Journal of the International Neuropsychological Society (2012), **18**, 749–756.

doi:10.1017/S1355617712000380

Robust Cognitive Change

Timothy A. Salthouse
University of Virginia, Charlottesville, Virginia

(Received July 14, 2011; Final Revision February 28, 2012; Accepted February 28, 2012)

Abstract

Two major challenges facing researchers interested in cognitive change are that measures of change are often not very reliable, and they may reflect effects of prior test experience in addition to the factors of primary interest. One approach to dealing with these problems is to obtain multiple measures of change on parallel versions of the same tests in a measurement burst design. A total of 783 adults performed three parallel versions of cognitive tests on two occasions separated by an average of 2.6 years. Performance increased substantially across the three sessions within each occasion, and for all but vocabulary ability these within-occasion improvements were considerably larger than the between-occasion changes. Reliabilities of the changes in composite scores were low, but averages of the three changes had larger, albeit still quite modest, reliabilities. In some cognitive abilities individual differences were evident in the relation of prior test experience and the magnitude of longitudinal change. Although multiple assessments are more time consuming than traditional measurement procedures, the resulting estimates of change are more robust than those from conventional methods, and also allow the influence of practice on change to be systematically investigated. (*JINS*, 2012, *18*, 749–756)

Keywords: Longitudinal, Aging, Neurocognitive, Reliable change, Measurement burst, Gains and losses

INTRODUCTION

Because it may help identify the beginning of neurodegenerative disease, and is important in monitoring recovery from trauma and the effects of interventions, there has been considerable interest in detecting cognitive change as early, and with the greatest sensitivity, as possible. However, investigation of cognitive change has been difficult because of the existence of measurement error, which contributes to low reliability of measures of change, and because effects of prior test experience may distort the estimates of change.

Low reliability of the measures of change is associated with at least two problems. First, the changes may not be meaningful if one cannot be confident that results from a second change assessment would be similar to those from the first change assessment. And second, when the tests do not have perfect reliability, which is almost always the case, regression toward the mean can occur and produce misleading estimates of change. That is, very high scores may be partially attributable to positive measurement error and very low scores may be partially attributable to negative measurement error. Extreme scores at the first occasion will, therefore, often be less extreme at the second occasion, such that individuals with the lowest initial values will tend to exhibit positive change, and individuals with the highest initial scores will tend to exhibit negative change.

Practice effects also complicate the detection and interpretation of change because some of the observed change may be attributable to effects of prior test experience and not to the factors of primary interest. Unless they are considered in the analyses, practice effects can obscure true declines, and inflate estimates of gains.

These problems have been addressed in different ways by researchers interested in group trends, and researchers interested in evaluating change within individuals. For example, when the interest is in group results, statistical models are often used in which change is represented as a latent construct that theoretically has no measurement error (e.g., Ferrer & McArdle, 2010). Furthermore, if the dataset involves three or more measurement occasions, or variable retest intervals, models can be specified in which effects of practice are distinguished from effects of other determinants of change (e.g., Ferrer, Salthouse, Stewart, & Schwartz, 2004; Salthouse, Schroeder, & Ferrer, 2004). The major limitations of these models are that moderately large samples are required to obtain stable estimates, and the results are applicable only to groups and not to individuals.

Many neuropsychologists and clinical psychologists are primarily interested in assessing the magnitude of change at

Correspondence and reprint requests to: Timothy A. Salthouse, Department of Psychology, University of Virginia, Charlottesville, Virginia 22904-4400. E-mail: salthouse@virginia.edu

the level of individuals, and determining the likelihood that the change is not attributable to chance, regression toward the mean, or practice effects. Individual change could be evaluated by simply comparing the change in a particular individual with the distribution of changes in a control sample. However, several researchers have proposed more elaborate methods of evaluating change with versions of the reliable change index (see reviews in Collie et al., 2004; Hinton-Bayre, 2010; Tempkin, Heaton, Grant, & Dikmen, 1999). The primary interest with these methods has been in evaluating whether change in the individual was atypical, and different from what would be expected by chance after consideration of measurement reliability, practice, and other factors.

