Salthouse, T.A., & Nesselroade, J.R. (2010). Dealing with short-term fluctuation in longitudinal research. *Journal of Gerontology: Psychological Sciences, 65B(6)*, 698–705, doi:10.1093/geronb/gbq060. Advance Access published on August 23, 2010.

Dealing With Short-term Fluctuation in Longitudinal Research

Timothy A. Salthouse and John R. Nesselroade

Department of Psychology, University of Virginia, Charlottesville.

Objectives. Many analytical methods are not very sensitive to change because of the difficulty of distinguishing short-term fluctuation from the developmental change of primary interest. The current project investigated one possible solution to this problem in the form of a measurement-burst design in which research participants perform several versions of each test at each measurement occasion.

Methods. Over 1,200 adults across a wide-age range performed different versions of cognitive tests on several sessions at each measurement occasion.

Results. Four methods of incorporating short-term variability were compared with respect to the magnitude of the correlations of the ability measures with each other and with respect to the magnitude of their relations with age.

Conclusions. The results revealed that more sensitive assessments of change can be obtained by taking short-term fluctuation into account with measurement-burst designs. In particular, capitalizing on the availability of multiple measures at each occasion to form latent constructs representing the level and change in cognitive performance may provide the most sensitive assessment of cognitive change.

Key Words: Cognitive change—Intra-individual variability—Measurement burst.

A number of articles have recently reported considerable within-person or intra-individual variability in cognitive performance with various reaction time tasks (e.g., Bunce, Handley, & Gaines, 2008; Duchek et al., 2009 ; Gorus, De Raedt, Lambert, Lemper, & Mets, 2008; MacDonald, Hultsch, & Dixon, 2008), as well as other types of cognitive tasks (e.g., Nesselroade & Salthouse, 2004; Salthouse, 2007; for a review, see Hultsch, Strauss, Hunter, & MacDonald, 2008). The discovery that people differ in the degree to which their performance varies has led to interest in examining relations of a variety of individual difference characteristics to measures of variability in addition to measures of average level of performance (see earlier citations and Nesselroade & Salthouse; Salthouse, 2007; Salthouse, Nesselroade, & Berish, 2006). The existence of substantial within-person cognitive variability also has important implications for longitudinal research because when short-term fluctuation is large, it can be difficult to distinguish the long-term change of primary interest from short-term fluctuation (cf. Salthouse et al., 2006). That is, estimates of longitudinal change might not be very accurate if the "noise" associated with short-term fluctuation is large relative to the "signal" corresponding to true longitudinal change.

One possible solution to the problem of distinguishing short-term fluctuation from meaningful longitudinal change was proposed by Nesselroade (1991) in the form of a measurement-burst design in which each individual is assessed multiple times at each occasion. The reasoning was that a single assessment can be assumed to reflect merely one sample from a distribution of many possible assessments, and multiple assessments can be expected to provide better estimates of the hypothesized distribution.

Figure 1 schematically portrays four ways in which multiple-assessment measurement-burst data could be analyzed in longitudinal research. The simplest method, represented in the top left panel, merely involves averaging the scores across the multiple assessments at each occasion and then using the difference between the averages as the measure of longitudinal change. The rationale for this method is that aggregation across several assessments at each measurement occasion should minimize the influence of short-term fluctuation, such that the difference between averages would be more sensitive than the difference between single assessments.

A second possible method of analyzing burst data, illustrated in the top right panel, was discussed by Salthouse (2007) (see Nesselroade & Salthouse, 2004; Salthouse et al., 2006; Salthouse, Kausler, & Saults, 1986), who suggested that the multiple measures at each occasion could be used to calibrate the longitudinal change in terms of each individual's own within-burst (across-assessment) variability. The motivation for expressing the difference relative to each person's variability is based on the assumption that a greater absolute change is needed to have the same meaning for someone with large short-term fluctuation compared with someone with small short-term fluctuation. To the extent that people differ in the magnitude of within-person variability, therefore, these variability-adjusted differences might be more sensitive than absolute differences.

