	6.0	2.2	1.4	0.8	0.64	NA
Matrix reasoning	7.9	3.7	1.6	0.8	0.43	14.5
Letter sets	11.1	2.1	1.3	0.8	0.62	24.0
Shipley	12.9	3.5	1.7	0.9	0.49	19.3
Spatial relations	10.1	3.7	2.4	1.2	0.65	23.7
Paper folding	6.6	2.7	1.7	1.0	0.63	24.9
Form boards	7.9	4.0	3.1	2.1	0.78	26.3
Recall	35.0	6.4	3.5	1.8	0.55	24.1
Logical memory	48.6	10.2	5.9	3.6	0.58	41.1
Paired associates	3.1	1.7	0.9	0.6	0.53	22.4
Digit symbol	78.4	21.0	5.8	3.6	0.28	9.8
Letter comparison	10.3	2.4	1.2	0.7	0.50	17.9
Pattern comparison	16.1	4.3	2.1	1.1	0.49	18.7

Note. Every participant had a mean and a standard deviation for each variable across the three sessions. The second and third columns contain the between-person means and standard deviations of the means, and the fourth and fifth columns contain the between-person means and standard deviations of the standard deviations. The sixth column is the ratio of the average within-person standard deviation (column 4) to the between-person standard deviation of the mean (column 3), and the last column expresses the average within-person standard deviation in years of cross-sectional age differences.

higher. It is interesting that this pattern is different from that typically found with reaction time variables, in which there is usually a strong positive correlation between the mean and measures of variability.

With the exception of the vocabulary variables, all of the correlations between age and the means were negative, and there was only a slight reduction in their magnitude after controlling for the variation in the standard deviation (i.e.,

Table 5
Correlations of Mean and Within-Person Standard Deviation With Age, Study 1 (N = 90)

Variable	Correlation *M.SD*	Age correlations			
		M	*M.SD*	*SD*	*SD.M*
Vocabulary	−.55	.11	.14	.07	.13
Picture vocabulary	−.15	.13	.13	−.02	.00
Synonym vocabulary	−.16	.26	.26	−.04	.00
Antonym vocabulary	−.06	.18	.18	−.14	−.13
Matrix reasoning	.22	−.63	−.61	−.14	.00
Letter sets	−.25	−.41	−.40	.05	−.06
Shipley	−.15	−.57	−.56	.02	−.00
Spatial relations	.06	−.63	−.62	−.09	−.07
Paper folding	−.19	−.66	−.64	.18	.07
Form boards	.62	−.69	−.44	−.50	−.10
Recall	−.38	−.60	−.53	.25	.03
Logical memory	−.22	−.43	−.43	.01	−.09
Paired associates	−.08	−.54	−.55	−.08	−.15
Digit symbol	.13	−.70	−.68	−.19	−.14
Letter comparison	.18	−.55	−.53	−.10	−.00
Pattern comparison	.20	−.69	−.66	−.30	−.22

Note. *M.SD* refers to the correlation with the mean after statistically controlling the variance in the standard deviation, and *SD.M* refers to the correlation with the standard deviation after statistically controlling the variance in the mean.

M.SD). Medians for the 13 nonvocabulary variables were −.62 for raw correlations of the means with age and −.56 for the age–mean correlations after controlling for the variation in the standard deviation. However, most of the age–standard deviation correlations were close to zero, with many of them negative, indicating smaller short-term variability with increased age. Furthermore, there was very little change in the magnitude of the correlations between age and the standard deviation after controlling for the variation in the mean (i.e., *SD.M*). Medians for the 13 nonvocabulary variables were −.09 for raw correlations and −.07 for correlations after controlling for the variation in the mean.

Table 6 contains reliabilities of the variables at each session (or version) and reliabilities of the across-session (within-person) means and standard deviations. The first value in each cell is the estimate from Study 1, and the second is the estimate from Study 2. Reliability estimates for the variables at each session (or version) were derived from coefficient alpha on the basis of items within each test. Most variables had good reliability, with the exception of the Version B synonym vocabulary, antonym vocabulary, and letter sets variables.

