Parallel but not equivalent: Challenges and solutions for repeated assessment of cognition over time

Published in final edited form as:
J Clin Exp Neuropsychol. 2012 ; 34(7): 758–772. doi:10.1080/13803395.2012.681628.

Parallel But Not Equivalent: Challenges and Solutions for Repeated Assessment of Cognition over Time

Alden L. Gross[2,3], **Sharon K. Inouye**[2,3], **George W. Rebok**[1,5], **Jason Brandt**[1,5], **Paul K. Crane**[6], **Jeanine M. Parisi**[1], **Doug Tommet**[2], **Karen Bandeen-Roche**[4], **Michelle C. Carlson**[1], **Richard N. Jones**[2,3], and **For the Alzheimer's Disease Neuroimaging Initiative***

[1]Department of Mental Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD

[2]Institute for Aging Research, Hebrew SeniorLife, Boston, MA

[3]Department of Medicine, Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA

[4]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD

[5]Department of Psychiatry and Behavioral Sciences, Johns Hopkins University School of Medicine

[6]Department of Internal Medicine, University of Washington

Abstract

Objective—Analyses of individual differences in change may be unintentionally biased when versions of a neuropsychological test used at different follow-ups are not of equivalent difficulty. This study's objective was to compare mean, linear, and equipercentile equating methods and demonstrate their utility in longitudinal research.

Study Design and Setting—The Advanced Cognitive Training for Independent and Vital Elderly (ACTIVE, N=1,401) study is a longitudinal randomized trial of cognitive training. The Alzheimer's Disease Neuroimaging Initiative (ADNI, n=819) is an observational cohort study. Nonequivalent alternate versions of the Auditory Verbal Learning Test (AVLT) were administered in both studies.

Results—Using visual displays, raw and mean-equated AVLT scores in both studies showed obvious nonlinear trajectories in reference groups that should show minimal change, poor equivalence over time (ps≤0.001), and raw scores demonstrated poor fits in models of within-person change (RMSEAs>0.12). Linear and equipercentile equating produced more similar means in reference groups (ps≥0.09) and performed better in growth models (RMSEAs<0.05).

Conclusion—Equipercentile equating is the preferred equating method because it accommodates tests more difficult than a reference test at different percentiles of performance and performs well in models of within-person trajectory. The method has broad applications in both clinical and research settings to enhance the ability to use nonequivalent test forms.

*Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.ucla.edu/wpcontent/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Corresponding author: Alden L. Gross, PhD; Phone: 617-971-5386; aldengross@hsl.harvard.edu.

Keywords

equating; equipercentile; neuropsychology; longitudinal analysis; alternate forms; parallel forms

Introduction

The identification of cognitive decline requires repeated assessment of the same individual over time. Alternate forms of a test are often used to minimize practice effects under the assumption that the forms are equivalent. Thus, the ability to equate alternate forms of neuropsychological tests is highly important in both clinical and research settings. Importantly, inferences about differences between groups, or change within persons over time, may be erroneous or biased when versions of a test are not equivalent.

Methods for test equating are particularly important for neuropsychological assessments in population-based research and clinical practice. Alternate versions of a test are considered equivalent if they produce the same mean and variance in a sample (Borsboom, 2005). Tests that are parallel but nonequivalent could be similar in content and length, but yield different scores for the same individual because they differ in item difficulty or administration characteristics. A relevant example is provided by word list-learning tests, for which use of alternate forms is common. Word lists in particular have an extensive history in Western psychology (Ebbinghaus, 1895/1964; Underwood, 1963). Different versions are intended to be equivalent, but often are not. One reason for non-equivalence is test length. For example, tests are often shortened due to time limitations or to reduce participant burden, as was done for the CVLT (Mitrushina et al., 2005), and equipercentile equating methods were used to equate a 9 item version with the original 16 item version. Other reasons for non-equivalence of forms are more complicated. In word list-learning tests, the word's frequency of use in the English language (Anderson & Bower, 1972; Battig & Montague, 1969; Fuller et al., 1997), number of syllables, serial position in the list, and imagery value (Paivio, 1968) all contribute to its difficulty. Fuller and colleagues (1997), for example, reported that two forms of the Auditory Verbal Learning Test (AVLT, described in Methods), Lists B and C (see Lezak et al., 2004), are not of equivalent difficulty. In a systematic review, Hawkins and colleagues (2004) showed recall differences of about three words (8% difference) across alternate forms that were studied.

Test equating is an analytical approach to adjust nonequivalent tests. The goal of test equating is to define a transformation of a variable that returns the same cumulative probability plot as the other variable being compared. There are many ways to equate tests. Given relevant characteristics of two distributions, equating methods can be applied in almost any setting in which multiple scores for each person are on different metrics. Two tests that measure the same outcome are equivalent if they place individuals in the same *relative position* in a group (Livingston, 2004). Widely used methods include mean, linear, and equipercentile equating. These methods differ by how relative position is defined. In mean equating, relative position is defined by the absolute difference from the sample mean of a test, and each individual's score is changed by the same amount to equate the sample mean to that of a reference test (Kolen & Brennan, 1995). In linear equating, relative position is defined in terms of standard deviations from the group mean. Linear equating is accomplished by adjusting scores from the new form to be within the same number of standard deviations of the mean of the original form. A formula is provided in the Methods section. Equipercentile equating defines relative position by a score's percentile rank in the group. It is accomplished by identifying scores on two measures with the same percentile rank and transforming the score on a new test to the corresponding score on the reference form with the same percentile rank.

The reason for administering alternate versions of a word list is to reduce retest or practice effects, but alternate forms do not completely correct for retest. Practice effects are attributable to general testing factors that arise from repeated exposure to the same task (e.g., learning to take tests) in addition to retention of particular test content (e.g., recall for specific words) (Crawford et al., 1989). Verbal learning tests are a mainstay in memory research and assessment, but are particularly susceptible to practice effects under repeated administration, such as in longitudinal research and in clinical settings when a patient must be re-evaluated over time. Alternate forms cannot account for general testing factors that contribute to practice effects, and may actually introduce greater problems in examining within-person trajectories in longitudinal studies if forms are not equivalent.

The objectives of the present study were to compare three equating methods – mean, linear, and equipercentile equating – and demonstrate their utility in longitudinal research. To demonstrate the generalizability of the approaches, examples are presented from two large studies of older adults, a cognitive intervention study and an observational study of predictors of conversion to Alzheimer's disease. Equating methods were contrasted using visual displays, tests of mean equivalence over time in reference groups, and with models of person-level growth.

Methods

Study samples

Participants were drawn from two large-scale, multi-site cohorts of older adults, the Advanced Cognitive Training for Independent and Vital Elderly (ACTIVE) study and the Alzheimer's Disease Neuroimaging Initiative (ADNI). These studies were selected because both provide longitudinal data using word list-learning measures collected using similar methods. Although the data are similar and nonequivalent forms pose similar challenges, the objectives of these studies differ considerably, which highlights the generalizability of equating approaches.

ACTIVE is a longitudinal randomized trial of cognitive training in cognitively intact, community-dwelling adults age 65 and older (Ball et al., 2002; Jobe et al., 2001; Willis et al., 2006). Participants (n=2,802) were randomized to one of four intervention groups (memory, reasoning, speed of cognitive processing, and no-contact control) after a baseline assessment, and followed up immediately after training and after one, two, three, five, and ten years. For the sake of parsimony, the present study used data from two of these groups, the memory-trained (n=703) and no-contact control (n=698) groups, which were collected at baseline before the intervention, immediately following training, and at the one, two, three, and five year follow-up assessments.

ADNI began in 2003 as a five-year observational cohort study of Alzheimer's disease (AD), with the primary goal of assessing the extent to which serial magnetic resonance imaging, positron emission tomography, other biological markers, and cognitive tests can be used to predict progression to mild cognitive impairment (MCI) and AD. Further information is available at http://www.loni.ucla.edu/ADNI. The present study used baseline and 6, 12, 18, 24, and 36 month follow-up data for normal subjects (n=229) and MCI (n=397) and AD (n=193) patients. MCI patients were assessed at all waves. Normal healthy controls were not followed at 18 months, and AD patients were not followed at the 18 or 36 month waves. Data from a 48 month wave were not included in the present study because data collection was still underway. Data, which are continuously updated, were downloaded for the present study on March 15, 2011.

