Journal of Gerontology: PSYCHOLOGICAL SCIENCES
2006, Vol. 61B, No. 3, P144–P151

Short-Term Variability in Cognitive Performance and the Calibration of Longitudinal Change

Timothy A. Salthouse, John R. Nesselroade, and Diane E. Berish

Department of Psychology, University of Virginia, Charlottesville.

Recent studies have documented that normal adults exhibit considerable variability in cognitive performance from one occasion to another. We investigated this phenomenon in a study in which 143 adults ranging from 18 to 97 years of age performed different versions of 13 cognitive tests in three separate sessions. Substantial within-person variability was apparent across 13 different cognitive variables, and there were also large individual differences in the magnitude of within-person variability. Because people differ in the amount of short-term variability, we propose that this variability might provide a meaningful basis for calibrating change in longitudinal research. Correlations among the measures of within-person variability were very low, even after we adjusted for reliability, and there was little evidence that increased age was associated with a larger amount of within-person variability.

IT IS often assumed, at least implicitly, that people can be characterized as having a fixed level of a cognitive ability that can be accurately evaluated with a single assessment. However, to the extent that performance on cognitive tasks varies from one occasion to the next, as has been reported in numerous recent studies (e.g., Hertzog, Dixon, & Hultsch, 1992; Hultsch, MacDonald, Hunter, Levy-Bencheton, & Strauss, 2000; Li, Aggen, Nesselroade, & Baltes, 2001; Nesselroade & Salthouse, 2004; Rabbitt, Osman, Moore, & Stollery, 2001; Rapport, Brines, Axelrod, & Thiesen, 1997; Salinsky, Storzbach, Dodrill, & Binder, 2001; Salthouse & Berish, 2005; Shamni, Bosman, & Stuss, 1998), single assessments may represent only one of many possible levels of performance that could have been observed for that individual, and hence they are potentially misleading (cf. Hultsch & MacDonald, 2004; Nesselroade, 1991).

Much of the prior research on within-person variability has focused on reaction time or other speed variables. These types of variables are often used as measures of transient states of arousal or alertness, and thus it is not surprising to find that they exhibit within-person variability. Only a few studies have investigated within-person variability with cognitive variables measured in terms of accuracy rather than in time or speed. In one such study, Hultsch and colleagues (2000) reported results on two memory tasks (word recognition and story recognition) across four occasions from 15 healthy older adults and 30 adults with dementia or osteoarthritis. In another study, Li and associates (2001) examined performance on three memory tasks (digit span, text memory, and spatial recognition) in 25 older adults across 25 sessions. Both studies reported substantial across-occasion variability for the measures of memory accuracy. However, in neither study was there any mention of how the test versions administered on different occasions were equated for difficulty, and therefore it is possible that some of the within-person (across-occasion) variability in those studies was attributable to differences in the difficulty of the versions. Only the study by Li and associates provided information about the reliability of the measures of within-person variability of cognitive performance, and the estimates were disappointingly low. That is, these researchers reported correlations between the variabilities computed across the first and the second half of the occasions, and between the variabilities computed across the odd- and even-numbered occasions, but both sets of correlations were quite low (i.e., $Mdn = 0.20$ and $Mdn = 0.32$, respectively). Finally, another limitation of the previous studies is that the sample sizes of normal healthy adults were fairly small, and only a narrow range of ages was represented.

We designed the study reported here to address these limitations by using a moderately large sample of adults ($N = 143$) across a wide age range (18 to 97 years), who each performed a battery of 13 different cognitive tests on multiple occasions. Different test versions were administered on each occasion, but we carried out adjustments for version differences on the basis of data from another group of individuals who performed the versions in a counterbalanced order across sessions.

There were two major issues of interest in this study. One concerned the magnitude of within-person variability in different measures of cognitive performance, and the implications that short-term variability might have for the interpretation of longitudinal change. Of particular interest is whether longitudinal change either could be difficult to detect or could be confused with short-term fluctuation, if the magnitude of short-term within-person variability is large relative to any systematic within-person change that may be occurring. To the extent that this might be the case, it is worth considering whether an individual's within-person variability might be used to help calibrate the magnitude of his or her change (Salthouse, Kausler, & Saults, 1986). Because some of the participants in this study had performed several of the tasks 3 years earlier, we were able to compare different ways of evaluating longitudinal change.

The second major issue concerned the nature of within-person variability, particularly whether it is merely random noise or is a reflection of a meaningful individual difference characteristic that might be uniquely informative about the performance capabilities of the individual. One type of information relevant to this issue is the reliability of the within-person variability measures, because only if they were reliable would it be useful to characterize people as systematically

Table 1. Characteristics of Participants

	Age Group						
	18–39		40–59		60–97		Age
Variable	*M*	*SD*	*M*	*SD*	*M*	*SD*	Correlation
N	38	—	49	—	59	—	—
Age	27.5	5.7	51.9	4.9	70.8	9.1	—
Proportion female	.58	—	.69	—	.48	—	−.10
Education	15.2	2.0	16.1	2.0	15.4	3.0	.03
Health	1.6	0.8	1.8	0.8	2.2	0.9	.30*
Anxiety	13.4	3.3	12.3	2.6	12.1	2.6	−.19
Scaled scores							
Vocabulary	13.5	2.3	12.9	2.3	12.8	2.9	−.18
Digit symbol	12.3	2.3	12.8	2.8	12.1	3.4	.00
Logical memory	11.5	2.7	12.1	2.0	12.0	3.0	.11
Word recall	13.5	3.5	13.6	3.1	13.6	3.3	−.02

Notes: Education is reported in years, and health was a self-rating on a scale ranging from 1 = excellent to 5 = poor. Anxiety is the average state anxiety score (Spielberger et al., 1970) across the three occasions. The scaled scores were based on the age-adjusted values in the WAIS III (Wechsler, 1997a) and WMS III (Wechsler, 1997b) that are scaled to have a mean of 10 and a standard deviation of 3.