Although these methods are useful for determining whether the change observed in a given individual differs from what might be expected, it is important to recognize that they do not solve the problem that measures of change often have low reliability. To illustrate, one method of assessing the reliability of change is to determine the correlation between two separate changes involving different versions of the same tests in the same individuals. Correlations of this type were computed in the current study for the reliable change index (Jacobson & Truax, 1991) and for a regression-based change index (McSweeny, Naugle, Chelune, & Luders, 1993). The correlations with the reliable change index for a word recall measure and for a digit symbol measure were .19 and .24, respectively, and those with the regression-based change were .29 and .27, respectively. Each of these correlations was significantly greater than zero, but they are quite low when considered as estimates of reliability. Furthermore, the other measures in the current project had similar correlations, and, therefore, the unreliability of reliable change indices is not specific to only a few neuropsychological measures.

The rationale for the current study was that more sensitive and reliable assessments of individual change might be obtained by capitalizing on various design features. For example, if multiple tests of each construct are available, measurement error can be reduced by relying on the principle of aggregation (Rushton, Brainerd, & Pressley, 1983), and conducting analyses on composite scores rather than on scores from individual tests. Furthermore, if a measurement burst design is used in which three parallel versions of the tests are administered at each longitudinal occasion, separate changes can be computed based on the longitudinal contrast with the first test version, with the second test version, and with the third test version. The availability of multiple changes from the same individuals allows two conceptually distinct aspects of change to be evaluated. First, the separate changes can be averaged to produce an aggregate measure of change that should be a better estimate of the individual's true change, and have greater reliability, than any single measure of change. And second, the separate measures of change can be examined to determine the consistency of the change. Average and consistency are conceptually distinct because, for example, the same average change could be achieved with a consistent pattern of three moderate negative changes, or with an inconsistent pattern of two small positive changes and one large negative change.

Another advantage of administering multiple test versions at each occasion is that change on the first version can be compared with change on later versions to examine the effects of prior testing experience on change. That is, because each successive test version is associated with progressively more test experience, changes on later test versions can be assumed to be less affected by practice effects than change on the first test version.

The analyses in the current study capitalized on a measurement-burst design in which parallel versions of each cognitive test were administered across three sessions completed within a period of approximately 2 weeks. The first session on the first occasion was designated 11, the second session on the first occasion 12, the third session on the first occasion 13, the first session on the second occasion 21, etc. The availability of scores on three sessions at each occasion, therefore, allowed three separate longitudinal changes to be computed: change on the first session (i.e., from 11 to 21), change on the second session (i.e., from 12 to 22), and change on the third session (i.e., from 13 to 23). Another feature of the design was that each of five cognitive abilities was represented by either three or four separate tests, which allowed the analyses to be conducted on more reliable composite variables rather than scores from individual tests. Finally, because the different types of changes could vary across individuals, relations of change were examined with three individual difference variables: age, years of education, and Mini Mental State Exam (MMSE; Folstein, Folstein, & McHugh, 1975) score on the second occasion (T2).

METHODS

Sample

The sample consisted of 783 healthy adults from the Virginia Cognitive Aging Project (Salthouse, 2007, 2010) who each completed two three-session measurement bursts separated by an average interval of approximately 2.5 years. Most participants completed the three sessions of each measurement burst within a period of approximately 2 weeks. As reported in Salthouse (2010), the participants with longitudinal data had somewhat higher average cognitive scores than did participants who only completed one occasion, with the exception of the younger adults for whom the returning participants had somewhat lower initial scores than the non-returners. All data were collected with the approval of the local Institutional Review Board.

To examine possible age differences in change, the sample was divided into four age groups: ages 18–39, 40–59, 60–79, and 80–95. Characteristics of the individuals in each group are reported in Table 1, where it can be seen that there were slightly more years of education, but lower ratings of health, among the older participants. The age-adjusted scaled scores were somewhat higher at older ages, suggesting that