Measures

The AVLT (Rey, 1964; Schmidt, 2004) was administered in both the ACTIVE and ADNI studies. During administration of the AVLT, participants are read a list of 15 unrelated words and asked to recall as many words as they can remember. The same list is repeated over five trials, followed by an interference trial with a new 15-word list, a short-delay free recall trial, and a long-delay free recall trial thirty minutes later. The present study used the sum of recall from the five immediate AVLT recall trials. The administration of the AVLT in ACTIVE differed from that in ADNI in two ways. First, the long-delay free recall trial was dropped due to time constraints. Second, the ACTIVE protocol modified the test for group administration by having participants write down responses instead of speaking them.

In ACTIVE, the original AVLT List A and interference list B (Taylor, 1959) were used at the baseline and third annual visits, lists by Geffen and colleagues (1994) were used in the immediate post-training and fifth annual visits, lists by Crawford and colleagues (1989) were used at the first annual visit, and lists described by Jones-Gotman and colleagues were used for the second annual visit (Lezak et al., 2004, pp. 423). In ADNI, lists from Taylor (1959) were used at the baseline, 12, and 24 month visits and lists by Crawford and colleagues (1989) were used during 6, 18, and 36 month visits. These visits are delineated in Tables and Figures with letters.

Statistical analyses

To account for differences in test difficulty, we conducted mean, linear, and equipercentile equating in ACTIVE and ADNI separately following similar procedures (Kolen & Brennan, 1995). We adapted weighted versions of each equating procedure to preserve aging, cohort, and group differences using a two-stage approach. In the first stage, we selected an equating sample from which to collect necessary characteristics of test distributions and derive the equating algorithm. The goal of this stage was to define a sample of participants at each follow-up visit whose memory ability was equivalent, such that any differences between visits could be attributed to form differences and not aging or group differences. In the second stage, we applied the equating algorithm to the full study sample in a way that preserved attrition, aging, cohort, and group differences but eliminated form differences. Equated scores were then compared visually using plots of mean recall over time and cumulative probability plots and statistically using tests of equivalence of means in reference groups as well as estimates of within-person change using latent growth models.

Stage 1: Defining the equating sample—An important assumption underlying any application of equating is that the populations producing responses on differently scaled tests at each time point must have the same underlying ability. To preserve differences in memory performance attributable to attrition, normal aging, and training status or diagnostic group, the equating sample was restricted in ACTIVE to control participants and in ADNI to MCI patients, respectively. Although the healthy control group in ADNI was the preferred reference group, the MCI group was used as the equating sample because by design ADNI assessed them at all waves and they provided a better coverage of AVLT scores observed across both healthy control and AD groups. Although the ADNI MCI group served as the equating sample, we subsequently restandardized the entire ADNI sample to make the healthy control group the reference sample by subtracting a model-implied mean difference in equated performance at each study visit between healthy controls and other participants. Equated performance at each time point was estimated from a weighted mixed effects model of test scores on indicators for time. Annotated code is available from the authors upon request.

To adjust for attrition over time, we used inverse propensity score weights in the equating sample to model the probability of dropout (Rosenbaum & Rubin, 1983). To preserve normal aging effects when the equating algorithm was applied to the full sample, we further restricted the equating sample to participant visits with ages that were common across all visits (ages 72 to 90 in ACTIVE and 63 to 90 in ADNI). Because ACTIVE and ADNI included long follow-up periods, we estimated analytic weights using a direct adjustment procedure for age to ensure the same age distribution at each study visit. This removes aging effects in the equating sample, but preserves aging in the full sample for individual differences analyses using equated test scores.

After procedures in this first stage, the only differences in memory performance between groups should be those attributable to form differences. These form differences are addressed to varying degrees with equating algorithms, key elements of which were carried over to be used on the full sample in the second stage.

Stage 2: Apply equating algorithms—Once equating samples were selected in ACTIVE and ADNI, equating algorithms were derived and applied to the full samples. These algorithms involve means at each study wave for mean equating (described earlier), means and standard deviations for linear equating, and test score percentiles for equipercentile equating. Linear equating was accomplished using the following formula:

$$Test2_{adjusted} = (SD_{Test1} / SD_{Test2}) * (Test2_i - \overline{\mathbf{Test2}}) + \overline{Test1}$$

Here, $Test2_{adjusted}$ is the linear-equated AVLT score for a follow-up Test2 visit for person i. The raw Test2 distribution had mean $\overline{\mathbf{Test2}}$ and standard deviation SD_{Test2}. The baseline test had mean $\overline{Test1}$ and standard deviation SD_{Test1}. The means and deviations were identified in the equating sample but, as for all equating methods, was applied to all participants in the second stage of equating.

The goal of equipercentile equating is to define a non-parametric transformation of one variable, in the present study a follow-up AVLT recall score, that returns the same cumulative probability plot as the baseline test. Test scores at follow-up visits were scaled to baseline scores with the same percentile rank. For example, a score of 47 on the baseline AVLT in ACTIVE had a percentile rank of 43.7%. The first annual AVLT score with that percentile ranking is 44, demonstrating that for this score range, the first annual test was more difficult than the baseline test by 3 words, assuming individuals did not truly decline after only 10 weeks. This example is over-simplified because the adaptation used for the present study does accommodate normal age-related decline. Additionally, our equipercentile equating algorithm used a loglinear function to smooth out equated score distributions (Albano, 2011).

Evaluation of equating methods with visual displays—Line graphs showing mean AVLT scores over time, or time trend plots, and cumulative probability plots were constructed to compare equated scores. Equivalent tests should yield identical cumulative probability plots in the reference sample. To account for data missing at random conditional on indicators for time and group, estimated means from random effects models were used in time trend plots. Cumulative probability plots show the cumulative proportion of the sample (y-axis) who recalled up to a given number of words on the AVLT (x-axis).

Evaluation of equating methods with tests of mean equivalence over time—The equivalence of test score means in reference groups (ACTIVE control, ADNI healthy controls) was tested for raw, mean, linear, and equipercentile equated scores using χ^2 tests for nested confirmatory factor analysis models. In the first of two models for each equating set, trial recall means at each study visit were constrained to be equal. In the second model,

means were freely estimated. Twice the difference in the log likelihood follows a chi-squared distribution. These tests are similar to repeated-measures ANOVA that tests for differences in means over time, but they are less stringent because they do not make the assumption that variances around the means are equal at each visit.

Evaluation of equating methods with models of within-person longitudinal trajectories—We used multiple group latent growth models to model person-level changes in recall over time (McArdle & Bell, 2000; Muthén, 1997; Muthén & Curran, 1997). Latent factors represent initial or baseline status and trajectories of change over time. These parameters are formed from observed scores at each study visit. We fixed factor loading paths from the intercept to observed recall sum scores for each assessment at 1, and factor loadings from the latent slope to values corresponding to a linear trajectory in time. In ACTIVE, a second intercept factor was also included to accommodate immediate training gains between baseline and post-training for trained participants (Bollen & Curran, 2006). We constrained its factor loadings to 0 at the baseline visit and to 1 at follow-up visits and its variance to 0.

Latent growth curve models and factor analyses were conducted using the Mplus (version 6.11) software package (Muthén & Muthén, 1998–2010). The models accommodate data missing at random conditional on observed covariates (Donders, van der Heijden, Stijnen, & Moons, 2006). Models using different equating methods were compared using standard model fit statistics including the root mean square error of approximation (RMSEA; Steiger, 1989) and comparative fit index (CFI; Hu & Bentler, 1999). These fit statistics were of key importance because they provide a measure of how much the model-estimated baseline levels and trajectories fit to observed trajectories using each method of equating. An RMSEA less than 0.05 and CFI greater than 0.95 are considered indicators of excellent model fit (Hu & Bentler, 1999). Graphical displays and equating algorithms were generated using Stata 12.0 (StataCorp, 2011) and R software packages (R Development Core Team, 2009).