Estimates of the reliability of the across-session means and standard deviations were obtained by determining scores for the odd-numbered and even-numbered items in each test in each session, computing means and standard deviations of the scores for the odd-numbered items and for the even-numbered items across the three sessions, and then estimating the reliability of the parameters across the three sessions with those values as items in coefficient alpha. Inspection of the values in columns 5 and 6 of Table 6 reveals that the estimated reliabilities of the means were all quite high, with a median in Study 1 of .94, but the estimated reliabilities of the standard deviations were much lower, with a Study 1 median of only .26.

Finally, regression equations were created to predict the score on the O version of each test from the scores on the A and B versions. The median R^2 in these equations was .57, indicating moderately strong relations between the levels of performance across versions. The intercept and slope parameters from these equations were used in Study 2 to equate the average performance across versions. To illustrate, the intercept and the slope of the equation predicting the O word-recall score from the A word-recall score were 15.21 and 0.61, respectively, and thus the adjusted A score for an observed A score of 30 would be 33.5 (i.e., 15.21 + [0.61 × 30]).

Study 2

The participants in Study 2 all received the three test versions in the OAB order. Data from 143 of these participants were entered in the analyses reported in the Salthouse et al. (2006) study. These participants were not administered three of the tests, Letter Sets (Ekstrom, French, Harman, & Dermen, 1976), Shipley Abstraction (Zachary, 1986), and Form Boards (Ekstrom et al., 1976), and thus the sample size for these variables was only 1,457. All of the analyses reported below are based on adjusted scores for the A and B versions computed with the regression-based method described above.

Results and discussion. Table 7 contains means and between-person standard deviations by session for the 16 cognitive variables, as well as the correlations with age. The overall pattern is very similar to that from Study 1, as reported in Table 3. Specifically, there were slight increases in many of the means across sessions, but the age correlations for most variables were similar in each session. The median age correlations for the vocabulary variables were .23, .07, and .13 for Sessions 1, 2, and 3, respec-

Table 6
Estimated Reliabilities of Single Assessments and of the Across-Session Means and Standard Deviations

	(Version)/session						Across-session			
	(O)1		(A)2		(B)3		*M*		*SD*	
Variable	Study 1	Study 2	Study 1	Study 2	Study 1	Study 2	Study 1	Study 2	Study 1	Study 2
Vocabulary	.92	.88	.91	.83	.90	.85	.97	.95	.65	.47
Picture vocabulary	.91	.85	.85	.80	.84	.81	.94	.94	.15	.35
Synonym vocabulary	.78	.82	.70	.65	.61	.59	.90	.87	−.11	.28
Antonym vocabulary	.81	.79	.80	.60	.62	.59	.90	.85	.25	.10
Matrix reasoning	.85	.82	.86	.81	.85	.85	.94	.92	.26	.08
Letter sets	.73	.74	.70	.68	.45	.60	.86	.86	−.02	.22
Shipley	.87	.84	.87	.81	.90	.83	.95	.94	.24	.08
Spatial relations	.87	.90	.73	.70	.75	.70	.92	.91	.28	.23
Paper folding	.77	.73	.73	.68	.81	.80	.93	.85	.38	.08
Form boards	.92	.87	.92	.90	.76	.72	.95	.93	.83	.57
Recall	.92	.89	.95	.92	.92	.91	.97	.96	.57	.54
Logical memory	.73	.84	.76	.85	.72	.83	.85	.82	.17	.07
Paired associates	.86	.83	.84	.83	.82	.87	.94	.91	.52	.39
Letter comparison	.83	.87	.90	.84	.79	.84	.94	.95	.34	.26
Pattern comparison	.93	.86	.87	.89	.92	.90	.97	.92	.02	.44

Note. Reliability estimates from the individual (versions) sessions were based on coefficient alpha across individual items, and those across sessions correspond to coefficient alpha for the scores based on odd-numbered and on even-numbered items in each session. No reliability estimates were available for the digit symbol variable, because it was based on a single timed score in each session.