Table 1. Characteristics of participants

	Age group				
	18–39	40–59	60–79	80–95	Age corr.
N	148	313	268	54	NA
Age	26.8 (6.8)	51.2 (5.5)	68.9 (5.7)	83.4 (3.5)	NA
Proportion Females	.59	.73	.63	.44	−.04
Self-rated health	2.3 (0.9)	2.2 (0.9)	2.4 (0.9)	2.5 (0.9)	.10*
Years of Education	14.7 (2.2)	15.7 (2.6)	16.1 (2.7)	16.5 (2.8)	.21*
Scaled Scores					
Vocabulary	12.4 (2.9)	12.1 (3.1)	13.3 (2.7)	14.0 (2.2)	.16*
Digit Symbol	11.1 (2.6)	11.4 (2.9)	11.5 (2.6)	12.8 (2.5)	.16*
Logical Memory	11.4 (2.8)	11.6 (2.9)	12.4 (2.7)	12.5 (3.1)	.15*
Word Recall	11.8 (3.7)	12.0 (3.7)	12.6 (3.3)	11.4 (3.5)	.05
T1 MMSE	28.7 (1.5)	28.7 (1.6)	28.4 (1.7)	27.5 (2.1)	−.15*
T2 MMSE	28.7 (1.6)	28.6 (1.6)	28.4 (1.8)	27.2 (2.3)	−.17*
T1-T2 Interval (years)	2.5 (0.9)	2.5 (0.9)	2.6 (0.9)	2.4 (0.7)	−.01

Note. *$p < .01$. NA indicates the value is not applicable. Values in parentheses are standard deviations. Self-rated health was a rating on a scale ranging from 1 for "excellent" to 5 for "poor." Scaled scores are age-adjusted scores from the WAIS III (Wechsler, 1997a) and WMS III (Wechsler, 1997b), in which the means and standard deviations in the nationally representative normative sample were 10 and 3, respectively. MMSE is the Mini-Mental Status Exam (Folstein, Folstein & McHugh, 1975). The final column contains the correlation of age with the variable in the entire sample.

the older participants in the sample had higher initial levels of functioning relative to their age peers than the younger participants. Approximately 85% of the participants in the total sample reported their ethnicity as Caucasian, approximately 7% as African-American, with the remainder distributed across other ethnicities, or reporting more than one ethnicity.

Cognitive Tests

In each session participants performed 16 cognitive tests designed to represent five cognitive abilities. Episodic memory was assessed with the Logical Memory and Word List Recall tests from the Wechsler Memory Scale III (Wechsler, 1997b), and a Paired Associates test developed locally (Salthouse, Fristoe, & Rhee, 1996). Perceptual speed was measured with the Digit Symbol (Wechsler, 1997a), and Letter Comparison and Pattern Comparison tests (Salthouse & Babcock, 1991). Vocabulary was measured with WAIS III Vocabulary (Wechsler, 1997a), Woodcock-Johnson Picture Vocabulary (Woodcock & Johnson, 1989), and Antonym and Synonym Vocabulary (Salthouse, 1993) tests. Reasoning was assessed with the Raven's Advanced Progressive Matrices (Raven, 1962) test, the Shipley Abstraction (Zachary, 1986) test, and the Letter Sets test from the Educational Testing Service Kit of Factor-Referenced Cognitive Tests (Ekstrom, French, Harman, & Dermen, 1976). And finally, spatial visualization (space) was assessed with the Spatial Relations test from the Differential Aptitude Test Battery (Bennett, Seashore, & Wesman, 1997), and the Paper Folding and Form Boards tests (Ekstrom et al., 1976). Descriptions of the tests, as well as information about reliability, and validity in the form of confirmatory factor analyses indicating the pattern of relations of variables to ability constructs, are contained in other articles (Salthouse, 2004, 2005, 2010; Salthouse & Ferrer-Caja, 2003; Salthouse, Pink, & Tucker-Drob, 2008).

Three different versions of each of the 16 tests were created with different items but an identical format. Because the versions could differ in mean performance, a separate sample of 90 adults between 20 and 79 years of age performed the three versions of the tests in a counterbalanced order to remove the confounding of test version and presentation order desirable in a study designed to investigate individual differences (Salthouse, 2007). That is, when the primary interest is differences between people, it is preferable to treat everyone the same and avoid between-person counterbalancing. Regression equations in the counterbalanced sample were used to predict performance in the original version from the scores on the second or third versions, and the intercepts and slopes of these equations were then used to adjust the scores of every participant on the second and third versions to remove the order-independent version differences in the means. To illustrate, the intercept and the slope for the function predicting matrix reasoning scores on the first version from those on the second version were 1.86 and .84, respectively. Applying these parameters to an individual with a session 2 score of 12 would result in an adjusted score of $(1.86 + .84 \times 12) = 11.94$.