*$p < .01$.

differing in their degree of across-occasion variability. A second type of relevant information is the magnitude of correlations among measures of within-person variability for different cognitive variables. The rationale is that if the correlations were found to be moderately high, then the influences contributing to variability are unlikely to reflect measurement error or determinants that are specific to a particular variable.

When investigating within-person variability, researchers need to consider two aspects: the number of occasions of measurement and the type of assessment in each occasion. For the first aspect, a minimum of two occasions is needed to evaluate within-person variability, but if a researcher is interested in attempting to determine the asymptotic level of performance, then 50 or more sessions may be required (e.g., Kliegl, Smith, & Baltes, 1989; Salthouse & Somberg, 1982). The optimum number of occasions obviously depends on the specific goals of the study, because if a researcher is interested in decomposing the variability in terms of factors related to learning, or to cyclical variations, then he or she will need a relatively large number of occasions. However, pragmatic considerations usually mean that there is a trade-off between the number of occasions available from each individual and the number of individuals with some estimate of within-person variability. Because our earlier work (Nesselroade & Salthouse, 2004) indicated that three occasions provided sufficient information to warrant comparisons of individual differences in measures of within-person variability, we had each participant in the current study perform the tests on three occasions.

The second aspect that researchers need to consider in studies of within-person variability is how they can be confident that the assessments on different occasions are equivalent, and that any variation observed from one occasion to the next is not attributable to differences in the particular items or versions that are administered on a given occasion. This is seldom a concern with reaction-time tasks and certain memory tasks in which the trials or items are either identical or very similar, and are sampled randomly within and across occasions. However, version equivalence is more of a concern with tests of other types of cognitive abilities, and there are a number of possible solutions to this problem. One approach is to use exactly the same version of the tests on each occasion. Although this obviously eliminates version differences, it may lead to an underestimate of within-person variability if at least some of the performance on later occasions is determined by one's memory for specific items from earlier occasions, or to an overestimate if awareness of the repeated items leads to feelings of resentment or boredom. Another approach is to use different versions of the tests on each occasion, and merely assume that they are equivalent without any explicit evaluation. However, to the extent that the versions are not truly equivalent, some of the observed variability may be attributable to version differences rather than to fluctuations within the individual. What might be the ideal approach would be to use different versions on each occasion that have been precisely equated by means of a procedure such as item response theory. Although this method would ensure that none of the across-occasion variation is attributable to differences in the tests, researchers would need a considerable amount of data to determine the characteristics of each item prior to conducting the research of primary interest. A compromise approach, and the one we adopt in this study, is to use different versions of the tests on each occasion, but to collect data from another sample of individuals who performed the versions in counterbalanced order to allow the difficulty levels of the different versions to be adjusted statistically.

Methods

Participants

Characteristics of the 143 participants, divided into three age groups for ease of description, are summarized in Table 1. An inspection of the table reveals that increased age was associated with slightly poorer self-ratings of health, but that there was no relation of age to the scaled scores for vocabulary, digit symbol, logical memory, or word recall. The means of the scaled scores ranged from 11.5 to 13.6, indicating that the participants were functioning well above the average levels of the nationally representative normative sample. However, the absence of significant relations between age and the scaled scores suggests that there was no confounding between age and degree of selectivity or representativeness relative to the normative sample.

Materials

The test battery consisted of 13 tests that we selected to represent four distinct cognitive abilities. We chose the tests, which are briefly described in Table 2, because prior research indicated that the measures were all reliable, and each had a strong loading on its respective ability factor in a confirmatory factor analysis (Salthouse, 2004; Salthouse & Ferrer-Caja, 2003). We developed alternate versions of the tests by using other items from the same tests, such as even-numbered items from the Advanced Ravens Progressive Matrices, or from similar tests, such as the Wechsler Abbreviated Scale of Intelligence (WASI; 1999) vocabulary and Neuropsychological Assessment Battery (Stern & White, 2003) story memory, or by creating new items (e.g., digit symbol, paper folding, spatial

Table 2. Description and Source of Variables

Variable	Description	Source
Vocabulary	Provide definitions of words	Wechsler (1997a)
Picture vocabulary	Name the pictured object	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)
Digit symbol	Use a code table to write the correct symbol below each digit	Wechsler (1997a)
Letter comparison	Same or different comparison of pairs of letter strings[a]	Salthouse & Babcock (1991)
Pattern comparison	Same or different comparison of pairs of line patterns[a]	Salthouse & Babcock (1991)
Matrix reasoning	Determine which pattern best completes the missing cell in a matrix	Raven (1962)
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)
Logical memory	No. of idea units recalled across three stories	Wechsler (1997b)
Free recall	No. of words recalled across Trials 1–4 of a word list	Wechsler (1997b)
Paired associates	No. of response terms recalled when presented with a stimulus term	Salthouse et al. (1996)

[a]There are two separately timed parts to this test, and thus the scores on each part can be treated as separate items when computing estimates of reliability.

relations, letter comparison, pattern comparison, synonym vocabulary, antonym vocabulary, and picture vocabulary). In all cases the number of items in the new versions of the tests was the same as in the original versions, but there was no overlap of specific items across versions. We designated the three versions as O for the original version and as A and B for each of the two new versions, respectively.