Table 7
Mean Levels of Performance (and Standard Deviations) Across Sessions After Adjusting for Version Differences, Study 2 (N = 1,600, 1,457), and Correlations With Age

Variable	Session 1			Session 2			Session 3		
	M	*SD*	*r*	*M*	*SD*	*r*	*M*	*SD*	*r*
Vocabulary	52.4	9.5	.06	52.4	6.8	.04	52.3	7.1	.13
Picture vocabulary	19.2	5.0	.27	18.5	3.8	.01	18.6	4.4	.10
Synonym vocabulary	7.3	2.7	.28	7.1	2.1	.19	7.1	2.0	.24
Antonym vocabulary	6.7	2.8	.18	6.6	1.7	.10	6.8	1.9	.12
Matrix reasoning	7.8	3.5	−.56	8.0	3.0	−.53	7.9	2.9	−.57
Letter sets	11.2	2.8	−.36	11.2	2.4	−.27	11.4	2.1	−.26
Shipley abstraction	13.4	3.6	−.46	13.4	2.8	−.46	13.7	2.7	−.49
Spatial relations	8.8	5.0	−.38	10.00	3.3	−.32	10.6	3.7	−.37
Paper folding	6.1	2.7	−.47	7.0	1.7	−.41	7.3	1.4	−.43
Form boards	7.1	4.5	−.48	7.5	3.2	−.52	8.4	2.9	−.39
Recall	35.1	6.4	−.42	35.8	4.4	−.49	35.9	5.0	−.44
Logical memory	44.6	9.7	−.28	44.5	7.9	−.35	45.7	6.6	−.30
Paired associates	3.1	1.8	−.41	3.3	1.2	−.39	3.4	1.3	−.43
Digit symbol	72.6	18.4	−.60	77.6	18.3	−.47	78.9	18.5	−.48
Letter comparison	10.4	2.5	−.51	10.8	2.2	−.48	10.9	2.2	−.48
Pattern comparison	15.6	3.8	−.55	16.4	3.9	−.58	17.0	3.9	−.59

Note. The Letter Sets, Shipley, and Form Boards tests were not administered to all participants, and thus the sample size for these variables was 1,457 instead of 1,600.

tively, and those for the other variables were −.47, −.47, and −.44 for Sessions 1, 2, and 3, respectively.

Means and standard deviations of the within-person means and standard deviations across the three sessions are reported in Table 8. Also in this table are ratios of the mean standard deviation (average within-person variability) to the standard deviation of the mean (between-person variability). Once again, the patterns in these data were very similar to those from Study 1 (cf. Table 4). Specifically, the median coefficient of variation for the within-person standard deviation was larger than that for the mean (i.e., .63 vs. .22), the median ratio of within-person variability to between-person variability was .57, and the average level of within-

Table 8
Means and Standard Deviations of Means and Within-Person Standard Deviations Across the Three Sessions and Two Estimates of the Magnitude of the Within-Person Variability, Study 2

Variable	Mean		*SD*		*M(SD)/SD(M)*	Years cross-sect. age diff.
	M	*SD*	*M*	*SD*		
Vocabulary	52.4	7.1	3.4	2.4	0.48	NA
Picture vocabulary	18.7	4.1	2.0	1.3	0.49	NA
Synonym vocabulary	7.2	1.9	1.3	0.8	0.68	NA
Antonym vocabulary	6.7	1.8	1.3	0.8	0.72	NA
Matrix reasoning	7.9	2.9	1.4	0.8	0.48	12.9
Letter sets	11.2	2.0	1.3	1.0	0.65	25.2
Shipley	13.5	2.7	1.3	0.9	0.48	15.1
Spatial relations	9.8	3.5	2.3	1.3	0.66	22.6
Paper folding	6.8	1.7	1.2	0.7	0.71	18.0
Form boards	7.6	3.1	2.1	1.2	0.68	17.6
Recall	35.6	4.7	2.6	1.7	0.55	17.9
Logical memory	44.9	7.1	4.2	2.5	0.59	29.4
Paired associates	3.3	1.3	0.8	0.5	0.62	19.7
Digit symbol	76.3	16.5	8.2	6.9	0.50	13.8
Letter comparison	10.7	2.1	1.0	0.8	0.48	14.5
Pattern comparison	16.3	3.4	1.6	1.0	0.47	13.9