Results

Table 1 shows baseline characteristics and AVLT test scores at each follow-up time for ACTIVE and ADNI samples. ACTIVE participants were mostly white females aged 65–94 and cognitively intact at baseline. On average, ADNI participants were younger, more highly educated, and a higher proportion of them were males compared to the ACTIVE sample.

Evaluation of equating methods with visual displays

Mean recall over time for ACTIVE control and memory-trained participants under different equating methods are plotted in each panel of Figure 1. Figure 2 provides similar information using the ADNI MCI diagnostic group. In ACTIVE, plots of raw scores give the impression that both groups start at about the same level at the baseline visit, decline up to two years after training, and then recover inexplicably (Figure 1). In ADNI, the MCI group zigzags in performance at every other visit by approximately 0.3 standard deviations. AVLT test scores in Table 1 show that raw score trends in Figure 1 generalize to all intervention and diagnostic groups. The effect of nonequivalent forms is demonstrated more rigorously using cumulative probability plots in the Appendix (see Legends for a detailed interpretation). They reveal different difficulty levels across the score distribution for different waves: participants in both ACTIVE and ADNI performing at the 50th percentile on follow-up tests had systematically lower scores than participants at the 50th percentile of

the baseline test. Scores at the lowest and highest performance percentiles were more comparable across visits.

Mean equating in ACTIVE produced more similar means, but overcompensated for differences in forms at the third and fifth annual visits (Figure 1). Mean equating in ADNI produced a plot of a declining trajectory, although some residual form differences remained at the 24 month visit (Figure 2). Cumulative probability plots for mean equating did not overlap as well as for other equating methods (Appendix).

Linear equating, like mean equating, revealed boosted performance at one year followed by decline in ACTIVE but also suggests improvement at the immediate post-training visit and a plateau in performance after the third annual visit (Figure 1). Linear equating in ADNI nearly eliminated any indication of decline in the mean level of performance over time, which should be expected in a sample of MCI patients.

Equipercentile equating produced the smoothest trajectories in ACTIVE, with an expected pre-post training gain and age-related cognitive decline (Figure 1). In ADNI, the wave pattern over time was successfully removed while an average decline of 1.8 AVLT words through three years was still apparent (Figure 2).

Although the graphical displays demonstrate the superiority of equating methods in relation to raw scores, they also indicate residual imprecision of these methods: in ACTIVE, the second visit was a post-test assessment only 10 weeks after the first. The equating methods should theoretically produce equal means between the baseline and immediate post-test assessments because there was almost no attrition or aging in the control group. Mean equating does not because of small residual form differences that persist after equating. Linear and equipercentile equating produce means at baseline and immediate post-training in ACTIVE very close to each other (Figure 1).

Cumulative probability plots are shown in the Appendix. Plots for equated scores demonstrate excellent overlap, especially for equipercentile equating in ACTIVE (Figure A1) and linear and equipercentile equating in ADNI (Figure A2). Thus, equipercentile equating visually demonstrates the best adjustment for learning effects and smoothes out mean trajectories.

Evaluation of equating methods with tests of mean equivalence over time

Negligible change over time was assumed in the ADNI healthy control and ACTIVE control groups. In ADNI, raw (χ^2=14.8, df=4, p=0.001) and mean equating (χ^2=16.7, df=4, p<0.001) produced significantly different means over time, but linear (χ^2=6.8, df=5, p=0.18) and equipercentile equating (χ^2=7.6, df=4, p=0.09) produced statistically equivalent means over time in the healthy control group. In the ACTIVE control group, there were significant differences in raw AVLT recall sum score means (χ^2=45.3, df=5, p<0.001), mean-equated means (χ^2=58.2, df=5, p<0.001), linear-equated means (χ^2=63.6, df=5, p<0.001), and equipercentile-equated means (χ^2=31.9, df=5, p<0.001).

By the fifth year visit, n=749 of 1,401 (53%) participants were still in the study sample and attrition did not differ by intervention group. Follow-up in ADNI after three years was higher (n=591/819, 72%). Our propensity adjustment for sample attrition, which had a larger effect in the ACTIVE study, is likely responsible for the lack of equivalence of equipercentile-equated means, which as shown in Figure 1 demonstrates a smooth declining trajectory.

Evaluation of equating methods with models of within-person longitudinal trajectories

Results from ACTIVE and ADNI are shown in Tables 2 and 3, respectively, for raw, mean, linear, and equipercentile-equated scores. Means and variances of growth parameters in the Tables characterize level and variability in within-person trajectory of AVLT performance over time. An immediate contrast is in poor RMSEA and CFI model fits using raw AVLT scores, shown in Tables 2 and 3, that improve dramatically given any equating method but are best after equipercentile equating. The model with equipercentile-equated AVLT scores in ADNI fit perfectly with the data (RMSEA: 0.0; CFI=1.0).

Besides fit statistics, three key substantive inferences change depending on the type of equating used. First, using raw scores in ACTIVE, the immediate training "boost" appears to be in the negative direction for memory-trained participants (pre-post change: −5.0 words) but flips to a positive direction after equating (Table 2). Second, in both ACTIVE and ADNI, annual memory decline in all groups, as indicated by slope means, is overestimated or underestimated to varying degrees using raw scores, while equated scores show less annual decline in ACTIVE control participants (Table 2) and ADNI MCI and AD participants (Table 3). A third substantive change in inferences is the correlation between initial recall and aging trajectory is overestimated using raw scores relative to equated scores. Based on model fits and substantive knowledge of trajectories of cognitive aging, latent growth models suggest differences in test difficulty are handled best with either linear or equipercentile equating.

Discussion

The present study investigated different methods of equating AVLT word list versions in longitudinal aging research. We adapted accepted test equating methods using a novel approach to the study of longitudinal cognitive aging. These methods are broadly applicable to within- and between-group comparisons of test performance data in both research and clinical settings. Equipercentile equating uses observed percentiles of a distribution, and is a more generalizeable non-parametric transformation than linear equating, which assumes normally distributed variables whose distributions are fully characterized by a mean and standard deviation. Graphical displays clearly show equipercentile equating accommodates tests that are more difficult than the reference test at different percentiles of performance, and models of within-person change show it also satisfactorily adjusts for practice, or retest, effects. Importantly, an implicit assumption of mean, linear, or equipercentile equating is that the populations producing two sets of scores, whether they are the same people followed over time or two different groups, have the same underlying ability. Because this may not be a valid assumption for older adults followed for years, the present study described equating procedures that used age standardization to preserve aging effects, propensity weighting to adjust for attrition, and restriction to preserve group differences due to diagnostic and intervention group membership.

ACTIVE is the largest study of cognitive training among older adults to date, and ADNI is a $60 million public-private partnership that is being used to stimulate innovative methods for evaluating progression of AD in clinical trials. The roller coaster trajectory in ACTIVE and waves in ADNI are attributable to nonequivalent AVLT forms used at different study visits. The ACTIVE study cycled through four versions of the AVLT until repeating the baseline list at the third annual follow-up. ADNI cycled between two AVLT lists, which explains the wave-like pattern. These method artifacts may be present in other settings. Indeed, important form differences are seen for the Hopkins Verbal Learning Test (Brandt & Benedict, 2001) in the ACTIVE study (data not shown) and for the ADAS-Cog word list-learning task in ADNI (Crane et al., under review). Similar plots and statistics presented in this study can be replicated using these measures. The reason these studies used alternate word lists was to

reduce practice effects, but in doing so they introduced complications for making inferences about cognitive performance. All ACTIVE publications involving comparisons of within-person memory performance over time use equipercentile-equated scores (e.g., Gross et al., 2010a, 2010b, 2011; Parisi et al., 2011). Aside from work in ACTIVE, we are not aware of equipercentile equating being used in longitudinal settings with cognitive performance data. We believe the field can benefit by being aware of and adopting these equating methods. To date, most published studies that have used longitudinal neuropsychological data from ADNI have not examined the AVLT from visits in which different AVLT forms were administered (e.g., Hinrichs et al., 2011, Murphy et al., 2010, Petersen et al., 2010). In other studies using ADNI data, word lists are treated as components in composite measures (e.g., Beckett et al., 2010), but results of some studies are potentially susceptible to nonequivalent form differences (e.g., Carmichael et al., 2010; Okonkwo et al., 2011). Future work in ADNI should pay close attention to form differences on the AVLT and ADAS-Cog.