Procedure

The three versions of the tests were administered on successive occasions to all participants in the order O-A-B. We scheduled the three occasions at the individual's convenience, but most occasions were at approximately the same time of day within a 2-week period. Depending on the individual's schedule, the sessions occurred across successive days or within an interval extending to as many as 10 weeks.

Because the test versions could have differed in difficulty, we conducted a preliminary study in which 60 young adults performed the three versions of each test in different orders. That is, 10 individuals each performed the tests in the six possible orders (O-A-B, O-B-A, etc.). The results from the preliminary study revealed that there were significant version differences in most of the tests, and thus we used the following procedure to adjust the scores on the A and B versions of each test to approximately match the mean of the O version. First, we computed linear regression equations from the data in the preliminary study to predict the scores on the O version of the test from the score on either the A or the B version. Second, we used the intercept and slope parameters from these equations to create adjusted A and B scores for each participant in the current study. To illustrate, the regression equation predicting the digit symbol score on the O version from the score on the A version based on the data from the preliminary study was $DS_O = 13.45 + 0.83(DS_A)$. The application of these parameters to a participant in the current study with a score of 75 on version A would therefore result in an adjusted version A score of 75.7, or 13.45 + 0.83(75). Our rationale for the adjustment procedure was that the results from a study in which the test versions were administered in counterbalanced order can be used to equate the difficulty of the new versions to that of the original version (see endnote).

Results

Table 3 contains means, standard deviations, and estimated (coefficient alpha) reliabilities of the original (O version) and adjusted (A and B versions) scores for each test. It can be seen that most of the reliability estimates were above .7, with the exception of the synonym and antonym measures for the A and B versions. Although not presented in Table 3, the correlations between the scores on different versions of the same test were all moderately high; the median was .74, and it was .88 after we applied the standard formula to adjust each correlation for unreliability of the scores.

Table 3. Adjusted Means, Between-Person Standard Deviations, and Estimated Reliabilities

	Original			A			B		
Construct Variable	*M*	*SD*	Reliability	*M*	*SD*	Reliability	*M*	*SD*	Reliability
Vocabulary									
Vocabulary	51.9	9.2	.92	55.2	6.9	.85	54.6	5.3	.81
Picture vocabulary	20.0	4.5	.84	17.9	2.8	.71	18.4	3.5	.76
Synonym vocabulary	7.7	2.3	.78	6.7	1.7	.65	6.9	1.9	.52
Antonym vocabulary	6.7	2.9	.84	5.9	2.0	.61	6.0	2.1	.58
Perceptual speed									
Digit symbol	76.2	18.3	NA	83.0	15.8	NA	82.8	19.2	NA
Letter comparison	11.0	2.7	.89	11.9	1.8	.88	12.1	2.1	.81
Pattern comparison	17.0	4.2	.90	17.7	3.9	.90	18.5	3.9	.90
Fluid Cognition									
Matrix reasoning	8.2	3.3	.79	10.6	2.2	.79	8.2	3.9	.81
Spatial relations	9.6	4.8	.87	13.8	2.1	.66	15.1	2.0	.70
Paper folding	6.6	2.8	.77	8.5	1.7	.74	7.8	2.7	.84
Episodic memory									
Logical memory	43.9	9.2	.86	47.8	5.6	.84	48.6	6.9	.86
Word recall	36.8	6.4	.90	39.4	3.9	.91	38.7	4.2	.92
Paired associates	3.6	1.8	.82	4.5	0.7	.82	4.3	0.9	.89

Notes: For all variables except for the antonym vocabulary and matrix reasoning variables in version B, scores for versions A and B were adjusted with the regression equations derived from the results of the preliminary study. Reliability refers to coefficient alpha estimates of reliability. NA indicates that an estimate of reliability was not available.