Note. Every participant had a mean and a standard deviation for each variable across the three sessions. The second and third columns contain the between-person means and standard deviations of the means, and the fourth and fifth columns contain the between-person means and standard deviations of the standard deviations. The sixth column is the ratio of the average within-person standard deviation (column 4) to the between-person standard deviation of the mean (column 3), and the last column expresses the average within-person standard deviation in years of cross-sectional age differences. The Letter Sets, Shipley, and Form Boards tests were not administered to all participants, and thus the sample size for these variables was 1,457 instead of 1,600.

person variability corresponded to a median of 17.8 years of cross-sectional age difference. Reliabilities of individual variables and of across-session means and standard deviations, reported in Table 6, were also similar to those of Study 1, with highly reliable means but weak reliability for the standard deviations.

Table 9 contains relations of age to the within-person means and standard deviations before and after controlling for the variation in the other variable. It is apparent that these results closely resemble those from Study 1 (cf. Table 5), as the median age correlations for the variables other than vocabulary were −.51 and −.47 for the means, before and after controlling for the variation in the standard deviations, and .03 and −.06 for the standard deviations, before and after controlling for the variation in the means.

In summary, there were three major results of Study 2. First, there was a substantial amount of variation in performance of the same test from one session to the next, and there were large individual differences in the magnitude of this variation. Second, the measures of within-person variability were not very reliable, particularly compared with measures of the level of performance at each session or with the average level across sessions. And third, there were very few unique relations of age with the measures of within-person variability that were independent of influences of the mean. In each respect these results closely resemble those of Study 1, despite differences in the sizes of the samples (i.e., 90 in Study 1 and 1,600 in Study 2) and in the method of equating the test versions (i.e., counterbalancing across participants in Study 1 and statistical adjustment in Study 2).

Study 3

Because 51 adults had performed 13 of the tests in three sessions an average of 1.5 years earlier, the data from these individuals were used to investigate the stability of the estimates of within-person variability. The first occasion is labeled T1 (for Time 1) and the second is labeled T2 (for Time 2). The major results relevant to this question are summarized in Table 10. It can be seen that there were similar relations between age and the across-session means at both the first and the second occasion. However, most of the age relations on the across-session standard deviations were weak and were inconsistent from the first to the second occasion. Stability coefficients for the means across the retest interval were all, with one exception, greater than .8, and the median was .88. In contrast, only two variables had stability coefficients for the within-person standard deviations greater than .5, and the median was only .13.

The availability of longitudinal data from the measurement burst design also allowed the reliability of two types of change scores to be assessed. The traditional method of assessing change is the simple difference between the scores on two occasions, whereas the distribution-based method illustrated in Figure 1 represents change as the difference between the means of the distributions at each occasion divided by the pooled standard deviation. That is, for each individual the equivalent of a *t* test was conducted, in which the three scores at the first occasion were contrasted with the three scores at the second occasion, with the resulting *t* value serving as the measure of change. Each type of change was computed for the 13 variables available from every participant who had completed two measurement burst assessments.

The reliability estimate for the traditional change score was obtained by creating separate *T2–T1* difference scores for odd-numbered and even-numbered items from the first session at each occasion and then using the two difference scores as items in coefficient alpha. The reliability estimate for the distribution-based measure involved a similar procedure with two distribution-referenced differences, one based on odd-numbered items and one based on even-numbered items, serving as the items in coefficient alpha.