If there are sequential relations on the performance measures across assessments within each occasion, possibly

Received December 14, 2009; Accepted July 19, 2010
Decision Editor: Rosemary Blieszner, PhD

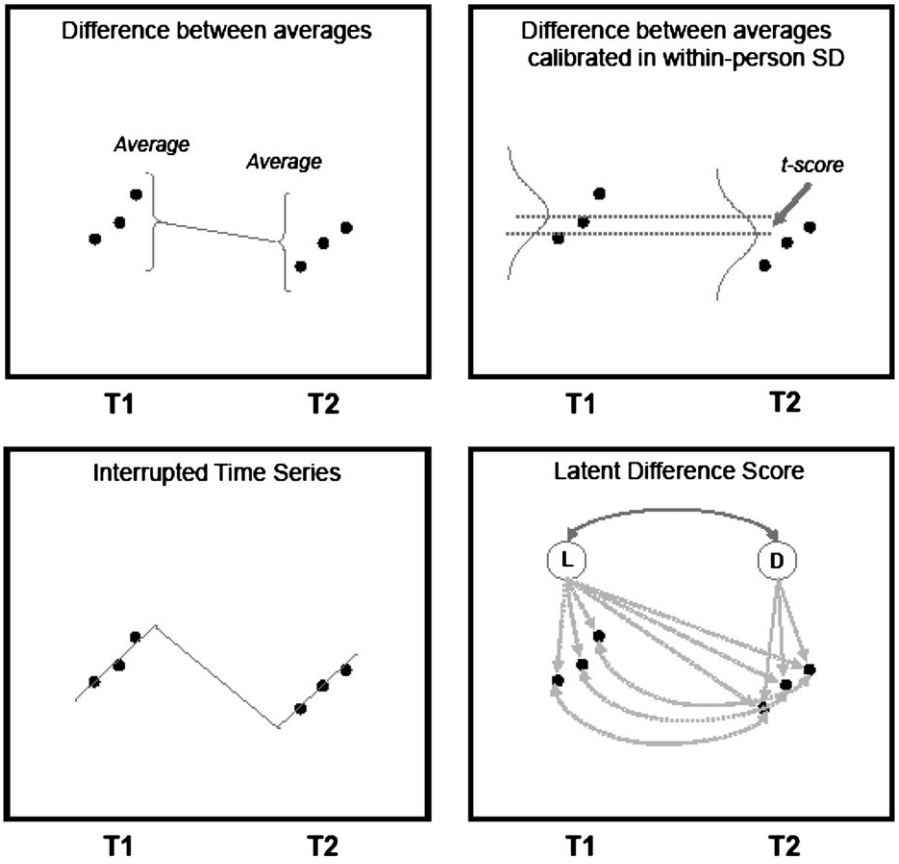

Figure 1. Schematic portrayal of four methods of analyzing three-assessment measurement-burst data in a longitudinal study. See text for details.

attributable to practice or fatigue, a third way in which longitudinal measurement-burst data could be analyzed is with a variant of an interrupted time series analysis, as illustrated in the bottom left panel of Figure 1. The version of this method used in the current project consisted of first computing the linear regression equations relating test performance to the successive assessments within each measurement burst. The parameter estimates from these two equations were then used to generate a predicted value for the third assessment on the first occasion and for the first assessment on the second occasion. Subtraction of the former value from the latter then served as the estimate of change based on the systematic relations within each measurement burst.

A fourth possibility for analyzing measurement-burst data, portrayed in the bottom right panel of Figure 1, is to treat the scores from the different assessments at each occasion as multiple indicators of a latent construct and then to examine across-occasion differences with a latent difference score structural equation model (i.e., McArdle & Nesselroade, 1994). Because latent constructs represent only the reliable variance that is shared among the indicator variables, latent difference score models have the potential to minimize problems of low reliability of difference scores.

A primary goal of the current report was to examine five methods of assessing longitudinal change, which differ in whether, and how, they deal with short-term fluctuation (which can be defined as the nonidentical performance on presumably equivalent tests performed in close temporal proximity). In order to provide a baseline for comparison, one method is the traditional difference between single scores across the two occasions. The initial measurement at each occasion was used for this purpose. The other four methods are those outlined above, namely, the difference between the average scores at each occasion, the difference between the average scores for a given individual divided by his or her average within-occasion standard deviation, the difference between predicted values based on regression equations for successive scores within each measurement burst, and a latent difference score derived from constructs based on the three assessments at each occasion.