RESULTS

All original and adjusted test scores were converted into *Z*-scores units based on the T11 distributions, and then composite scores were formed by averaging the *Z*-scores for the three (or four for vocabulary) variables representing each ability in each session. Longitudinal change was computed by subtracting the T1 composite score from the T2 composite

score for corresponding sessions (i.e., T21–T11, T22–T12, and T23–T13). Average change was computed by averaging the changes across the three sessions. Consistency of change was evaluated both in terms of correlations between changes in different sessions, and with a measure of the degree of agreement of the dichotomous outcome (i.e., increase or decrease) across sessions. Correlations are appropriate for evaluating similarity of continuous variables, but other methods are desirable to evaluate the consistency of a binary classification. This latter type of consistency was computed by categorizing the individual as having improved or declined according to whether the T2x–T1x difference in the composite score for a given session (x) was positive (coded 1) or negative (coded 0). The change codes were then summed across the first (T21–T11), second (T22–T12), and third (T23–T13) sessions. A sum of 0, therefore, indicates a consistent pattern of decline (or loss) in the composite scores in all three sessions, and a sum of 3 indicates a consistent pattern of improvement (or gain), with intermediate values corresponding to decline in one or two of the sessions.

An initial set of analyses examined reliabilities of the first session changes and of the average change across the three sessions. Correlations between change on the first (i.e., T21–T11) and second (i.e., T22–T12) sessions were used as alternate-form estimates of reliability for the session 1 change, and coefficient alphas based on the changes in the three sessions were used as the estimate of internal consistency reliability for the average change. The correlations between sessions 1 and 2 and the coefficient alphas for each ability were, respectively: memory .25 and .46, speed .31 and .59; vocabulary .15 and .32; reasoning .19 and .31; and space .23 and .36. The estimated reliability of the average changes (range from .31 to .59) were somewhat higher than those for the session 1 changes (range from .15 to .31), but all were quite low.

Table 2 contains correlations of age, years of education, and T2 MMSE scores with the change at each session, average change, consistent gain, and consistent loss. Inspection of the entries in the table reveals that the patterns in the three sessions were generally similar, with the exception of more negative relations of education on vocabulary change in the first session than in later sessions. Increased age was associated with more negative average changes in every ability, and higher T2 MMSE scores were associated with more positive change in memory ability and in vocabulary ability. Although smaller, the patterns with consistent gains and consistent losses were similar to those with average change.

The strongest change relations were with the age variable, and, therefore, subsequent analyses focused on age as the individual difference variable of interest. Figure 1 portrays relations of age to average change and to the two types of consistent change. Although the average and consistency measures are conceptually independent, it is noteworthy that the average change and the two measures of consistent change had similar age trends, albeit in the opposite direction for the consistent losses compared to the average and consistent gain measures.

Table 2. Correlations of age, education and T2 MMSE score with measures of T2-T1 change

	Session			Consistent		
Ability	1	2	3	Average	Gain	Loss
Memory						
Age	−.29*	−.23*	−.22*	−.34*	−.20*	.23*
Education	−.02	−.05	−.04	−.05	−.05	.06
T2 MMSE	.10*	.15*	.10	.13*	.12*	−.14*
Speed						
Age	−.17*	−.13*	−.14*	−.20*	−.20*	.11*
Education	−.03	−.08	−.06	−.07	−.05	.07
T2 MMSE	.05	.01	−.02	.02	.04	−.00
Vocabulary						
Age	−.18*	−.22*	−.20*	−.30*	−.21*	.16*
Education	−.14*	.02	−.06	−.09	−.10	−.00
T2 MMSE	.11*	.11*	.09	.17*	.12*	−.08
Reasoning						
Age	−.14*	−.15*	−.12*	−.18*	−.11*	.16*
Education	−.04	.02	−.07	−.03	−.05	−.04
T2 MMSE	.06	.07	−.02	.08	.06	−.10
Space						
Age	−.21*	−.13*	−.18*	−.26*	−.16*	.14*
Education	−.01	−.02	−.08	−.04	−.06	.04
T2 MMSE	.12*	.08	−.02	.11	.05	−.04

Note. *$p < .01$. The three individual difference variables were all continuous with age and education in years and MMSE (Folstein et al. 1975) at the second occasion in score points.