Table 9
Correlations of Mean and Within-Person Standard Deviation With Age, Study 2 (N = 1,600, 1,457)

Variable	Correlation *M.SD*	Age correlations *M*	*M.SD*	*SD*	*SD.M*
Vocabulary	−.45	.08	.09	.03	.06
Picture vocabulary	−.26	.15	.17	.07	.11
Synonym vocabulary	−.42	.28	.24	−.10	.02
Antonym vocabulary	−.32	.16	.17	.03	.08
Matrix reasoning	−.11	−.60	−.61	−.02	−.11
Letter sets	−.47	−.35	−.31	.08	−.08
Shipley	−.37	−.51	−.47	.15	−.05
Spatial relations	−.22	−.40	−.39	.04	−.05
Paper folding	−.42	−.50	−.42	.20	−.00
Form boards	.12	−.53	−.52	−.11	−.06
Recall	−.33	−.50	−.45	.16	−.01
Logical memory	−.23	−.35	−.35	−.02	−.11
Paired associates	−.11	−.47	−.47	−.00	−.06
Digit symbol	−.02	−.57	−.57	.05	.04
Letter comparison	.05	−.53	−.54	.01	.04
Pattern comparison	.04	−.62	−.62	−.14	−.15

Note. The Letter Sets, Shipley, and Form Boards tests were not administered to all participants, and thus the sample size for these variables was 1,457 instead of 1,600.

Table 10
Summary Statistics on 13 Variables With a Retest Interval of 1 or 2 Years, Study 3 (N = 51)

Variable	Age correlations		Stability coefficients	
	*M*1/*M*2	*SD*1/*SD*2	*M*1, *M*2	*SD*1, *SD*2
Vocabulary	−.12/−.13	.19/.19	.91	.71
Picture vocabulary	.08/−.01	−.38/.10	.89	.17
Synonym vocabulary	.21/.13	−.15/−.07	.89	.56
Antonym vocabulary	.02/−.02	−.05/.05	.83	.32
Matrix reasoning	−.59/−.64	.01/.01	.90	.13
Spatial relations	−.57/−.52	.27/.07	.92	−.02
Paper folding	−.48/−.53	.10/.10	.82	.19
Recall	−.45/−.56	.05/.23	.90	.17
Logical memory	−.38/−.36	.01/.11	.80	−.04
Paired associates	−.49/−.53	−.00/.11	.86	.05
Digit symbol	−.62/−.50	−.04/.18	.56	−.05
Letter comparison	−.54/−.61	.02/−.03	.88	.00
Pattern comparison	−.61/−.66	−.01/−.02	.87	−.01

	M and *SD* and reliability of change (*T2–T1*)					
	First assessment			Distribution referenced		
	M	*SD*	*Est. rel.*	*M*	*SD*	*Est. rel.*
Matrix reasoning	0.37	2.26	.18	0.20	0.79	.86
Spatial relations	1.96	2.21	−.24	0.39	0.58	.13
Paper folding	0.43	2.03	.13	0.12	0.70	.44
Recall	0.12	4.60	.67	−0.10	0.69	.67
Logical memory	2.10	7.89	.66	0.16	0.83	.58
Paired associates	−0.04	1.33	.41	−0.05	0.78	.51
Letter comparison	−0.18	1.73	.47	−0.18	0.88	.67
Pattern comparison	−0.32	2.72	.62	0.08	0.76	.62
Digit symbol	1.45	8.90		−0.14	1.02	

Note. Est. rel. = estimates of reliability.

Mean changes and estimated reliabilities computed in this manner are reported in the bottom of Table 10. Most of the changes were small, but many were positive, indicating a slightly higher level of performance in the second occasion compared with the first occasion. It is noteworthy that the estimates of reliability were at least as high for the distribution-based change measures as for the traditional change measures, as the medians were .44 for the traditional change scores, on the basis of the first session at each occasion, and .59 for the distribution-based scores.