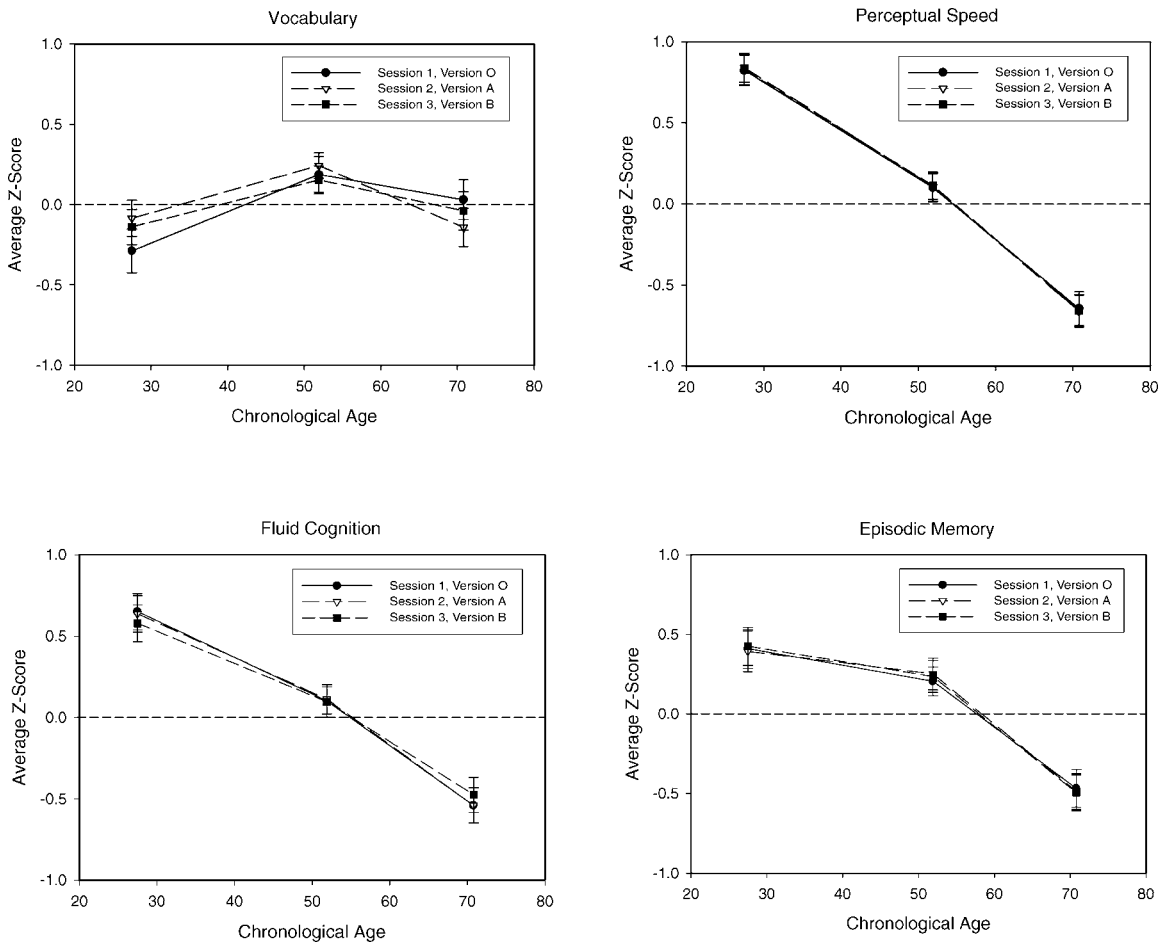

Figure 1. Mean composite scores across the three sessions or versions as a function of age. Bars above and below each point are standard errors.

Figure 1 portrays the mean composite scores, which we created by averaging the *z* scores for the variables assumed to represent the same construct (see Table 3), as a function of age and session or version. Two points should be noted about the results in this figure. First, the patterns of age relations on the composite variables are very similar to those reported in many previous studies (e.g., Salthouse, 2004; Salthouse & Ferrer-Caja, 2003), as there is an increase followed by a slight decrease in measures of vocabulary but generally monotonic declines for the other measures. Second, the patterns were nearly identical for the composite scores based on the three versions administered in different sessions. This latter finding suggests that any version or session differences that might exist were fairly similar in each age group.

Three ways of expressing within-person variability for the measures are presented in Table 4. The simplest method is with the standard deviation of the scores across the three occasions (versions). Although computationally straightforward, the standard deviation is not easily interpretable without a frame of reference. One such frame of reference is the variability apparent across people, and therefore a second method of expressing within-person variability is in terms of the ratio of the average within-person standard deviation to the between-person standard deviation on the first occasion. Values for this index of within-person variability are contained in the fourth column of Table 4. Another frame of reference is the variation associated with an individual difference variable such as age. The fifth column of Table 4 therefore expresses within-person variability relative to the differences associated with cross-sectional age variation for the measures with significant age correlations. Specifically, we computed the slope relating the score on the first occasion to years of age, and then we divided the slope into the average within-person standard deviation to determine the number of years of cross-sectional age difference corresponding to the average within-person standard deviation. To illustrate, the age regression slope for the digit symbol variable was –0.635, and when this number was divided into 5.4 it yielded a value of 8.5.

The median within-to-between ratio across the 13 measures was 0.46, which indicates that the variation for a given individual from one occasion to the next is almost one half as much as the variation from one person to the next. (This comparison is based on standard deviations as the measure of variability because they are in the same units of measurement as the variable. A comparison based on variances would obviously yield much smaller ratios, because the variance is the square of the standard deviation.) For the 9 measures with negative age correlations, the average within-person standard deviation