Means and standard errors for the ability composites on the first, second, and third sessions in the two longitudinal occasions are portrayed in Figure 2. To improve visibility in the figure, the sessions are separated by 1 year rather than the actual 1 week, and the longitudinal occasions are separated by 5 years rather than the actual 2.5 years. The lines connecting the data points represent the direction and magnitude of longitudinal change in each session and each age group. Parallel lines indicate that the longitudinal change was similar in direction and magnitude across sessions, whereas non-parallel lines suggest that the change differed across sessions.

Three general trends are apparent in Figure 2. First, all of the changes were positive at younger ages, and either close to zero or negative at older ages. Second, with the exception of speed and vocabulary abilities, the across-session improvement in performance was greater at older ages (i.e., the vertical separation of the changes were more pronounced with increased age). And third, most of the lines representing changes on different sessions were nearly parallel, suggesting similar direction and magnitude of the change in each session.

These observations were formally investigated with analyses of variance in which the three sessions and two times (occasions) were within-subjects factors, and age (four groups) was a between-subjects factor. Results of these analyses, and estimates of the means for the major contrasts, are reported in Table 3. The main effect of time represents the longitudinal change averaged across the three sessions and the four age groups. Significant effects of time were evident with reasoning

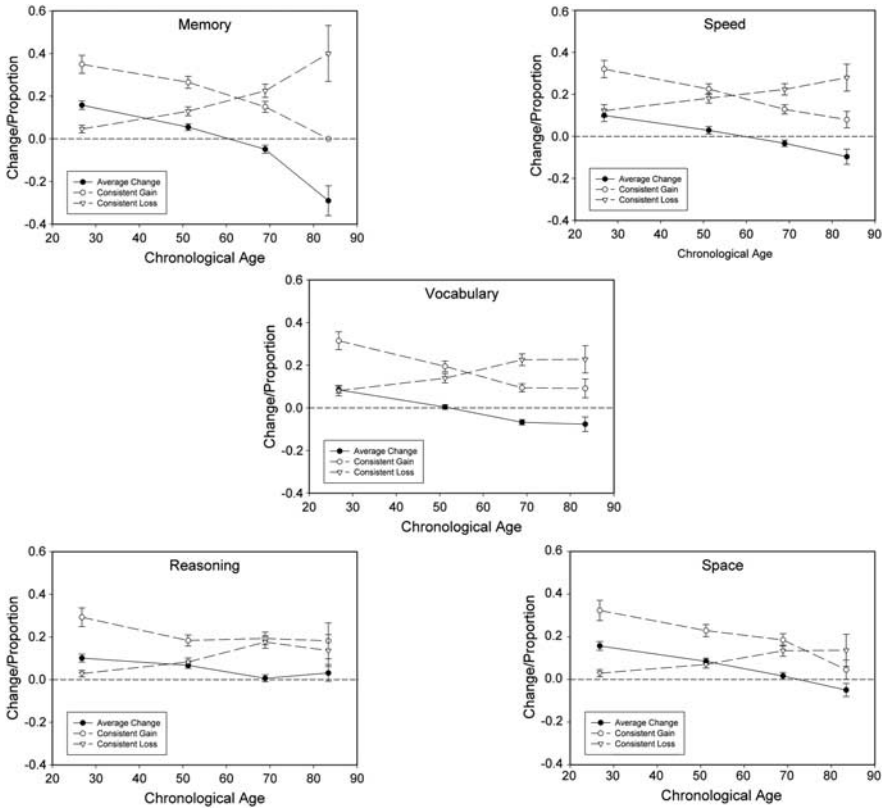

Fig. 1. Means and standard errors of average change in composite scores for five cognitive abilities across the three sessions (solid symbols and lines), and of the proportion of individuals with consistent gain or consistent loss across the three changes.

and space abilities, in the direction of increases from T1 to T2. The main effect of session indicates change across the three sessions averaged across the two occasions (time) and the four age groups. The session effects were significant in the direction of higher scores on successive sessions for every ability except vocabulary. The time × session interaction indicates whether the longitudinal change varied according to session when averaged across age groups. The interaction was only significant with space ability, in the direction of less positive change on later sessions.

All of the main effects of age were significant in these analyses, but they largely reflect cross-sectional trends because the scores were averaged across session and occasion. The age × time interactions were also significant for every ability, reflecting the negative correlations of age with average longitudinal change (cf., Table 2). The age × session interactions were significant in the direction of greater gains for older adults for all but vocabulary ability, in which the across-session gain was greater for younger adults. The age × time × session interaction indicates whether the longitudinal change differed according to both session and age. The interaction was significant for memory ability, in the direction of stronger age relations (i.e., greater gains at young ages and greater losses at older ages) for session 1 change than for change on sessions 2 or 3.