Equating methods are powerful tools, but their use comes with several caveats. First, measures should not be equated that have different meanings. For example, it is statistically possible to equate short-delay and long-delay recall trials, but the trials measure qualitatively different constructs. Relatedly, equating methods can equate test scores but do not address qualitative differences in behaviors, such as different strategies used on more difficult tests at different measurement occasions (Crawford et al., 1989; Light, 1991). A second limitation of equating is that populations that produce two sets of test scores must have the same underlying ability to be validly equated. This is an easy assumption to make when the same cognitively normal persons are being retested over time, but may not be achievable (or measurable) in all situations. The application of equating methods in the present study would have been fairly straightforward if we had assumed this. However, in studies with several years of longitudinal follow-up such as those in the present study, one can divide the equating task into two stages as we have done: identify a subset of observations as an equating sample in which underlying abilities can be assumed to be the same over time, then apply the equating algorithm derived in that sample to the full sample. A third limitation is that, in longitudinal settings, equating procedures assume the magnitude of retest effects is exchangeable across groups. This assumption may be unreasonable when comparing patients with different clinical syndromes or diseases, such as delirium or amnesia. Fourth, a limitation specific to equipercentile equating is that the outcome should be continuously distributed and have enough range to reliably distinguish different quantiles. Applying equipercentile equating to individual AVLT trial recall scores, for example, would be more challenging. This is not a concern in linear equating, which presumes a normally distributed outcome. Another limitation of this study is that we assumed that the underlying trajectory of change in AVLT performance is in fact linear over time. We used this assumption in growth models to assess the different equating methods. Previous work in ACTIVE has demonstrated memory follows a linear pace of change following the immediate post-training visit (Gross & Rebok, 2011; Parisi et al., 2011). The assumption of linear change in cognitive function among older adults is a commonly accepted fact in many other studies of older adults (e.g., Proust et al., 2006). Nevertheless, because true change is a latent and unobserved phenomenon, whether the AVLT in ACTIVE and ADNI in fact shows linear decline over time is uncertain. A final potential limitation specific to the ACTIVE study is that modifications in test administration of the AVLT from standard clinical administration limit the generalizability of findings from these data to clinical settings. However, our purpose in the present study was to illustrate equating methods and not to make inferences about training effects on memory function in ACTIVE, which have been reported elsewhere (e.g., Gross & Rebok, 2011; Parisi et al., 2011; Willis et al., 2006).

Mean, linear, and equipercentile equating, based in classical test theory, are not the only equating methods. Item response theory (IRT) methods can be used if populations producing

two sets of scores differ in the underlying ability being measured, but require some items in common between the tests to anchor the two groups with respect to each other (Livingston, 2004). Counterbalancing is a method of adjusting for form differences in the study design before analysis, but is useful only for making inferences about group differences and not within-person change (Cozby, 2009).

Although equipercentile equating proved to be ideal for the applications of the present study, the same procedure may not apply in all cases. Mean equating is intuitive and produces the same grand mean on two tests, but it does not change an individual's absolute difference from the mean. Thus, mean equating can lead to impossible or improbable scores among some individuals; for example, if two tests means are 60 and 50, and the maximum possible value is 100, then an individual scoring a 95 on the second test will have a mean-equated score of 105. Similar to mean equating, a limitation of linear equating is that extreme scores on a new test may yield equated scores outside the possible range of values in the original test. This is not a concern in equipercentile equating. The principal advantage of equipercentile equating over linear equating is that it does not assume the reference test is normally distributed, but there are cases in which that assumption is viable. The AVLT in ACTIVE and ADNI was approximately normally distributed, which explains the similarities in findings between linear and equipercentile equating.

Clinically, a patient's test scores can only be interpreted using appropriate reference norms, but normative values are unhelpful if normative test scores come from a different population from which the patient came. Tests shown to be equivalent in certain groups defined by education, sex, or age may not be equivalent in other subpopulations (Ivnik et al., 1990). For this reason, Schmidt (2004) reports AVLT word lists that produce similar scores for older adults in addition to which lists produce similar scores to other lists. Equating techniques require data from cohorts of individuals to carry out, so it would not be possible to perform similar analyses for any particular person being evaluated clinically. Nevertheless, important differences in form difficulty should be kept in mind, and if different forms are used across time, this should be documented. Data from studies similar to the one presented here may be useful to assist the practitioner in understanding whether change has occurred, and if so, its likely direction and magnitude. Ignoring differences in difficulty across forms in clinical settings could lead to unnecessary confusion at least and incorrect conclusions or diagnoses at worst. Finally, it is important to acknowledge that equated data contribute to only a small part of the clinical picture. A clinician's judgment of change will depend on multiple test results and findings, the clinical history, non-quantitative observations about the patient's abilities (Lezak et al., 2004), and on his or her expert judgment and prior experience (Mitrushina et al., 2005).

In conclusion, equating challenges are pervasive but often unrecognized in research studies and clinical practice. When prior knowledge about form equivalence is unavailable or unclear when planning a study, we recommend that researchers use the same form and apply established methods to control for practice effects (e.g., Ferrer et al., 2004, 2005; Rabbitt et al., 2004; Salthouse et al., 2004, 2008, 2010). Thorough data exploration is necessary both to recognize the need for equating and to understand the relative merits of different equating procedures. The replication of findings across two cohorts, utilizing special weighting adaptations, highlights the versatility and generalizability of the equating methods used in the present study.

The method of equipercentile equating may have broad applications in both clinical and research settings to enhance the ability to use nonequivalent test forms, to evaluate change over time, to quantify retest effects, and to align scores on different tests of the same construct (such as identifying cutpoints for dementia on cognitive screening tests).

Equipercentile equating is a well-accepted tool for comparing psychiatric diagnostic instruments (Furukawa et al., 2009; Leucht et al., 2005; Montoya et al., 2011; Noonan et al., 2011; Schennach-Wolff et al., 2010) and for identifying clinically relevant benchmarks and crosswalks on neuropsychological tests (Fong et al., 2009, 2011). The procedure represents a robust and innovative approach to better understanding longitudinal changes over time. The present study demonstrated an innovative application of equating methods for longitudinal settings in which participants or patients are followed over long periods of time.

Acknowledgments

Dr. Gross was supported by a National Institutes of Health Translational Research in Aging fellowship (T32AG023480-07). Dr. Inouye holds the Milton and Shirley F. Levy Family Chair in Alzheimer's Disease. This work was supported in part by Grant No. P01AG031720 (SKI) from the National Institute of Aging. Dr. Rebok is an investigator with Compact Disc Incorporated for the development of an electronic version of the ACTIVE memory intervention. He has received no financial support from them for ACTIVE. Dr. Brandt receives royalty income from Psychological Assessment Resources, Inc., on sales of the Hopkins Verbal Learning Test-Revised. Drs. Rebok's and Brandt's relationships are managed by the Johns Hopkins University according to its established conflict of interest policies.

The ACTIVE intervention trials are supported by grants from the National Institute on Aging and the National Institute of Nursing Research to Hebrew Senior Life (U01NR04507), Indiana University School of Medicine (U01NR04508), Johns Hopkins University (U01AG14260), New England Research Institutes (U01AG14282), Pennsylvania State University (U01AG14263), the University of Alabama at Birmingham (U01 AG14289), and the University of Florida (U01AG14276).

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott; Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Amorfix Life Sciences Ltd.; AstraZeneca; Bayer HealthCare; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles. This research was also supported by NIH grants P30 AG010129, K01 AG030514, and the Dana Foundation.