Table 4. Different Methods of Expressing Within-Person Variability

Variable	Within *SD*		Within/ Between	Age, in Years
	M	*SD*		
Vocabulary	3.5	2.4	.38	NA
Picture vocabulary	2.1	1.1	.47	NA
Synonym vocabulary	1.3	0.9	.57	NA
Antonym vocabulary	1.5	0.9	.52	NA
Digit symbol	5.4	2.9	.30	8.5
Letter comparison	1.0	0.6	.37	12.3
Pattern comparison	1.4	0.8	.33	9.2
Matrix reasoning	2.0	1.0	.61	18.2
Spatial relations	3.3	1.8	.69	27.3
Paper folding	1.5	0.8	.54	19.0
Logical memory	4.2	2.3	.46	29.2
Word recall	2.5	1.5	.39	16.3
Paired associates	0.8	0.5	.44	18.6

Notes: Within SD refers to the within-person standard deviation across the three occasions; within, between is the ratio of average within-person standard deviation to the between-person standard deviation on the first occasion; and years of age refers to the number of years of cross-sectional age difference on the first occasion corresponding to the within-person standard deviation. NA indicates that the estimate was not available because the variable did not have a significant negative correlation with age.

Table 5. Correlations of Within-Person Variability and Mean Performance With Each Other and With Age Before and After Control of the Other Variable

Variable	*SD*−*M*	Age Correlations			
		SD	*SD*·*M*	*M*	*M*·*SD*
Vocabulary	−.70*	.26*	.14	−.18	.01
Picture vocabulary	−.02	.17	.17	−.04	−.04
Synonym vocabulary	−.20	−.01	.03	.19	.19
Antonym vocabulary	−.14	.06	.06	−.01	.00
Digit symbol	−.10	.13	.08	−.65*	−.65*
Letter comparison	−.34*	.21	.00	−.61*	−.55*
Pattern comparison	−.12	−.08	.01	−.71*	−.70*
Matrix reasoning	−.48*	.18	−.18	−.65*	−.57*
Spatial relations	−.78*	.36*	−.01	−.49*	−.21*
Paper folding	−.58*	.37*	.06	−.54*	−.35*
Logical memory	−.39*	.03	−.13	−.39*	−.38*
Word recall	−.60*	.29*	−.00	−.49*	−.33*
Paired associates	−.70*	.23*	−.12	−.48*	−.33*

Notes: The age correlation for *SD*·*M* is the semipartial correlation between age and *SD* after controlling the variation in the mean, and the age correlation for *M*·*SD* is the semipartial correlation between age and mean after controlling for variation in the *SD*.

*$p < .01$.

was equivalent to the amount of variation apparent in cross-sectional comparisons across a period ranging from 8 to 29 years, with a median of 18.2.

The values in Table 4 indicate that there is considerable within-person variability in each of the measures of cognitive functioning. Moreover, the ratios of within-person to between-person variability were actually somewhat larger for the vocabulary, fluid cognition, and episodic memory variables than for the perceptual speed variables that are most similar to the types of variables examined in prior studies of within-person variability. Another point to note in this table is that people differ in the magnitude of their within-person variability. That is, the (between-person) standard deviation of the within-person (across-occasion) standard deviations, reported in the third column of Table 4, are all moderately large. The variation across occasions is therefore not simply a reflection of a situation in which everybody is affected to the same degree.

Table 5 contains correlations of the within-person standard deviation and mean across the three occasions, and the correlations of these variables with age before and after control of the other variable. The values in the second column indicate that most of the correlations between an individual's mean and his or her across-session variability were negative, which indicates that better (higher) performance was associated with smaller across-session variability.

An inspection of the third column in Table 5 reveals that increased age was associated with significantly larger within-person variability for five of the variables. However, in every case the correlation with age was substantially reduced, and no longer significantly different from zero, after we controlled the variation in the mean score by means of a semipartial correlation. Nine of the variables had significant (negative) correlations between age and the mean, and all were still statistically significant after we controlled the variation in the across-occasion standard deviation. This pattern suggests that increased age is associated with lower mean performance on many of the cognitive variables, but that any relations between age and the measure of within-person variability appear to be attributable to the relations both variables have with the mean.

We investigated the question of whether people can be characterized as more or less variable across different types of cognitive tests by computing correlations of the within-person standard deviations for the 13 cognitive tests. The correlations ranged from −.15 to +.48, but the median was only .05, and it was still only .09 when we ignored the sign of the correlation. We also computed correlations after partialling age from both variables, and in subsamples with a narrower range of ages. The median age-partialled correlation was .05, and the median correlations for participants 18 to 39, 40 to 59, and 60 to 97 years of age were, respectively, .06, −.01, and .11.

One possible reason for the low correlations is weak reliability of the within-person variability measures. We investigated this possibility by obtaining estimates of the reliability of the within-person standard deviations by using the standard deviations from scores on different pairs of sessions as the "items" in the coefficient alpha. That is, for each individual, we computed three standard deviations for each test variable based on the scores in sessions 1 and 2, the scores in sessions 1 and 3, and the scores in sessions 2 and 3, and we then treated these three standard deviations (each based on two scores) as items in the computation of the coefficient alpha.

The reliability estimates derived in this manner ranged from .42 to .71, with a median of .59. Although these values are lower than the psychometric standard of .70, they nevertheless indicate that the measures of within-person variability were sufficiently high to sustain meaningful correlational patterns if they were to exist in the data. What is most important is that, when we adjusted the correlations among the within-person standard deviations for these estimates of reliability, they were still quite small, with a median of only .10. At least on the basis of these results, therefore, it does not appear that individuals who exhibit large across-occasion variability in one cognitive measure are any more likely than the average individual to exhibit large across-occasion variability in other cognitive measures.