Two criteria were used to compare the analytical methods. Both criteria relied on the assumption that other things being equal, change measures with stronger relations to other variables can be inferred to be more sensitive and reliable than change measures with weaker relations. The first criterion was based on correlations of the changes with the changes in other variables representing the same cognitive ability and the second criterion was based on the correlations of the changes with a different type of variable, namely age.

The data to be described were obtained from a measurement-burst design in which three parallel versions of different cognitive tests were administered within about a 2-week period on each of two measurement occasions. The interval between occasions varied across participants, but averaged approximately 2.4 years, and was uncorrelated with age. At each assessment, the participants performed a battery of 12 cognitive tests selected to represent four different cognitive abilities (i.e., reasoning, spatial visualization, episodic verbal memory, and perceptual speed).

Longitudinal data are available from two groups of participants, one group with only the first version of each test in the initial occasion and a three-assessment burst in the second occasion (i.e., single burst) and another group with the three-assessment measurement burst on both occasions (i.e., double burst). The two groups were combined in some analyses, but only the data from the double burst participants were used in other analyses.

Method

Participants

Participants were recruited from newspaper advertisements, flyers, and referrals from other participants. Approximately 81% identified themselves as White, 10% identified themselves as African American, and 5% identified themselves as a mixture, with the remaining participants distributed in very small percentages across other ethnic categories. A total of 3,298 adults originally participated between 2001 and 2007, and 1,282 of them returned for retesting between 2004 and 2008. Characteristics of the returning participants through 2008, who are the primary focus of this report, are summarized in Table 1. Four tests, Vocabulary, Digit Symbol, Word Recall, and Logical Memory, were obtained from standardized test batteries (Wechsler, 1997a, 1997b), and therefore, scores on these tests could be compared with data from the nationally representative normative samples to evaluate the representativeness of the current samples of participants. Scaled scores in the normative sample are adjusted for age and have *M*s of 10 and *SD*s of 3.

Increased age was associated with somewhat lower self-ratings of health but with a greater amount of education and higher age-adjusted (scaled scores) levels of Vocabulary, Digit Symbol, Logical Memory, and Word Recall. The scaled score means indicate that the current sample was functioning about 0.5–1 *SD*s above the nationally representative sample used for the norms in these tests. Moreover, the positive age correlations indicate that older adults in the sample were functioning at somewhat higher levels relative to their age peers than were the younger adults. Although these sample characteristics may limit generalizations to a broader population, if anything, the higher level of functioning among the older adults may lead to underestimates of the actual age differences.

Table 1. Descriptive Characteristics of Participants With a Three-Assessment Measurement Burst Only at T2 (single burst) or With a Burst Assessment at Both T1 and T2 (double burst)

	M	*SD*	Age correlation
Single-burst participants			
N	862	NA	NA
Age at T1	53.2	16.6	NA
Proportion of female	0.66	NA	–0.05
Years of education	15.7	2.7	0.21*
Self-rated health	2.1	0.9	0.14*
Retest interval	2.6	1.2	–0.03
Scaled scores (at T1)			
Vocabulary	12.9	3.1	0.16*
Digit Symbol	11.5	2.8	0.10*
Logical Memory	12.1	2.9	0.17*
Word Recall	12.6	3.1	0.09
Double-burst participants			
N	420	NA	NA
Age at T1	54.7	18.7	NA
Proportion of female	0.63	NA	–0.01
Years of education	15.7	2.6	0.21*
Self-rated health	2.3	0.9	0.14*
Retest interval	2.2	0.6	0.00
Scaled scores (at T1)			
Vocabulary	13.0	2.9	0.11
Digit Symbol	11.6	2.8	0.18*
Logical Memory	12.0	2.8	0.21*
Word Recall	12.6	3.4	0.07

Note: The age range in the single-burst participants was 18–95 years and that in the double-burst participants was 18–91 years. Health was rated on a 5-point scale in which 1 represented *excellent* and 5 represented *poor*. Scaled scores have *M* of 10 and *SD*s of 3 in the normative samples (i.e., Wechsler, 1997a, 1997b). NA = not applicable.