A final set of analyses consisted of correlations between the average within-session change across the two occasions (i.e., average of 11 to 13 and 21 to 23), and the average between-occasion change across the three sessions (i.e., average of 11–21, 12–22, and 13–23). These correlations are reported in Table 4, along with coefficient alpha estimates of the reliability of the averages. The reliability estimates were modest (i.e., medians of .59 for the within-occasion changes and .36 for the between-occasion changes), but the correlations were even smaller among both the within-occasion changes (median of .11) and the between-occasion changes (median of .17). There was also no evidence of a relation between the within-occasion and between-occasion changes as the median was −.03 with the same ability and −.05 across different abilities.

DISCUSSION

Although sometimes referred to as indices of reliable change, several derived measures of change for common neuropsychological tests had weak correlations in parallel tests, and thus can be inferred to have low replicability. Because measures involving monotonic transformations of the observed change, such as subtraction (e.g., of a group practice effect) or division (e.g., by the standard deviation of the changes) of a constant will be perfectly correlated with simple changes, the low reliabilities will also apply to many variants of the original methods intended to account for practice or average age effects. In other words, although these various derived change measures are sometimes assumed to assess reliable change,

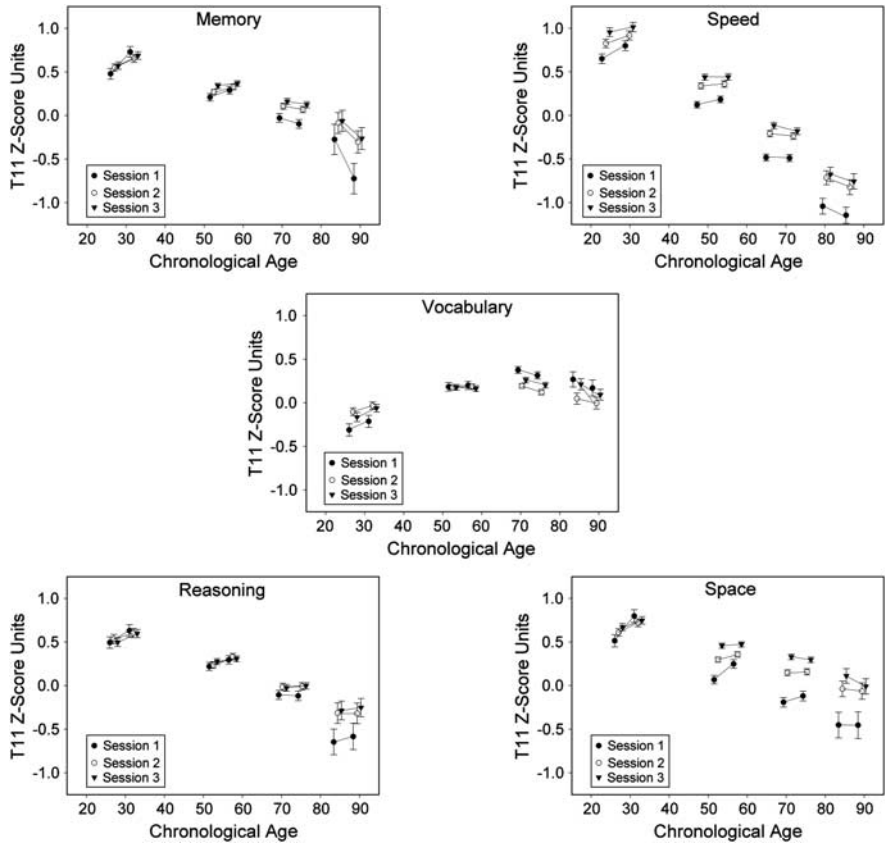

Fig. 2. Means and standard errors of composite scores on the T1 and T2 occasions for each of the three sessions.

many of them have low repeatability across parallel assessments of change.