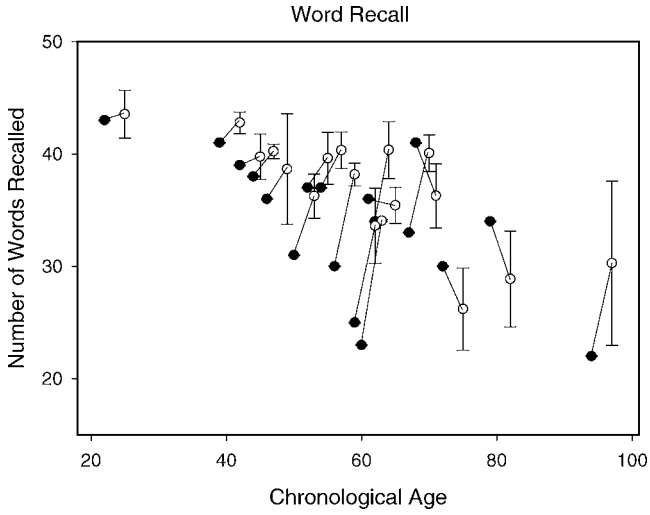

Figure 2. Scores in 2001 (solid circles), and means (open circles) and standard deviations (bars) across the three assessments in 2004 for individual participants on the Wechsler Memory Scale III word recall variable.

We also conducted an exploratory factor analysis on the within-person standard deviations. As we would expect from the low correlations, the factor analysis did not reveal much evidence of structure. The correlation matrix had six eigenvalues greater than one, and thus we extracted six factors (iterative principal axes) and rotated them by means of promax. There were three variables with loadings greater than .4 on the first factor (i.e., spatial relations, paper folding, and paired associates), two on the second factor (i.e., synonym vocabulary and antonym vocabulary), four on the third factor (i.e., vocabulary, logical memory, recall, and paired associates), two on the fourth factor (i.e., letter comparison and recall), and one each on the fifth (i.e., matrix reasoning) and sixth (i.e., pattern comparison) factors. The relatively large number of factors with modest loadings of the variables on each factor, and the diversity of the variables loading on each factor, suggests that there is little or no structure in the available measures of within-person variability. This finding is substantially different from that observed in analyses of the means, because there is clear evidence of a strong structure among the means of these variables (e.g., Salthouse, 2004; Salthouse & Ferrer-Caja, 2003).

The other major issue of interest in this study concerned the relation between short-term variability and longitudinal change. Because people vary in the magnitude of short-term within-person variability, it is possible that the same absolute value of longitudinal change could have quite different meanings for different people. For some individuals the change might be well within their normal range of fluctuation, but for others it might represent an extreme value. We could examine this possibility in the current data because 18 of the participants (age at first assessment, $M = 57.1$) had performed the original version of six of the tests 3 years earlier. The tests common across the 2001 and 2004 assessments were Vocabulary, Picture Vocabulary, Digit Symbol, Logical Memory, Word Recall, and Paired Associates tests. This sample is too small for meaningful statistical analyses, but it is useful for illustrating the point that change in the original units of measurement may not have the same functional meaning for different people.

The longitudinal patterns were generally similar for each variable and can be illustrated with the results from the Word Recall test. The values for individual participants on this variable are portrayed in Figure 2, with the solid circles indicating the 2001 score and the open circles with bars representing the mean and standard deviation, respectively, of the three scores in 2004. An inspection of the figure indicates that, for many of the individuals, the scores were higher at the later assessment. Although these gains could reflect true improvements in ability, we suspect that a substantial proportion of the performance gains are attributable to retest effects (e.g., Ferrer, Salthouse, Stewart, & Schwartz, 2004; Salthouse, Schroeder, & Ferrer, 2004).

The bars around the 2004 (Time 2, or T2) estimates in the figure confirm the finding that there is variability in performance from one occasion to another within a short interval, and the variation in the size of the bars indicates that the degree of across-occasion variability varies across individuals. These results imply that the same absolute longitudinal change may not have the same meaning for everyone. That is, a given magnitude difference from 2001 to 2004 may be small relative to the typical fluctuation for someone with relatively large within-person variability, whereas it could be large for a person who is much less variable.

This point can be illustrated by considering two individuals in Figure 2 who were 44 and 46 years of age in 2001. These two individuals had similar word recall scores of 38 and 36 in 2001, and similar means across the three assessments in 2004 of 40 and 39, respectively. However, their across-occasion standard deviations in 2004 were 0.7 and 4.9, respectively, which indicates that the slightly larger absolute difference for the 46-year-old individual (i.e., 3 vs 2) was actually substantially smaller than that of the 44-year-old when it is expressed in within-person standard deviation units (i.e., 0.6 vs 2.9).