References

Albano, A. equate: Statistical methods for test score equating. R package version 1.1-2. 2011. Retrieved December 19, 2011 from http://CRAN.R-project.org/package=equate

Anderson JR, Bower GH. Recognition and retrieval processes in free recall. Psychological Review. 1972; 79:97–123.

Ball K, Berch DB, Helmers KF, Jobe JB, Leveck MD, Marsiske M, et al. Effects of cognitive training interventions with older adults: A randomized controlled trial. Journal of the American Medical Association. 2002; 288:2271–2281. [PubMed: 12425704]

Battig WF, Montague WE. Category norms for verbal items in 56 categories: A replication and extension of the Connecticut category norms. Journal of Experimental Psychology, Monograph. 1969; 80:1–46.

Beckett LA, Harvey DJ, Gamst A, Donohue M, Kornak J, Zhang H, Kuo JH. Alzheimer's Disease Neuroimaging Initiative. The Alzheimer's Disease Neuroimaging Initiative: Annual change in biomarkers and clinical outcomes. Alzheimer's & Dementia: The Journal of the Alzheimer's Association. 2010; 6:257–264.

Bollen, KA.; Curran, PJ. Latent curve models: A structural equation approach. Hoboken, NJ: Wiley; 2006.

Brandt, J.; Benedict, RHB. Hopkins Verbal Learning Test–Revised: Professional manual. Odessa, FL: Psychological Assessment Resources; 2001.

Buschke H. Selective reminding for the analysis of memory and learning. Journal of Verbal Learning and Verbal Behavior. 1973; 12:543–550.

Borsboom, D. Measuring the Mind: Conceptual Issues in Contemporary Psychometrics. Cambridge, UK: Cambridge University Press; 2005.

Carmichael O, Schwarz C, Drucker D, Fletcher E, Harvey D, Beckett L, Jack CR Jr, Weiner M, DeCarli C. Alzheimer's Disease Neuroimaging Initiative. Longitudinal Changes in White Matter Disease and Cognition in the first year of ADNI. Archives of Neurology. 2010; 67:1370–1378. [PubMed: 21060014]

Cozby, PC. Methods in Behavioral Research: Tenth Edition. New York, NY: McGraw-Hill; 2009.

Crane PK, Carle A, Gibbons LE, Mungas D. for the Alzheimer's Disease Neuroimaging Initiative*. Development and assessment of a psychometrically sophisticated composite score for memory in the Alzheimer's Disease Neuroimaging Initiative (ADNI). (under review).

Crawford JR, Stewart LE, Moore JW. Demonstration of savings on the AVLT and development of a parallel form. Journal of Clinical and Experimental Neuropsychology. 1989; 11:975–981. [PubMed: 2592534]

Donders AR, van der Heijden GJ, Stijnen T, Moons KG. Review: A gentle introduction to imputation of missing values. Journal of Clinical Epidemiology. 2006; 59:1087–1091. [PubMed: 16980149]

Ebbinghaus, H. Memory: A contribution to experimental psychology. New York: Dover; 1895–1964.

Ferrer E, Salthouse T, McArdle J, Stewart W, Schwartz B. Multivariate modeling of age and retest in longitudinal studies of cognitive abilities. Psychology & Aging. 2005; 20(3):412–422. [PubMed: 16248701]

Ferrer E, Salthouse TA, Stewart WF, Schwartz BS. Modeling age and retest processes in longitudinal studies of cognitive abilities. Psychology & Aging. 2004; 19(2):243–259. [PubMed: 15222818]

Fong TG, Fearing MA, Jones RN, Shi P, Marcantonio ER, Rudolph JL, Yang FM, Kiely DK, Inouye SK. Telephone interview for cognitive status: Creating a crosswalk with the Mini-Mental State Examination. Alzheimers & Dementia. 2009; 5(6):492–497.

Fong TG, Jones RN, Rudolph JL, Yang FM, Tommet D, Habtemariam D, Marcantonio ER, Langa KM, Inouye SK. Development and validation of a brief cognitive assessment tool: the sweet 16. Archives of Internal Medicine. 2011; 171(5):432–437. [PubMed: 21059967]

Fuller KH, Gouvier WD, Savage RM. Comparison of List B and List C of the Rey Auditory Verbal Learning Test. The Clinical Neuropsychologist. 1997; 11:201–204.

Furukawa TA, Shear MK, Barlow DH, Gorman JM, Woods SW, Money R, Etschel E, Engel RR, Leucht S. Evidence-based guidelines for interpretation of the Panic Disorder Severity Scale. Depression and Anxiety. 2009; 26(10):922–929. [PubMed: 19006198]

Geffen GM, Butterworth P, Geffen LB. Test-retest reliability of a new form of the Auditory Verbal Learning Test (AVLT). Archives of Clinical Neuropsychology. 1994; 9:303–316. [PubMed: 14589623]

Gross AL, Rebok GW, Unverzagt FW, Willis SL, Brandt J. Word list memory predicts everyday function and problem–solving in the elderly: Results from the ACTIVE cognitive intervention trial. Aging, Neuropsychology, and Cognition. 2010; 18:129–146.

Gross AL, Rebok GW, Unverzagt FW, Willis SL, Brandt J. Cognitive predictors of everyday functioning in the elderly: Results from the ACTIVE cognitive intervention trial. Journal of Gerontology: Psychological Sciences. 2011; 66:557–566.

Gross AL, Rebok GW. Memory training and strategy use among older adults: Results from the ACTIVE cognitive intervention trial. Psychology & Aging. 2011; 26:503–517. [PubMed: 21443356]

Hawkins KA, Dean D, Pearlson GD. Alternative forms of the Rey Auditory Verbal Learning Test: A review. Behavioural Neurology. 2004; 15:99–107. [PubMed: 15706053]

Hinrichs C, Singh V, Xu G, Johnson SC. Alzheimers Disease Neuroimaging Initiative. Predictive markers for AD in a multi-modality framework: An analysis of MCI progression in the ADNI population. Neuroimage. 2011; 55:574–589. [PubMed: 21146621]

Hu L, Bentler PM. Cutoff criteria for fit indices in covariance structure analysis: Conventional versus new alternatives. Structural Equation Modeling. 1999; 6:1–55.

Ivnik RJ, Malec JF, Tangalos EG, Petersen RC, Kokmen E, Kurland LT. The Auditory Verbal Learning Test (AVLT): Norms for ages 55 and older. Psychological Assessment: A Journal of Consulting and Clinical Psychology. 1990; 2(3):304–312.

Jobe JB, Smith DM, Ball K, Tennstedt SL, Marsiske M, Willis SL, et al. ACTIVE: A cognitive intervention trial to promote independence in older adults. Controlled Clinical Trials. 2001; 22:453–479. [PubMed: 11514044]

Kolen, M.; Brennan, R. Test equating: Methods and practices. New York: Springer; 1995.

Leucht S, Kane JM, Kissling W, Hamann J, Etschel E, Engel RR. What does the PANSS mean? Schizophrenia Research. 2005; 79(2–3):231–238. [PubMed: 15982856]

Lezak, MD.; Howieson, DB.; Loring, DW. Neuropsychological Assessment. 2nd edition. New York: Oxford University Press; 2004.

Light L. Memory and aging: Four hypotheses in search of data. Annual Review of Psychology. 1991; 42:333–376.

Livingston, SA. Equating test scores (without IRT). Educational Testing Service; 2004.

McArdle, JJ.; Bell, RQ. Recent trends in modeling longitudinal data by latent growth curve methods. In: Little, TD.; Schnabel, KU.; Baumert, J., editors. Modeling longitudinal and multiple–group data: Practical issues, applied approaches, and scientific examples. Mahwah, NJ: Lawrence Erlbaum; 2000. p. 69-108.

Mitrushina, M.; Boone, KB.; Razani, J.; D'Elia, LF. Handbook of Normative Data for Neuropsychological Assessment. 2nd edition. NY: Oxford University Press; 2005.