Because it is based on three separate assessments, average change can be assumed to provide a better reflection of an individual's true change than change based on scores in a single session on each occasion. Although the average of three changes will generally be more reliable than change from a single session, it can still be influenced by extreme

Table 3. F-ratios for Mixed Effects ANOVA on composite scores with age as between-subjects variable

	Memory		Speed		Vocabulary		Reasoning		Space	
	F	*df*	*F*	*df*	*F*	*df*	*F*	*df*	*F*	*df*
Time (Longitudinal)	3.05	1,594	0.00	1,719	2.31	1,658	15.26*	1,545	15.35*	1,544
Session	31.15*	3,594	345.39*	3,719	0.72	2,1316	24.25*	2,1090	108.93*	2,1088
Time ∗ Session	0.30	2,1188	2.40	2,1438	0.05	2,1316	0.42	2,1090	15.70*	2,1088
Age	38.54*	3,594	185.71*	3,719	18.34*	3,658	38.49*	3,545	37.90*	3,544
Age ∗ Time	25.80*	3,594	9.05*	3,719	20.75*	3,658	5.35*	3,545	12.18*	3,544
Age ∗ Session	7.13*	6,1188	4.63*	6,1438	24.45*	6,1316	6.57*	6,1090	16.07*	6,1088
Age ∗ Time ∗ Session	3.46*	6,1188	0.44	6,1438	0.46	6,1316	1.44	6,1090	1.09	6,1088
Estimated Means (with SE)										
Time 1	.19 (.04)		.01 (.03)		.15 (.03)		.07 (.04)		.21 (.03)	
Time 2	.16 (.04)		.01 (.03)		.13 (.03)		.12 (.04)		.26 (.03)	
Session 1	.07 (.05)		−.18 (.03)		.14 (.03)		.02 (.04)		.05 (.04)	
Session 2	.20 (.03)		.06 (.03)		.13 (.02)		.13 (.03)		.27 (.03)	
Session 3	.24 (.03)		.14 (.03)		.15 (.02)		.14 (.03)		.39 (.02)	
T2-T1 Change										
Session 1	−.05 (.03)		.03 (.02)		−.01 (.02)		.07 (.02)		.13 (.02)	
Session 2	−.02 (.02)		−.01 (.02)		−.01 (.01)		.04 (.02)		.04 (.02)	
Session 3	−.03 (.02)		−.02 (.02)		−.02 (.02)		.05 (.02)		−.02 (.02)	

Note. *$p < .01$. Values for T2x-T1x change obtained from ANOVA on T2x-T1x difference scores.

Table 4. Correlations of within-occasion and between-occasion changes in the five cognitive abilities

	1	2	3	4	5	6	7	8	9	10
1 – Memory Within	(.49)									
2 – Speed Within	.05	(.33)								
3 – Vocabulary Within	.09	−.03	(.68)							
4 – Reasoning Within	.24*	-.01	.17*	(.59)						
5 – Space Within	.26*	.00	.13*	.43*	(.75)					
6 – Memory Between	−.03	−.02	.06	−.15*	−.16*	(.46)				
7 – Speed Between	.00	.08	.09	−.05	−.04	.13*	(.59)			
8 – Vocabulary Between	−.08	−.05	.13*	−.05	−.08	.30*	.15*	(.32)		
9 – Reasoning Between	−.08	.00	.07	−.09	−.04	.19*	.12*	.21*	(.31)	
10 – Space Between	−.12*	−.04	−.00	−.08	−.12*	.22*	.17*	.10	.17*	(.36)
Mean	.13	.31	−.01	.07	.33	.03	.01	−.01	.05	.07
SD	.37	.29	.37	.36	.48	.27	.29	.19	.21	.23

Note. *$p < .01$. Numbers in parentheses are coefficient alpha estimates of reliability based on two (Within) or three (Between) values.

values that could affect the magnitude of relations with other variables. However, extreme scores are less of a problem with measures of consistency of binary change outcomes across three separate assessments.

The availability of three separate changes also allowed change to be examined as a function of relevant test experience. It should be noted that the comparisons of changes across sessions 1, 2, and 3 each involved two additional measurements intervening between those used to assess change. That is, sessions 12 and 13 were administered between the 11 and 21 assessments used to evaluate change on session 1, and sessions 13 and 21 were administered between the 12 and 22 assessments used for change on session 2, and sessions 21 and 22 were administered between the 13 and 23 assessments used for change on session 3. This additional experience could have attenuated effects associated with the amount of test experience occurring before the first assessment used in the evaluation of change, in which case the current results may underestimate the effects of test experience on change.