General Discussion

The variation in performance of essentially the same test across three measurement sessions was found to be moderately large for a variety of different cognitive and neuropsychological variables, with an average magnitude corresponding to about 50% of the between-person standard deviation for the variable. Because the cross-sectional age differences for many cognitive variables, with the notable exception of vocabulary measures of acquired knowledge, range from about .02 to .03 standard deviation units per year (e.g., Salthouse, 2004), this level of within-person variability corresponds to the difference that would be expected across an age range of about 10 to 29 years (cf. Tables 4 and 8; Nesselroade & Salthouse, 2004; Salthouse et al., 2006).

Figure 2 summarizes the preceding information for the averages of the scores on variables representing four different cognitive abilities. The ordinate in the figure consists of units of between-person standard deviations; the first panel portrays the average within-person variability from Study 1, the second panel portrays the average within-person variability from Study 2, and the third panel portrays the differences expected over a 10-year interval based on the cross-sectional age trends. This figure clearly indicates that the magnitude of within-person variability is large, both in comparison with the variability apparent across people and relative to the cross-sectional age-related differences expected across a 10-year interval.

The existence of large within-person variability for parallel versions of the same test implies that there could be considerable imprecision in evaluations based on single assessments. Some indication of the degree of imprecision can be obtained by multiplying the average within-person standard deviation for a variable by 1.96 to determine the 95% confidence interval around the average. To express this information in a more meaningful form, the researcher can convert the confidence intervals to age-adjusted

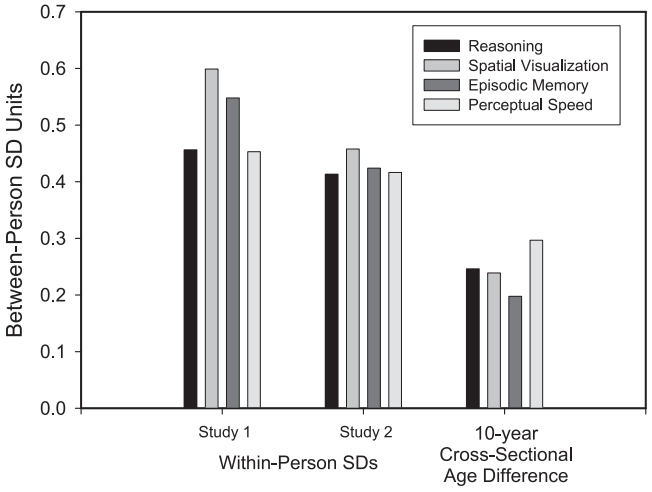

Figure 2. Average levels of within-person variability and of 10-year cross-sectional age differences for four cognitive abilities scaled in between-person standard deviation units.

scaled scores for variables from the Wechsler Memory Scale (Wechsler, 1997b). As an illustration, for the word recall variable, the mean at age 50 in the normative sample was 32, and the average within-person standard deviation was 3.5 in Study 1 and 2.6 in Study 2. These values yield 95% confidence intervals of 25 to 39 in Study 1 and 27 to 37 in Study 2, which correspond to scaled score ranges of 6 to 14 in Study 1 and 7 to 13 in Study 2. Because Wechsler-scaled scores have a mean of 10 and a standard deviation of 3, this level of short-term fluctuation corresponds to a range of between two and three standard deviations in scaled score units. The clear implication from these results is that evaluations based on a single assessment can be very imprecise.

It should be noted that moderately large within-person variability does not necessarily imply that retest reliability will be low. In fact, test–retest correlations will likely be high if the within-person fluctuations are small relative to differences in the average level of performance between people. However, retest reliability could be very low for people with high levels of within-person variability, and in some circumstances these individuals may be the ones for whom accurate classification is most critical.

Another noteworthy finding of these studies was that the measures of within-person variability were not very reliable. At least with the particular cognitive and neuropsychological variables examined in this project, within-person variability based on three measurement assessments does not appear meaningful as a distinct individual difference variable with unique predictive power. Reliability of the within-person variability measures would likely be increased by adding more assessments at each occasion, but it is important to realize that this additional experience could change the nature of what is being evaluated and would greatly increase the burden on the examinees.