Montoya A, Valladares A, Lizán L, San L, Escobar R, Paz S. Validation of the Excited Component of the Positive and Negative Syndrome Scale (PANSS-EC) in a naturalistic sample of 278 patients with acute psychosis and agitation in a psychiatric emergency room. Health and Quality of Life Outcomes. 2011; 9:18. [PubMed: 21447155]

Murphy EA, Holland D, Donohue M, McEvoy LK, Hagler DJ Jr, Dale AM, Brewer JB. Alzheimer's Disease Neuroimaging Initiative. Six-month atrophy in MTL structures is associated with subsequent memory decline in elderly controls. Neuroimage. 2010; 53:1310–1317. [PubMed: 20633660]

Muthén, BO. Latent variable modeling with longitudinal and multilevel data. In: Raftery, A., editor. Sociological methodology. Boston: Blackwell Publishers; 1997. p. 453-480.

Muthén BO, Curran PJ. General LongitudinalModeling of Individual Differences in Experimental Designs: A Latent Variable Framework for Analysis and Power Estimation. Psychological Methods. 1997; 2:371–402.

Muthén, LK.; Muthén, BO. Mplus user's guide: Sixth Edition. Los Angeles, CA: Muthén & Muthén; 1998–2010.

Noonan VK, Cook KF, Bamer AM, Choi SW, Kim J, Amtmann D. Measuring fatigue in persons with multiple sclerosis: creating a crosswalk between the Modified Fatigue Impact Scale and the PROMIS Fatigue Short Form. Quality of Life Research. 2011 [Epub ahead of print].

Okonkwo OC, Mielke MM, Griffith HR, Moghekar AR, O'Brien RJ, Shaw LM, Trojanowski JQ, Albert MS. Alzheimer's Disease Neuroimaging Initiative. Cerebrospinal Fluid Profiles and Prospective Course and Outcome in Patients With Amnestic Mild Cognitive Impairment. Archives of Neurology. 2011; 68:113–119. [PubMed: 21220682]

Paivio A. A factor-analytic study of word attributes and verbal learning. Journal of Verbal Learning and Verbal Behavior. 1968; 7:41–49.

Parisi JM, Gross AL, Rebok GW, Saczynski JS, Crowe M, Cook SE, Langbaum JB, Sartori A, Unverzagt FW. Modeling change in memory performance and perceptions: Findings from the ACTIVE study. Psychology & Aging. 2011; 26:518–524. [PubMed: 21463064]

Petersen RC, Aisen PS, Beckett LA, Donohue MC, Gamst AC, Harvey DJ, Jack CR Jr, Jagust WJ, Shaw LM, Toga AW, Trojanowski JQ, Weiner MW. Alzheimer's Disease Neuroimaging Initiative (ADNI): clinical characterization. Neurology. 2010; 74:201–209. [PubMed: 20042704]

Proust C, Jacqmin-Gadda H, Taylor JM, Ganiayre J, Commenges D. A nonlinear model with latent process for cognitive evolution using multivariate longitudinal data. Biometrics. 2006; 62:1014–1024. [PubMed: 17156275]

R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Austria: Vienna; 2009. ISBN 3-900051-07-0, URL: http://www.R-project.org.

Rabbitt P, Diggle P, Holland F, McInnes L. Practice and drop-out effects during a 17-year longitudinal study of cognitive aging. Journal of Gerontology, Series B: Psychological & Social Sciences. 2004; 59(2):84–97.

Rey, A. L'examen clinique en psychologie. Paris, France: Presses Universitaires de France; 1964.

Rosenbaum PR, Rubin D. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983; 70(1):41–55.

Salthouse TA. Influence of age on practice effects in longitudinal neurocognitive change. Neuropsychology. 2010; 24:563–572. [PubMed: 20804244]

Salthouse T, Schroeder D, Ferrer E. Estimating retest effects in longitudinal assessments of cognitive functioning in adults between 18 and 60 years of age. Developmental Psychology. 2004; 40(5): 813–8229. [PubMed: 15355168]

Salthouse TA, Tucker-Drob EM. Implications of short-term retest effects for the interpretation of longitudinal change. Neuropsychology. 2008; 22:800–811. [PubMed: 18999354]

Schennach-Wolff R, Obermeier M, Seemüller F, Jäger M, Schmauss M, Laux G, et al. Does clinical judgment of baseline severity and changes in psychopathology depend on the patient population? Results of a CGI and PANSS linking analysis in a naturalistic study. Journal of Clinical Psychopharmacology. 2010; 30:726–731. [PubMed: 21105273]

Schmidt, M. Rey Auditory and Verbal Learning Test: A handbook. Los Angeles, CA: Western Psychological Services; 2004.

StataCorp. Stata Statistical Software: Release 12. College Station, TX: StataCorp LP; 2011.

Steiger, JH. EZPATH: A supplementary module for SYSTAT and SYGRAPH. Evanston, IL: Systat; 1989.

Taylor, EM. The appraisal of children with cerebral deficits. Cambridge, MA: Harvard University Press; 1959.

Underwood BJ. Coding processes in verbal learning. Journal of Verbal Learning and Verbal Behavior. 1963; 1:250–257.

Willis SL, Tennstedt SL, Marsiske M, Ball K, Elias J, Koepke KM, et al. Long–term effects of cognitive training on everyday functional outcomes in older adults. Journal of the American Medical Association. 2006; 296:2805–2814. [PubMed: 17179457]

Appendix

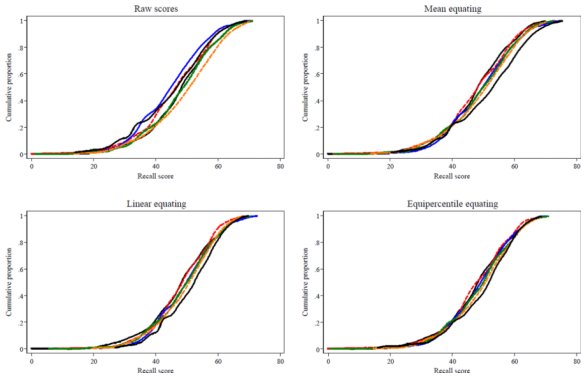

Figure A1.
Parallel but Nonequivalent Forms: Cumulative Probability Plots of Raw and Equated AVLT Scores in ACTIVE

Legend. Cumulative probability plots overlay distributions of raw or equated AVLT recall sum scores among control participants from each ACTIVE study visit (n=698). Plots for each visit are not labeled clearly because the purpose of this diagnostic plot is to assess degree of overlap of each visit's cumulative distribution; visits are plotted in the following colors (baseline: gray; immediate post-training: black; first annual: red; second annual: blue; third annual: green; fifth annual: orange). Results suggest linear and equipercentile equating produces more overlap than other methods. As an example of how to interpret this plot, in the raw scores panel, the blue line shows that about 30% of ACTIVE control participants recalled up to 39 words in year 1, and 100% of participants at all waves recalled 75 words or less (the test's ceiling).

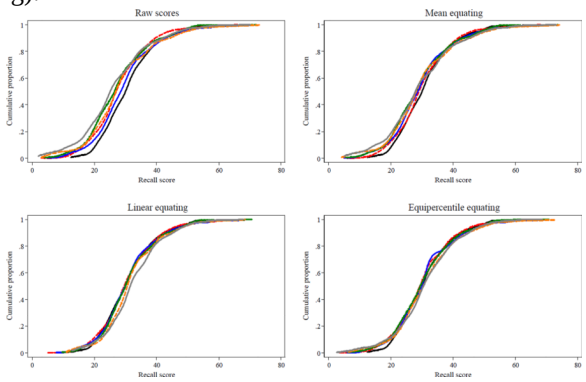

Figure A2.