*$p < .01$

Cognitive Tests

The cognitive tests and results of confirmatory factor analyses, indicating the pattern of relations of variables to ability constructs, have been reported in other publications (Salthouse, 2004, 2005, 2007; Salthouse, Pink, & Tucker-Drob, 2008; Salthouse & Tucker-Drob, 2008). The three tests representing each cognitive ability were Matrix Reasoning, Shipley Abstraction, and Letter Sets for reasoning; Spatial Relations, Paper Folding, and Form Boards for spatial visualization (space); Word Recall, Paired Associates, and Logical Memory for memory; and Digit Symbol, Pattern Comparison, and Letter Comparison for perceptual speed. The tests on different sessions within an occasion (i.e., measurement burst) involved different items, but the same tests were repeated in the same order on the second occasion.

Scores

Performance in each test was represented by the number of correct responses. Because the different versions of the tests had somewhat different means, scores on the second and third versions were adjusted for each participant with regression equations. The adjustment procedure was necessary because all participants received the test versions in the same order, and thus, test version was confounded

with presentation order (e.g., version A was always presented in Session 1, etc.). A separate sample of adults with a wide range of ages was therefore administered the task versions in counterbalanced order (e.g., 1/3 of the participants received version A in Session 1, 1/3 received version A in Session 2, and 1/3 received version A in Session 3). Regression equations were then computed relating scores on versions B and C to the scores on version A to determine relations among the mean scores in the versions when there was no confounding of test version and sequence order (see Salthouse, 2007 for further details). Finally, the parameters of these regression equations were used in the current study to adjust for version differences without distorting any possible sequence effects.

For some of the analyses, the scores were converted to *z*-score units relative to the first assessment. This was accomplished by first computing the mean and standard deviation of the scores at the first assessment in the first occasion. Each score was then subtracted from the first assessment mean and divided by the first assessment standard deviation to create *z*-scores based on the mean and standard deviations of the first assessment in the first occasion.

Terminology

Test scores were designated according to the occasion (i.e., T1 for Time 1 and T2 for Time 2) and assessment number (i.e., 1, 2, or 3) with the first assessment at the first occasion designated T11, the second assessment at the first occasion designated T12, etc. Longitudinal differences were labeled according to the type of contrast, with the difference between scores at the first assessment on each occasion designated T21 – T11, the difference between the averages in the two occasions designated T2avg – T1avg, the difference between averages divided by the individual's average within-occasion standard deviation designated [T2avg – T1avg]/*SD*, and the difference between scores predicted from within-burst regression equations designated PredT21 – PredT13.

Results

Within-Person Variability

After converting all variables into *z*-scores based on the T11 distribution, the standard deviations of each individual's three scores for each cognitive test on the second occasion were computed. The median within-individual *SD* was 0.43, with a range across tests from 0.34 to 0.52. To place these values in context, the median between-person *SD* of the first score at the first occasion (i.e., T11) was 0.97, and the median cross-sectional slope was 0.02. The median within-person variability was therefore about 22 times larger than the expected annual difference based on cross-sectional comparisons and about 45% the magnitude of the variability apparent across participants on the first assessment.

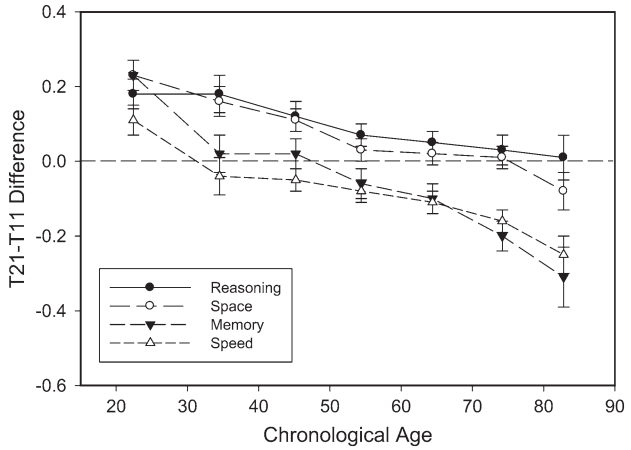

Figure 2. Means (and *SE*s) of T21 – T11 differences for composite variables representing the four cognitive abilities as a function of age.