There were very few relations of age to the measures of within-person variability that were independent of variations in the means. Some of the lack of unique age relations may be attributable to low reliability of the within-person variability measures. However, additional analyses were carried out with latent constructs created with the scores from odd- and even-numbered items as the manifest variables. Although at least theoretically the focus on the variance shared across variables minimizes measurement error and increases reliability, the age relations on the measures of within-person variability in these latent construct analyses were still very small.

It is important to emphasize that even if the reliability of the measures of within-person variability is low, the existence of moderate variability could have considerable practical significance, because it contributes to imprecise assessment and implies that detecting true longitudinal change may be difficult in the presence of this substantial short-term fluctuation. With respect to this latter point, regardless of whether the amount of fluctuation from one assessment to the next is an enduring characteristic of an individual, some of what is interpreted as long-term change may actually be a reflection of short-term fluctuation in performance. One solution to these problems is to rely on the use of measurement burst designs, in which there are multiple assessments at each occasion, and to then evaluate across-occasion change relative to each individual's distribution of scores within the occasions. This method of calibrating change is particularly desirable when there is evidence, such as that reported in the current studies, that the average within-person variability is relatively large, and that it varies in magnitude across people. Under circumstances such as these, a larger absolute change is needed to have the same meaning for someone who has more short-term variability. Expressing change in terms of each individual's own variability takes this into consideration in a manner analogous to the computation of an effect size.

Distribution-referenced scores are frequently used in psychology, because most measurement units are somewhat arbitrary; consequently, ratios or percentages that require ratio level measurement are seldom meaningful. Much of the past research on change has relied on the distribution of scores from different people (i.e., between-person variability) as the basis for evaluating the magnitude of change (e.g., Frerichs & Tuokko, 2005; Temkin, Heaton, Grant, & Dikmen, 1999). In some cases the reference scores have been measures of performance at the first occasion, and in other cases they have been the differences between single assessments of performance across two occasions. However, within-person change has typically been evaluated relative to the differences or changes in other people, and yet there is no necessary relation between the variability apparent across different people and the variability exhibited by a given individual across different assessments.

Scaling of scores based on distributions is necessarily somewhat sample specific, because the reference distributions can vary across samples. The amount of specificity would be even greater with the current proposal, because not only would the distributions differ across people but they could conceivably also differ across occasions, as the reliability of within-person variability is not very high. In other words, the change from the first to the second occasion and from the second to the third occasion would be scaled differently not only across people but possibly also within the same individual if the distributions of scores varied across occasions. Nevertheless, the proposed procedure has the advantages of distinguishing short-term fluctuation from true change and of taking individual differences in the magnitude of the fluctuation into account in the analysis of change. Furthermore, this is apparently the only method that can be used to evaluate the statistical significance of change within a single individual without reference to information from other individuals.

The discovery that performance exhibits considerable variation from one assessment to the next in nearly identical tests suggests that it may be useful to consider a reconceptualization of cognitive abilities. That is, it may be more meaningful to think of an individual's cognitive ability as consisting of a distribution of many potential levels of performance rather than as corresponding to a single discrete level that is highly stable over short intervals. The performance observed in a given assessment may therefore represent one sample from the distribution, but other assessments need not have the same value, because they would correspond to different samples from that distribution.

The relatively large amount of variability over short intervals observed in the current studies is surprising, given that the participants were generally healthy and functioning at high levels. Problems of imprecise assessment and ambiguity in the interpretation of change would likely be even greater among certain clinical groups for whom short-term variability might be larger. What can be done to minimize these problems? The approach advocated here is to adopt a measurement burst procedure, in which the individuals receive several parallel assessments within a relatively short period of time. Not only would the multiple assessments provide a more stable evaluation for conventional comparisons through the principle of aggregation, but the across-assessment variation in performance could be used as the basis for calibrating change and could allow true change to be distinguished from short-term fluctuation.