The preceding example suggests that, depending on the method used to calibrate change, different conclusions might be reached about the magnitude and correlates of change. In order to examine this issue more systematically, Table 6 contains information relevant to different methods of calibrating change for the six variables with longitudinal data. In addition to containing the means and between-person standard deviations for the 2001 (Time 1, or T2) score, the mean across the three T2 scores, and the standard deviation of the T2 scores, Table 6 also summarizes three ways of expressing the within-person change from T1 to T2. The simplest and most frequently used method of representing change is the difference between the relevant scores at each occasion (i.e., the mean of the T2 scores minus the T1 score) in the original units of measurement. A second method is the difference scaled in T1 between-person standard deviation units (e.g., Ivnik et al., 1999; Schaie, 1996), and a third method is the difference relative to each individual's T2 within-person standard deviation (Salthouse et al., 1986).

An examination of the entries in Table 6 reveals that the different methods of evaluating change yield different estimates of the magnitude of change, and perhaps more importantly, of the magnitude of individual differences in change. To illustrate, the between-person standard deviation for the absolute difference in the paired associates variable was approximately one half the average 3-year change (i.e., 0.8 vs 1.5), but when we

Table 6. Retest Statistics

Variable	T1	T2 *M*	T2 *SD*	Difference	Difference (T1 *SD*)	Difference (*SD*)
Vocabulary	52.1 (13.2)	52.3 (6.2)	3.8 (3.2)	0.2 (8.8)	0.01 (.67)	−0.2 (1.8)
Picture vocabulary	19.7 (5.2)	18.4 (3.7)	2.8 (1.4)	−1.3 (2.2)	−0.25 (.43)	−0.2 (2.4)
Digit symbol	72.6 (18.3)	75.4 (16.7)	5.8 (3.6)	2.8 (7.0)	0.15 (.38)	1.4 (2.7)
Word recall	33.9 (6.1)	36.9 (4.8)	2.5 (1.7)	3.0 (4.6)	0.50 (.76)	8.4 (28.8)[a]
Logical memory	43.3 (10.9)	45.9 (7.1)	3.7 (2.0)	2.6 (5.5)	0.24 (.50)	1.0 (2.4)
Paired associates	2.5 (1.6)	4.0 (1.1)	0.7 (0.5)	1.5 (0.8)	0.94 (.50)	4.7 (8.4)

Notes: For the table, $N = 18$. T1 is the score at the 2001 assessment; T2 *M* is the mean across the three assessments in 2004; T2 *SD* is the within-person standard deviation across the three assessments in 2004; Difference is the T2 *M* − T1 difference; Difference (T1 *SD*) is the difference between T2 *M* and T1 in T1 between-person *SD* units; and Difference (*SD*) is the difference between T2 *M* and T1 in person-specific T2 *SD* units. Values in parentheses are between-person standard deviations.

[a]The presence of an extreme outlier inflated the between-person standard deviation because the value without the outlier was 2.3.

scaled change in within-person standard deviation units, the between-person standard deviation was almost double the average change (i.e., 8.4 vs 4.7). The rank ordering of individuals in terms of the amount of change can also differ according to method of assessment. In fact, the correlations between the absolute difference and the difference scaled in within-person standard deviation units were .66 for vocabulary, .81 for picture vocabulary, .71 for digit symbol, .49 for word recall, .75 for logical memory, and .24 for paired associates.

Discussion

The results of this study provide answers to the two major issues that motivated us to initiate the project. First, the short-term within-person variability in accuracy measures of cognitive functioning is substantial, and for many variables it is nearly one half as large as that for between-person variability. The existence of moderate to large within-person variability for each of the 13 cognitive variables implies that single assessments may not be very informative about an individual's true level of functioning. This further compounds the psychometric problems associated with evaluations of within-person change associated with normal aging, disease, trauma, or some type of intervention. Accuracy of measurement is sometimes evaluated with retest correlations because they indicate the stability of the ordering of individuals. However, it is important to note that retest correlations could be moderately large because of the substantial variation across people in their average levels, and yet there could still be considerable fluctuation around these levels. Indeed, this situation is apparent in Figure 2 because people maintain approximately the same relative positions from T1 to T2 (i.e., the 3-year stability coefficient for the word recall variable was .67, and the median across the six variables was .81), but the bars corresponding to within-person variability indicate that the assessments still exhibited substantial variability.

The existence of within-person variability in cognitive functioning has important implications for the evaluation of change. Most contemporary researchers assess change in the original units of measurement, and thus they implicitly assume that the units have the same meaning for everyone. A number of researchers have attempted to evaluate individual change relative to the variability that exists across people (e.g., Ivnik et al., 1999; Schaie, 1996), but this is not ideal because between-person variability is only a crude approximation of within-person variability. Salthouse and colleagues (1986) proposed that researchers might obtain more sensitive assessments of change by expressing the change for each individual relative to his or her own across-occasion variability. This method is analogous to the computation of an effect size for each individual, and it shares the property that normal variability is taken into account when the magnitude of an effect is specified.

Conclusions about the magnitude of change, about between-person variation in change, and about correlates of change will therefore vary depending on how the change is assessed. The optimal method to be used in assessing change will obviously depend on the specific question of interest. For example, a comparison in the absolute units of measurement may be more meaningful if the variable is scaled in a ratio level of measurement such as units of time, or if it represents progress toward an absolute criterion. However, change calibrated relative to each individual's within-person variability may be more meaningful if the goal is to investigate change in units that are functionally equivalent in different people.