Within-Burst Analyses

Linear regression equations were computed between the three scores at each measurement occasion and their sequence position (i.e., 1, 2, or 3) for every participant in the double burst sample. The average slopes of these regression equations were positive for every cognitive variable at both occasions, indicating that performance improved across successive sessions within each occasion.

The regression equations from each occasion were used to generate two predicted values for each test for every participant. One value corresponded to the predicted value for the T13 assessment, representing the final level of performance on the initial occasion. The other value corresponded to the predicted value for the T21 assessment, representing the initial level of performance on the second occasion. An estimate of across-occasion change was derived by subtracting the predicted T13 value from the predicted T21 value (i.e., PredT21 – PredT13).

Longitudinal Changes

Differences between single scores (T21 – T11), between averages (T2Avg – T1Avg), between averages scaled in the average within-person standard deviation ([T2Avg – T1Avg]/*SD*), and between predicted values (PredT21 – PredT13) were computed for each variable. The differences between single scores were based on the entire sample, and the remaining differences were based on only the double-burst sample.

In order to illustrate the age trends in longitudinal change, Figure 2 portrays the T21 – T11 differences for composite scores representing each cognitive ability as a function of age decade. Note that the longitudinal changes are positive

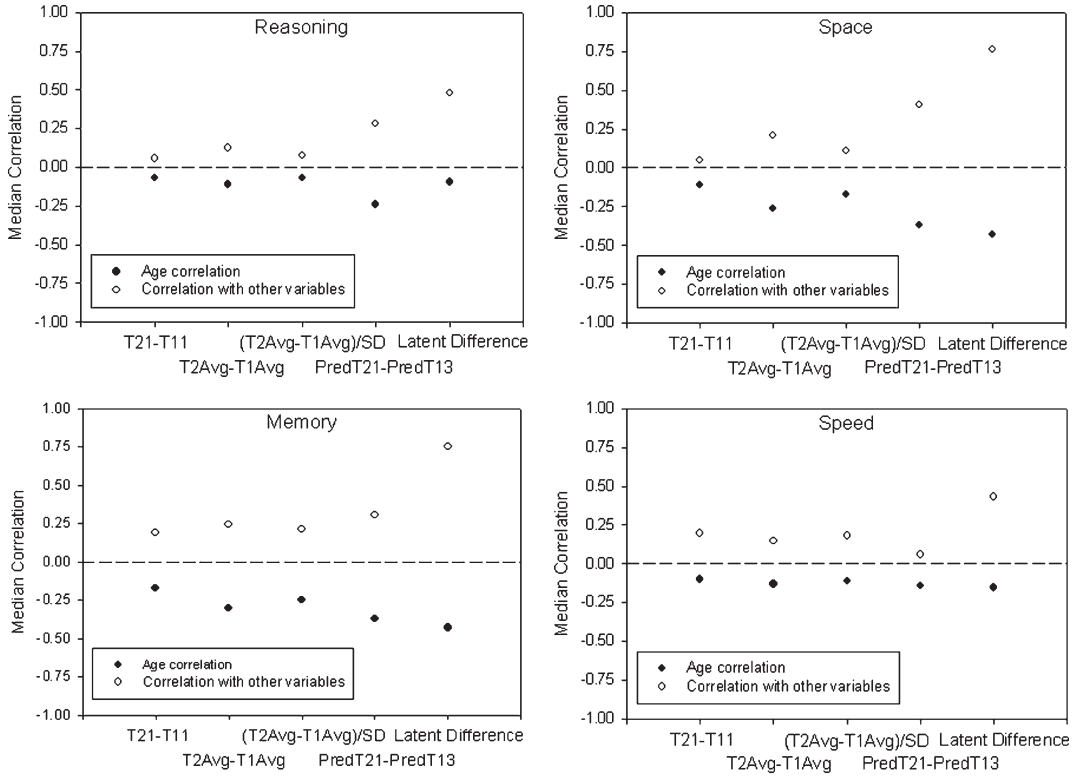

Figure 3. Median correlations among changes in variables representing the same cognitive ability and median age correlations, based on five different methods of analyzing change.

at young ages and become increasingly more negative (or less positive) at older ages and that the overall age trends were approximately linear.