In conclusion, the current studies add to the evidence that there is considerable within-person variability in many different cognitive and neuropsychological variables, and that there appear to be large individual differences in the magnitude of this variability. However, they also reveal that within-person variability does not appear to be a stable and enduring (i.e., reliable) characteristic of the individual, and that it has few unique relations to age. Nevertheless, the existence of within-person variability complicates the assessment of cognitive and neuropsychological functioning and raises the possibility that single measurements may not be sufficient for precise evaluations of individuals, or for sensitive detection of change.

References

Allaire, J. C., & Marsiske, M. (2005). Intraindividual variability may not always indicate vulnerability in elders' cognitive performance. *Psychology and Aging, 20,* 390–401.

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

Burton, C. L., Strauss, E., Hultsch, D. F., Moll, A., & Hunter, M. A. (2006). Intraindividual variability as a marker of neurological dysfunction: A comparison of Alzheimer's disease and Parkinson's disease. *Journal of Clinical and Experimental Neuropsychology, 28,* 67–83.

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.

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

Hultsch, D. F., MacDonald, S. W., Hunter, M. A., Levy-Bencheton, J., & Strauss, E. (2000). Intraindividual variability in cognitive performance in older adults: Comparison of adults with mild dementia, adults with arthritis, and healthy adults. *Neuropsychology, 14,* 588–598.

Li, S. C., Aggen, S. H., Nesselroade, J. R., & Baltes, P. B. (2001). Short-term fluctuations in elderly people's sensorimotor functioning predict text and spatial memory performance: The MacArthur successful aging studies. *Gerontology, 47,* 100–116.

Murtha, S., Cismaru, R., Waechter, R., & Chertkow, H. (2002). Increased variability accompanies frontal lobe damage in dementia. *Journal of the International Neuropsychological Society, 8,* 360–372.

Nesselroade, J. R. (1991). The warp and woof of the developmental fabric. In R. Downs, L. Liben, & D. Palermo (Eds.), *Views of development, the environment, and aethestics: The legacy of Joachim F. Wohlwill* (pp. 213–240). Hillsdale, NJ: Erlbaum.

Nesselroade, J. R., & Salthouse, T. A. (2004). Methodological and theoretical implications of intraindividual variability in perceptual–motor performance. *Journal of Gerontology: Psychological Science, 59B,* P49–P55.

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

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). What and when of cognitive aging. *Current Directions in Psychological Science, 13,* 140–144.

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. (2005). Correlates of within-person (across-occasion) variability in reaction time. *Neuropsychology, 19,* 77–87.

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., Kausler, D. H., & Saults, J. S. (1986). Groups versus individuals as the comparison unit in cognitive aging research. *Developmental Neuropsychology, 2,* 363–372.

Salthouse, T. A., Nesselroade, J. R., & Berish, D. E. (2006). Short-term variability and the calibration of change. *Journal of Gerontology: Psychological Sciences, 61,* P144–P151.

Shipley, B. A., Der, G., Taylor, M. D., & Deary, I. J. (2006). Cognition and all-cause mortality across the entire adult age range: Health and Lifestyle Survey. *Psychosomatic Medicine, 68,* 17–24.

Strauss, E., MacDonald, S. W. S., Hunter, M., Moll, A., & Hultsch, D. F. (2002). Intraindividual variability in cognitive performance in three groups of older adults: Cross-domain links to physical status and self-perceived affect and beliefs. *Journal of the International Neuropsychological Society,* 893–906.

Temkin, N. R., Heaton, R. K., Grant, I., & Dikmen, S. S. (1999). Detecting significant change in neuropsychological test performance: A comparison of four models. *Journal of the International Neuropsychological Society, 5,* 357–369.

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

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

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

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

Received April 28, 2006
Revision received January 8, 2007
Accepted January 25, 2007 ■