Parallel but Nonequivalent Forms: Cumulative Probability Plots of Raw and Equated AVLT Scores in ADNI

Legend. Cumulative probability plots overlay distributions of raw or equated AVLT recall sum scores among MCI patients from each ADNI study visit (n=397). Plots for each visit are not labeled labeled clearly because the purpose of this diagnostic plot is to assess degree of overlap of each visit's cumulative distribution; visits are plotted in the following colors (baseline: black; 6 month: red; 12 month: blue; 18 month: green; 24 month: orange; 30 month: yellow; 36 month: gray). Results suggest equipercentile equating produces more overlap than other methods. As an example of how to interpret this plot, in the raw scores panel, the gray line shows that about 10% of ADNI MCI participants recalled up to 12 words at 36 months, and 100% of participants at all waves recalled 75 words or less (the test's ceiling).

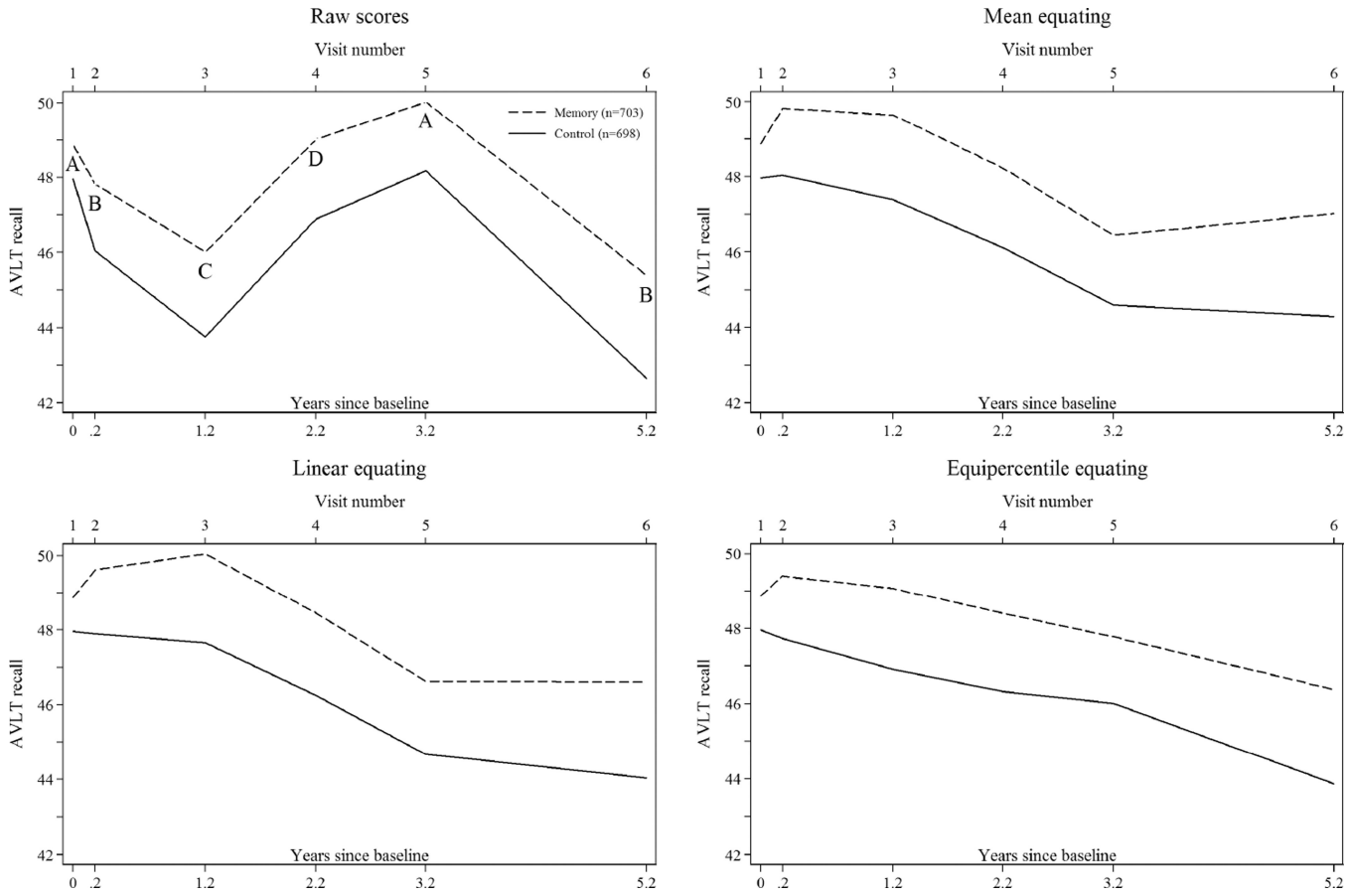

Figure 1.

Parallel but Nonequivalent Forms: Plots of Raw and Equated AVLT Scores Over Time in ACTIVE (N=1,401)

Legend. Time trend plots present means of AVLT scores by study visit in the ACTIVE control and memory-trained groups. Means in the time trend plots are adjusted for selective attrition using random effects models that assume data are missing at random conditional on indicators for time and group. Letters correspond to AVLT list versions administered at a visit: in ACTIVE, the baseline and year 3, and post-training and year 5, study visits used the same AVLT form.

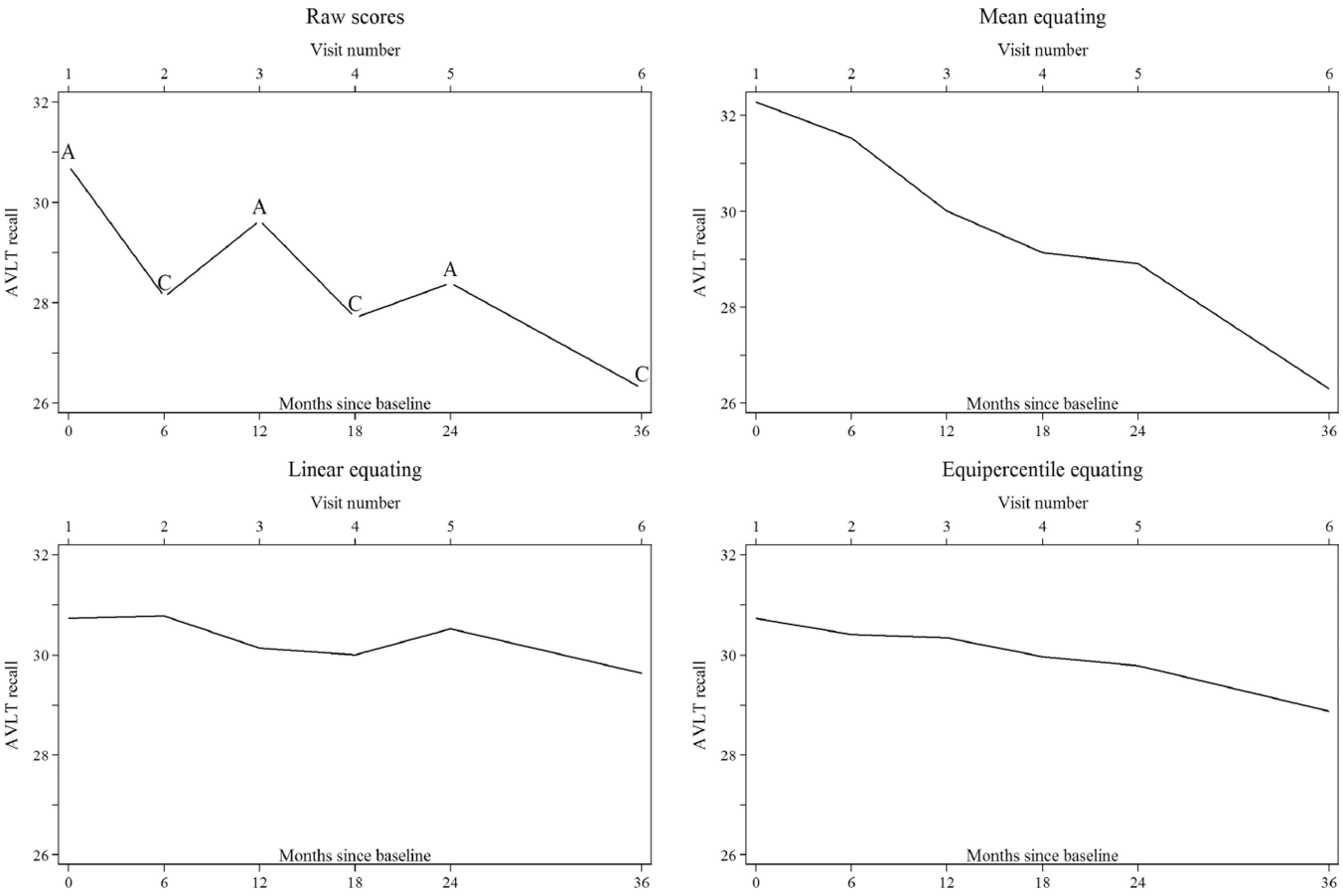

Figure 2.