Latent Difference Score

The latent difference score analysis was based on the analytical model introduced by McArdle and Nesselroade (1994). The model, which is schematically illustrated in the bottom right panel of Figure 1, consists of two latent constructs, a latent level (L) construct based on scores for the three assessments at both measurement occasions and a latent difference (D) construct based on scores for the three assessments at the second measurement occasion in addition to the scores from the first measurement occasion. Order of measurements within occasions were ignored in this model. A covariance was specified between the latent level and latent difference constructs, and covariances were also specified between residuals for the same test version on the two occasions. Although not essential to identify the model, all the regression coefficients from the latent constructs to the observed variables were fixed to 1, whereas all the variances and covariances were freely estimated.

Sensitivity of Changes

Correlations among the change scores for the variables representing the same ability were computed and then the median determined for each ability. Next correlations of the change scores with age were computed and medians determined for the variables representing the same ability. The latent change models with the Shipley Abstraction variable did not yield admissible solutions, and therefore, the medians for the Reasoning ability are based on only one value for the between-variable correlations and on two values for the age-variable correlations.

The two sets of medians for each type of longitudinal change are portrayed for the four cognitive abilities in the four panels of Figure 3. In the format of this figure, greater sensitivity corresponds to a larger deviation of the values from zero. Although there is clearly variation across abilities, several consistent trends are apparent. For example, with each ability the largest median correlation with the changes in other variables representing the same ability occurred with changes derived from the latent difference method. At least with this criterion, therefore, changes based on latent difference scores appear to be the most sensitive of the five types of change measures examined.

A second consistent pattern in Figure 3 is that the age correlations were all negative, indicating less positive, or more negative, longitudinal changes with increasing age. Nonlinear age relations on the change scores were also examined. In order to minimize collinearity, prior to these analyses, the age variable was centered and then the age-centered variable was squared to create a quadratic age

Table 2. Estimates of Effect Sizes for Mean Change Correlations Relative to T21 – T11 Changes

	Correlations with changes in other variables	Correlations with age
T2Avg – T1Avg	0.62	0.75
(T2Avg – T1Avg)/*SD*	0.29	0.41
PredT21 – PredT13	1.52	1.23
Latent difference	5.87	1.42

Note: Values were obtained by subtracting the mean correlation based on the T21 – T11 change from each type of change and then dividing the difference by the standard deviation of the T21 – T11 changes.

term. Both the age-centered and squared age-centered variables were entered simultaneously in multiple regression analyses to determine the standardized beta coefficients for the linear and quadratic age relations in the prediction of each cognitive variable. The quadratic trend was significant in only 2 of the 60 comparisons, and in both cases (i.e., the difference between averages and the difference between predicted values), they involved the word recall variable. The generally linear age relations suggest that longitudinal cognitive change occurs continuously across adulthood and not abruptly at some discrete age.

There was less variation across change measures in the magnitude of the correlations with age than in the magnitude of the correlations among measures representing the same cognitive ability. It is nevertheless worth noting that the median age correlations were largest for changes based on the latent difference score model in the two abilities with the largest age correlations, that is, Space and Memory.

The following procedure was used to provide quantitative estimates of the relative sensitivity of the various measures of change. First, means and standard deviations of the Fisher *r*-to-*z* transformed correlations were computed across all abilities for each type of change. Second, after converting the *z*-scores back to correlations, the mean correlation with each type of change was expressed in standard deviation units of the T21 – T11 change by subtracting the mean T21 – T11 change from the mean target change and dividing by the standard deviation of the T21 – T11 change. These values, which are analogous to effect sizes relative to the conventional method of computing longitudinal changes, are summarized in Table 2.

Inspection of the entries in Table 2 reveals that all the effect sizes were positive, indicating that changes based on multiple observations at each occasion were more sensitive than changes based on a single observation at each occasion. However, with both types of correlations, the effect size estimates were largest for changes based on the latent difference score method and considerably larger for correlations among changes in variables representing the same cognitive ability.

Discussion

As expected from the results of Salthouse (2007), the magnitude of short-term fluctuation in cognitive performance was substantial. One indication of the size of the phenomenon is available in a comparison of the median within-person standard deviation, which was 0.43, with the median slope for the cross-sectional age relation, which was 0.02 *SD* per year. With this particular contrast therefore, the short-term fluctuation is nearly 22 times greater than the annual differences expected on the basis of cross-sectional comparisons. The existence of sizable variability in measures of cognitive functioning that are often assumed to reflect stable trends is surprising, and both Salthouse (2007) and Salthouse and colleagues (2006) speculated about the nature and causes of this variability.