The results in Figure 2, and confirmed by the session effects in Table 3, indicate that for many abilities performance improved with additional test experience on successive sessions within the same longitudinal occasion. Moreover, the time × session interaction was significant for space ability, and the age × time × session interaction was significant for memory ability, thus indicating that the magnitude of longitudinal change can vary according to the amount of relevant test experience. In particular, inferences about the relations between age and change in memory will differ according to whether change is evaluated on the first session, or on subsequent sessions after the individuals have had some experience with the tests. These results are consistent with the interpretation that when only a single assessment is available at each occasion, as in most longitudinal research, change from the first to the second occasion will likely represent an unknown mixture of gains associated with additional test experience, and losses occurring over the T1 to T2 interval.

One method of trying to distinguish the different contributions to change is to determine change after an initial practice period, as proposed with the dual-baseline procedure (e.g., Beglinger et al., 2005; McCaffrey & Westervelt, 1995). For example, in the current project the focus could be on change from either T12 or T13 to T21, instead of from T11 to T21. Although the dual-baseline procedure will likely minimize the influence of practice on the first occasion, it neglects the possibility that practice effects could also occur on the second occasion. An advantage of the measurement burst design is that the multiple assessments at each occasion allow evaluation of change with the same amount of practice at each occasion. In particular, across-occasion change on the second and third assessments in each occasion can be assumed to be less affected by test experience than change in the first assessment. The effects of practice on change for a given individual can, therefore, be evaluated by comparing change on the first assessment with change on the second or third assessment. To illustrate, the results of this study suggest that older individuals benefit more than younger individuals from additional test experience with memory ability as they exhibited less negative changes on sessions 2 and 3 than on session 1.

In summary, change at the level of the individual is limited by low reliability and unknown involvement of practice. However, reliability can be improved by aggregation of several test scores to form composite scores, and by relying on multiple changes to determine both average change and consistency of change. In addition, influences of practice can be estimated by comparing change with different amounts of prior test experience. This type of measurement burst design is more time consuming and expensive than traditional designs with a single assessment at each occasion, but procedures such as this may be necessary to obtain measures of change at the level of the individual that not only provide estimates of the influence of practice, but also allow both average change and consistency of change to be evaluated.

ACKNOWLEDGMENTS

The project was supported by Award Number R37AG024270 from the National Institute on Aging. The content is solely the

responsibility of the author and does not necessarily represent the official views of the National Institute on Aging or the National Institutes of Health. There are no conflicts of interest.

REFERENCES

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

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

Collie, A., Maruff, P., Makdissi, M., McStephen, M., Darby, D.G., & McCrory, P. (2004). Statistical procedures for determining the extent of cognitive change following concussion. *British Journal of Sports Medicine*, *38*, 273–278.

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., & McArdle, J.J. (2010). Longitudinal modeling of developmental changes in psychological research. *Current Directions in Psychological Science*, *19*, 149–154.

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.

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

Hinton-Bayre, A.D. (2010). Deriving reliable change statistics from test-retest normative data: Comparisons of models and mathematical expressions. *Archives of Clinical Neuropsychology*, *25*, 244–256.

Jacobson, N.S., & Truax, P. (1991). Clinical significance: A statistical approach to defining meaningful change in psychotherapy research. *Journal of Consulting and Clinical Psychology*, *59*, 12–19.

McCaffrey, R.J., & Westervelt, H.J. (1995). Issues associated with repeated neuropsychological assessments. *Neuropsychology Review*, *5*, 203–221.

McSweeny, A.J., Naugle, R.I., Chelune, G.J., & Luders, H. (1993). "T scores for change": An illustration of a regression approach to depicting change in clinical neuropsychology. *The Clinical Neuropsychologist*, *7*, 300–312.

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

Rushton, J.P., Brainerd, C.J., & Pressley, M. (1983). Behavioral development and construct validity: The principle of aggregation. *Psychological Bulletin*, *94*, 18–38.

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

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

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

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

Salthouse, T.A. (2010). Influence of age on practice effects in longitudinal neurocognitive change. *Neuropsychology*, *24*, 563–572.

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

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

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

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

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.

Tempkin, 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: The Psychological Corporation.

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

Woodcock, R.W., & Johnson, M.B. (1989). *Woodcock–Johnson psycho-educational battery—Revised.* Allen, TX: DLM.

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