Parallel but Nonequivalent Forms: Plots of Raw and Equated AVLT Scores Over Time in ADNI MCI Participants (N=397)

Legend. Time trend plots present means of AVLT scores by study visit in the ADNI MCI group. Means in the time trend plots are adjusted for selective attrition using random effects models that assume data are missing at random conditional on indicators for time and group. Letters correspond to AVLT list versions administered at a visit: in ADNI, the baseline, 12 month, and 24 month visits used the same form and the 6 month, 18 month, and 36 month visits used a different form.

Table 1

Descriptive characteristics of cohorts used in the present study

	ACTIVE		ADNI		
	Memory-trained	**Control**	**Healthy control**	**Mild Cognitive Impairment**	**Alzheimer's disease**
	n=703	**n=698**	**n=229**	**n=397**	**n=193**
Age, *mean (SD)*	73.5 (6.0)	74.1 (6.1)	67.8 (11.8)	67.6 (11.7)	67.9 (12.6)
Years of Education, *mean (SD)*	13.6 (2.7)	13.4 (2.7)	16.0 (2.9)	15.7 (3.0)	14.7 (3.1)
Sex, *n (% female)*	537 (76.0)	514 (74.1)	110 (48.0)	141 (35.5)	91 (47.2)
Ethnicity, *n (% white)*	521 (74.3)	500 (72.2)	210 (91.7)	371 (93.5)	181 (93.8)
MMSE score, *mean (SD)*	27.3 (2.1)	27.3 (2.0)	29.1 (1.0)	27.0 (1.8)	23.3 (2.1)
Auditory Verbal Learning Test (AVLT) scores					
First follow-up (Baseline)	48.8 (10.6)	47.9 (11.0)	43.2 (9.3)	30.7 (9.0)	23.0 (7.8)
Second follow-up	48.1 (10.9)	46.1 (11.5)	41.3 (10.2)	28.1 (9.3)	20.2 (7.9)
Third follow-up	47.1 (10.7)	44.6 (10.8)	43.8 (10.4)	29.6 (10.3)	20.0 (8.1)
Fourth follow-up	50.3 (10.3)	47.9 (11.2)	--	27.7 (10.1)	--
Fifth follow-up	51.5 (11.0)	49.4 (11.5)	44.6 (10.6)	28.4 (11.6)	16.8 (8.9)
Sixth follow-up	47.6 (11.4)	45.2 (12.0)	40.4 (10.3)	26.3 (11.6)	--

Legend. The first follow-up in ACTIVE and ADNI was the baseline visit. The second follow-up visit was at immediate post-training in ACTIVE and at six months in ADNI. The third, fourth, and fifth follow-up visits were at one, two, and three years after baseline in both ACTIVE and ADNI. The sixth follow-up visit in ACTIVE was five years after initial training.

Table 2

AVLT Growth Parameters Using Different Equating Methods: Results from ACTIVE (N=1,401)

	Raw scores	Equating method		
		Mean	Linear	Equipercentile
	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)
Means				
Control group				
Baseline (initial level)	47.0 (0.4)	48.1 (0.4)	48.1 (0.4)	47.9 (0.4)
Immediated pre-post training change	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
Slope	−0.5 (0.1)	−0.9 (0.1)	−0.9 (0.1)	−0.7 (0.1)
Memory-trained group				
Baseline (initial level)	48.9 (0.4)	48.9 (0.4)	48.9 (0.4)	48.9 (0.4)
Immediated pre-post training change	−5.0 (1.3)	5.7 (1.3)	5.9 (1.3)	4.1 (1.2)
Slope	0.0 (0.1)	−0.8 (0.1)	−0.8 (0.1)	−0.6 (0.1)
Group differences (Memory trained - Control)				
Baseline (initial level)	1.9 (0.6)	0.8 (0.6)	0.8 (0.6)	0.9 (0.5)
Immediated pre-post training change	−5.0 (1.3)	5.7 (1.3)	5.9 (1.3)	4.1 (1.2)
Slope	0.4 (0.1)	0.1 (0.1)	0.1 (0.1)	0.1 (0.1)
Variances				
Baseline (initial level)	91.5 (5.5)	91.6 (5.4)	93.3 (5.4)	89.0 (5.3)
Immediated pre-post training change	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
Slope	0.4 (0.2)	0.6 (0.2)	0.6 (0.2)	0.5 (0.2)
Correlation (Baseline, Slope)	0.2 (0.1)	0.1 (0.1)	0.1 (0.1)	0.0 (0.1)
Model fit statistics				
RMSEA	0.148	0.051	0.058	0.023
CFI	0.889	0.987	0.983	0.997

Legend. Multiple group latent growth models in ACTIVE of raw AVLT recall and mean, linear, and equipercentile equated scores. Results suggest equipercentile equated scores fit the hypothesized model best and that estimates of pre-post training change, long-term change, and correlation between initial level and slope are sensitive to the equating method used. Models assumed a linear function for time and accommodated nonlinear change in trajectory between baseline and post-training in the memory-trained group. The slope terms reflect annual change in AVLT recall, in units of total words recalled. The pre-post change parameter's mean in the control group and variance in both groups were fixed to 0 to identify the model.

Table 3

AVLT Growth Parameters Using Different Equating Methods: Results from ADNI (N=819)

	Raw scores	Equating method		
		Mean	Linear	Equipercentile
	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)
Means				
Healthy control group				
Baseline (initial level)	42.4 (0.6)	44.5 (0.6)	43.2 (0.6)	43.3 (0.6)
Slope	0.8 (0.3)	0.0 (0.3)	0.2 (0.2)	−0.2 (0.2)
Mild cognitive impairment (MCI) group				
Baseline (initial level)	30.2 (0.4)	32.3 (0.4)	30.7 (0.4)	30.8 (0.4)
Slope	−1.4 (0.2)	−2.2 (0.2)	−0.6 (0.2)	−0.9 (0.2)
Alzheimer's disease (AD) group				
Baseline (initial level)	22.6 (0.5)	24.7 (0.5)	23.0 (0.5)	23.2 (0.5)
Slope	−3.2 (0.3)	−4.2 (0.3)	−1.4 (0.3)	−1.7 (0.3)
Group differences (MCI - AD)				
Baseline (initial level)	7.6 (0.7)	7.6 (0.7)	7.7 (0.7)	7.7 (0.7)
Slope	1.8 (0.4)	1.9 (0.4)	0.9 (0.3)	0.8 (0.3)
Variances				
Initial level	59.4 (8.1)	61.9 (7.6)	65.9 (7.7)	65.5 (8.0)
Slope	−0.7 (2.7)	4.6 (1.4)	2.9 (1.1)	3.5 (1.1)
Correlation (Baseline, Slope)	0.3 (0.1)	0.3 (0.1)	−0.1 (0.1)	0.0 (0.1)
Model fit statistics				
RMSEA	0.12	0.07	0.06	0.00
CFI	0.95	0.98	0.99	1.00

Legend. Multiple group latent growth models in ADNI of raw AVLT recall and mean, linear, and equipercentile equated scores. Results suggest equipercentile equated scores fit the hypothesized model best and that estimates of pre-post training change, long-term change, and correlation between initial level and slope are sensitive to the equating method used. Models assumed a linear function for time. The slope terms reflect annual change in AVLT recall, in units of total words recalled. Latent growth parameter variances were held constant over diagnostic group.