Whether one conceptualizes intra-individual variability as "noise" or "signal," individual differences in the magnitude of intra-individual variability could have important implications for the interpretation of differences and changes in cognitive performance. Of particular relevance in the current context is that because the magnitude of short-term fluctuation varies across people, it could impact the meaning of longitudinal changes since the same absolute difference corresponds to a larger proportion of an individual's short-term fluctuation for someone with small within-person variability than for someone with large within-person variability.

This fluctuation is likely to affect the sensitivity of longitudinal change because some of what is interpreted as change could be attributable to vagaries of sampling at each occasion. What might be the ideal approach to dealing with this problem is to identify the determinants of the short-term fluctuation and then control them to minimize their influences. In the current project, all the within-occasion assessments for a given participant were carried out in the same season of the year and most were at the same time of day in very similar rooms and with the same standardized protocol. The fluctuations are therefore unlikely to be attributable to factors in the physical environment but instead to factors within the individual such as quantity or quality of sleep, diet, exercise, motivation, level of stress, etc. Unfortunately, relatively little is currently known about the causes of either the endogenous or exogenous factors contributing to short-term fluctuation in cognitive performance and even less about how to control their influences.

Four methods of dealing with short-term fluctuation in longitudinal research were examined in the current report: differences between averages, differences between averages calibrated in each participant's own within-person variability, differences between scores predicted from within-occasion regression equations, and latent difference scores. One method of evaluating the sensitivity of longitudinal change consisted of computing correlations among the changes for variables representing the same ability. The rationale was that the observed correlations among the changes provide a lower bound estimate of reliability of the changes. This method revealed that the largest correlations were with changes derived from the latent difference score procedure.

The second method used to evaluate change sensitivity consisted of comparing correlations of the changes with a

different variable, namely, age. All the changes, including changes based on single scores, had qualitatively similar patterns of age relations. Nevertheless, the largest average age correlations were with changes based on the latent difference score procedure.

Change estimates based on calibration of the changes between averages in terms of one's own variability were intermediate in age sensitivity between the changes in averages and the changes in single scores. This relatively low sensitivity may be attributable to the fact that the estimates of within-person variability were based on only three assessments and thus are not as precise as they might be if based on more assessments. Calibration of change in terms of each individual's own variation might have been more sensitive if there had been a greater number of assessments to yield more precise estimates of short-term fluctuation. However, depending on the causes of the short-term fluctuation, it is also possible that little gain in sensitivity could be achieved with this method.

The change measure with the greatest sensitivity was that based on the latent difference model. This is not surprising because in this model, change is represented by a latent construct formed from variance shared among variables at the second occasion after controlling the variance in a latent construct formed from the variance shared among variables at both the first and second occasions. Rather than treating variability in performance within the same occasion as a problem, the latent difference method exploits the variation to derive latent constructs representing only the systematic variance among the variables at each occasion. Moreover, because only systematic or reliable variance can be shared, latent constructs have no measurement error. Another advantage of analyses based on latent constructs is that they can easily accommodate missing data. Although change derived from latent difference models can be quite sensitive, it should be recognized that there are some limitations of this procedure. For example, only group-level estimates of the mean and variance of change are available with no information at the level of individuals, and relatively large sample sizes are needed for these types of analyses. Moreover, in some cases, the estimates cannot be derived because of violations of the underlying assumptions, as was the case with the Shipley Abstraction reasoning variable in the current project.