Although large individual differences in amount of within-person variability are evident in every variable in Tables 4 and 6 and in Figure 2, there were no significant correlations between age and the measures of within-person variability after we adjusted for influences associated with the mean. This pattern is similar to that recently reported by Salthouse and Berish (2005) in several analyses, and it seems to suggest that information about an individual's short-term variability may not have any unique predictive power beyond what is available from his or her mean level of performance. However, it should be noted that in samples of older adults, Hultsch and colleagues (2000) and Rabbitt and associates (2001) have reported that within-person variability in reaction time was correlated with level of performance in other cognitive tasks. Therefore, it may be the case that within-person variability only provides unique information with variables assessing performance speed rather than performance accuracy, or in samples of individuals likely to be experiencing substantial change in level of cognitive functioning.

In our study there was also little evidence of structure in the measures of within-person variability either in the raw correlations or in the exploratory factor analysis, before or after we adjusted for reliability of the measures. This pattern suggests that influences contributing to across-occasion variability in these variables are specific to particular variables and are not shared across different variables, even those assumed to reflect the same cognitive ability. In other words, even though the reliability estimates suggest that within-person variability is

not simply random fluctuation, there is no evidence in these data that people who exhibit high across-occasion variability for one cognitive variable exhibit high across-occasion variability for other cognitive variables. Li and associates (2001) also failed to find much evidence of structure among the measures of within-person variability for several memory measures assessed across 25 occasions. Important questions for future research are the determination of what is responsible for the across-occasion variability in cognitive performance, and why there are such weak relations among the within-person variability measures from different cognitive tasks.

In conclusion, there is now considerable evidence that calls into question the adequacy of the classical notion of a fixed true score as an ideal focus of measurement efforts. Theoretical concepts and analytical methods should therefore reflect this shift of thinking if progress it to be made in describing, measuring, and explaining behavior and behavior change.

Acknowledgments

This research was supported by the National Institute on Aging under Grant RO1 AG 19627 to T. A. Salthouse. We thank the following people for the scheduling and testing of participants and the entering and checking of data: James Darragh, Samantha Norton, Malaika Schiller, Sara Shelley, Rachel Spiotto, Catherine Thrasher, and David Yost.

References

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

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.

Ferrer, E., Salthouse, T. A., Stewart, W., & Schwartz, B. (2004). Modeling age and retest processes in longitudinal studies of cognitive abilities. *Psychology and Aging, 19*, 243–259.

Hertzog, C., Dixon, R. A., & Hultsch, D. F. (1992). Intraindividual change in text recall of the elderly. *Brain and Language, 42*, 248–269.

Hultsch, D. F., & MacDonald, S. W. S. (2004) Intraindividual variability in performance as a theoretical window onto cognitive aging. In R. A. Dixon, L. Backman, & L-G. Nilsson (Eds.), *New frontiers in cognitive Aging* (pp. 65–88). New York: Oxford University Press.

Hultsch, D. F., MacDonald, S. W. S., 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.

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.

Kliegl, R., Smith, J., & Baltes, P. B. (1989). Testing-the-limits and the study of adult age differences in cognitive plasticity of a mnemonic skill. *Developmental Psychology, 25*, 247–256.

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.

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 aesthetics: 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 Sciences, 59B*, P49–P55.

Rabbitt, P., Osman, P., Moore, B., & Stollery, B. (2001). There are stable individual differences in performance variability, both from moment to moment and from day to day. *Quarterly Journal of Experimental Psychology, 54A*, 981–1003.

Rapport, L. J., Brines, D. B., Axelrod, B. N., & Thiesen, M. E. (1997). Full scale IQ as a 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.

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 the 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., & 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., 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., & Somberg, B. L. (1982). Skilled performance: The effects of adult age and experience on elementary processes. *Journal of Experimental Psychology: General, 111*, 176–207.

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

Shammi, P., Bosman, E., & Stuss, D. T. (1998). Aging and variability in performance. *Aging, Neuropsychology, and Cognition, 5*, 1–13.

Spielberger, C. D., Gorsuch, R. L., & Lushere, R. E. (1970). *Manual for the State–Trait Anxiety Inventory—Form X*. Palo Alto, CA: Consulting Psychologists Press.

Stern, R. A., & White, T. (2003). *Neuropsychological Assessment Battery administration, scoring, and interpretation manual*. Lutz, FL: Psychological Assessment Resources.

WASI (Wechsler Abbreviated Scale of Intelligence) manual. San Antonio: The Psychological Corporation.

Wechsler, D. (1997a). *Wechsler Adult Intelligence Scale* (3rd ed.). San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1997b). *Wechsler Memory Scale* (3rd ed.). San Antonio, TX: The Psychological Corporation.

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

Received June 15, 2005
Accepted October 10, 2005
Decision Editor: Thomas M. Hess, PhD

Endnote

Although it would have been desirable to use individuals of the same age and ability range as those in the primary study for this calibration study, college students were more readily available and could be compensated with credit towards a course requirement rather than with money. These young adults had somewhat higher average levels of performance than the age-heterogeneous sample in the primary study on many of the variables. However, most of the relations between the predictor (i.e., version A or B) and criterion (i.e., version O) variables in the primary sample were linear (i.e., for the linear relation, median R^2 = was .532; for the quadratic relation, median R^2 = .004; and for the cubic relation, median R^2 = .002), indicating that it is reasonable to extrapolate from one region of the distribution to the entire distribution.