The focus in the current study was on conceptually simple versions of each analytic method, and more complex, and potentially more powerful, versions could clearly be considered. For example, instead of unit-weighted aggregation to form composite scores, factor analyses could be performed to allow each session score to be weighted according to its contribution to the factors representing performance on each occasion. In addition, the interrupted time series model might be reconfigured as a multilevel model (e.g., Hoffman, 2007), and the latent difference method could be elaborated to allow weightings of session scores to vary in their contributions to the latent constructs representing level and change. Although these more complex variants may turn out to have greater sensitivity and power than the relatively simple methods used in this study, they all capitalize on the availability of multiple scores at each occasion to provide more sensitive assessment of change without capturing the intuition that the estimate of change for an individual should be considered less precise when the observations entering into change are more variable. The method of dividing the across-occasion difference by the average within-occasion variability does incorporate this property, but the estimates of within-occasion variability are not very precise when they are based on only three measurements and thus the resulting variability-adjusted changes were not very sensitive. Although it may not be practical to calibrate change in terms of one's within-occasion variability, it is still the case that across-occasion change is likely to be less meaningful for an individual with large within-occasion variability compared with an individual with lower level of within-occasion variability. Additional methods that combine estimates of the magnitude of change with information about the precision of those estimates should therefore continue to be explored.

The assessment of cognitive change remains a prominent theme of gerontological research, but there are many approaches to this topic. We examined four that can be used within the context of measurement-burst designs. Multiple assessments in measurement-burst designs are associated with an increase in cost and participant burden, and it is reasonable to ask whether the benefits exceed the costs. Unfortunately, there is unlikely to be a simple answer to this question. On one hand, it is true that the age trends with these more expensive and time-consuming procedures are generally similar to those with the traditional contrast of single scores. On the other hand, some indices of sensitivity, such as the correlations among changes in variables from the same ability, reveal greater sensitivity with changes based on the latent difference score model, which relies on multiple assessments at each occasion to derive latent constructs.

In conclusion, the results described in this report confirm and further elaborate the existence of large short-term fluctuation in cognitive performance which could affect the interpretation of longitudinal change because different estimates of change could be obtained depending upon the particular scores that happen to be compared across the two occasions. Aggregation of measures across several assessments at each occasion results in greater precision in the estimates of change, and the current results suggest that in many cases, the precision will be greatest with changes derived from the latent difference score model or from conceptually similar latent growth curve models.

Funding

This research was supported by National Institute on Aging grant R37AG024270 to T.A.S.

Correspondence

Address correspondence to Timothy A. Salthouse, PhD, Department of Psychology, University of Virginia, Charlottesville, VA 22904. Email: salthouse@virginia.edu

References

Bunce, D., Handley, R., & Gaines, S. O. (2008). Depression, anxiety, and within-person variability in adults aged 18 to 85 years. *Psychology and Aging*, *23*, 848–858.

Duchek, J. M., Balota, D. A., Tse, C.-S., Holtzman, D. M., Fagan, A. M., & Goate, A. M. (2009). The utility of intraindividual variability in selective attention tasks as an early marker for Alzheimer's disease. *Neuropsychology*, *23*, 746–758.

Gorus, E., de Raedt, R., Lambert, M., Lemper, J.-C., & Mets, T. (2008). Reaction times and performance variability in normal aging, mild cognitive impairment, and Alzheimer's disease. *Journal of Geriatric Psychiatry and Neurology*, *21*, 204–218.

Hoffman, L. (2007). Multilevel models for examining individual differences in within-person variation and covariation over time. *Multivariate Behavioral Research*, *42*, 609–629.

Hultsch, D. F., Strauss, E., Hunter, M. A., & MacDonald, S. W. S. (2008). Intraindividual variability, cognition, and aging. In F. I. M. Craik & T. A. Salthouse (Eds.), *Handbook of aging and cognition* (3rd ed., pp. 491–556). New York: Psychology Press.

McArdle, J. J., & Nesselroade, J. R. (1994). Structuring data to study development and change. In S. H. Cohen & H. W. Reese (Eds.), *Life-span developmental psychology: Methodological innovations* (pp. 223–268). Hillsdale, NJ: Erlbaum.

MacDonald, S. W. S., Hultsch, D. F., & Dixon, R. A. (2008). Predicting impending death: Inconsistency in speed is a selective and early marker. *Psychology and Aging*, *23*, 595–607.

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 Sciences*, *59B*, P49–P55.

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

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

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

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.

Salthouse, T. A., Pink, J. E., & Tucker-Drob, E. M. (2008). Contextual analysis of fluid intelligence. *Intelligence*, *36*, 464–486.

Salthouse, T. A., & Tucker-Drob, E. M. (2008). Implications of short-term retest effects for the interpretation of longitudinal change. *Neuropsychology*, *22*, 800–811.

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.