

Aging, Neuropsychology, and Cognition

A Journal on Normal and Dysfunctional Development

ISSN: 1382-5585 (Print) 1744-4128 (Online) Journal homepage: <http://www.tandfonline.com/loi/nanc20>

Examining the reliability of ADAS-Cog change scores

Joseph H. Grochowalski, Ying Liu & Karen L. Siedlecki

To cite this article: Joseph H. Grochowalski, Ying Liu & Karen L. Siedlecki (2015): Examining the reliability of ADAS-Cog change scores, *Aging, Neuropsychology, and Cognition*, DOI: [10.1080/13825585.2015.1127320](https://doi.org/10.1080/13825585.2015.1127320)

To link to this article: <http://dx.doi.org/10.1080/13825585.2015.1127320>



Published online: 28 Dec 2015.



Submit your article to this journal [↗](#)



Article views: 7



View related articles [↗](#)



View Crossmark data [↗](#)

Examining the reliability of ADAS-Cog change scores

Joseph H. Grochowalski, Ying Liu and Karen L. Siedlecki

Department of Psychology, Fordham University, Bronx, NY, USA

ABSTRACT

The purpose of this study was to estimate and examine ways to improve the reliability of change scores on the Alzheimer's Disease Assessment Scale, Cognitive Subtest (ADAS-Cog). The sample, provided by the Alzheimer's Disease Neuroimaging Initiative, included individuals with Alzheimer's disease (AD) ($n = 153$) and individuals with mild cognitive impairment (MCI) ($n = 352$). All participants were administered the ADAS-Cog at baseline and 1 year, and change scores were calculated as the difference in scores over the 1-year period. Three types of change score reliabilities were estimated using multivariate generalizability. Two methods to increase change score reliability were evaluated: reweighting the subtests of the scale and adding more subtests. Reliability of ADAS-Cog change scores over 1 year was low for both the AD sample (ranging from .53 to .64) and the MCI sample (.39 to .61). Reweighting the change scores from the AD sample improved reliability (.68 to .76), but lengthening provided no useful improvement for either sample. The MCI change scores had low reliability, even with reweighting and adding additional subtests. The ADAS-Cog scores had low reliability for measuring change. Researchers using the ADAS-Cog should estimate and report reliability for their use of the change scores. The ADAS-Cog change scores are not recommended for assessment of meaningful clinical change.


ARTICLE HISTORY

Received 20 May 2015
Accepted 27 November 2015

KEYWORDS

ADAS-Cog; Alzheimer's disease; reliability; change scores; mild cognitive impairment

The Alzheimer's Disease Assessment Scale Cognitive Subscale (ADAS-Cog; Rosen, Mohs, & Davis, 1984) is a cognitive measure often used in studies that measure intervention and treatment efficacy for persons with Alzheimer's disease (AD). Despite its use to assess change over time, no studies to date have examined the reliability of the ADAS-Cog change scores across multiple administrations. A *change score* is the difference between a person's scores from one administration to another, used as a measure of growth or change in the trait measured by the scale. This study estimates the reliability of change scores on the ADAS-Cog, which is necessary (but not sufficient) to ensure that ADAS-Cog change scores are valid for measuring cognitive change or treatment efficacy

CONTACT Joseph H. Grochowalski  jgrochowalsk@fordham.edu

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.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.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

© 2015 Taylor & Francis

in clinical studies. We assess the effects of modifications (i.e., adding more subtests), subtest weighting, and score uses on change score reliability.

The ADAS measures broad areas, from cognitive ability to delusions and motor ability. Most research, however, has utilized only the cognitive subscale of the ADAS. The original ADAS-Cog, referred to as the ADAS-11 (Rosen et al., 1984), consists of 11 subtests: spoken language ability, comprehension of spoken language, recall of test instructions, word-finding difficulty, following commands, naming: objects and fingers, constructions: drawing, ideational praxis, orientation, word recall, and word recognition. Seven of the subtests (referred to as *scored subtests*) are scored by summing the number of errors made on each test, and four of the subtests (referred to as *rated subtests*) are scored as individual ratings by clinicians. Each subtest generates a subscore, and the total score on the ADAS-11 is an unweighted sum of all 11 subscores. A lower total score on the ADAS-11 indicates better cognitive performance overall (i.e., fewer errors are made and rating scores reflect limited or no impairment in performance). The ADAS-Cog has been shown to have high test-retest reliability (Rosen et al., 1984) and moderately high internal consistency (Weyer, Erzigkeit, Kanowski, Ihl, & Hadler, 1997).

However, scores from the ADAS-11, especially on the non-memory-related subtests, often suffer from the ceiling effect in persons with mild cognitive impairment (MCI; e.g., Mohs et al., 1997; Pyo, Elble, Ala, & Markwell, 2006). To increase the sensitivity of the ADAS-Cog at lower levels of cognitive impairment, variations of the ADAS-11 were developed by including additional subtests. Among them, ADAS-13, with additional delayed recall and digit cancellation tasks (Mohs et al., 1997), is frequently used, and is also the scale used in the current study.

Other modifications to the ADAS-Cog involve alternate weighting schemes. The total score of the ADAS-Cog is an unweighted sum of subscores, which implies arbitrary weighting due to the number of subtests mapped onto the different domains. For example, the three verbal memory subtests on the ADAS-13 (i.e., word recall, delayed word recall, word recognition) account for only 23% of the total score, while general cognitive subtests can account for as much as 46%. There is no explicit theoretical rationale for this allocation of subtest weights in the total score, so the interpretation of the composite scale score may be suspect due to arbitrary weighting.

While existing psychometric studies of the ADAS-Cog have focused on single administrations, the scale is often administered longitudinally throughout the progression of AD. A change score is often used to describe the degree of change in cognitive impairment over a given period of time. Change scores can be used for several purposes, including comparing relative change across persons, or examining how much an individual score has changed.

The level of change score reliability depends on how the score will be used. Many clinical studies assess treatment efficacy by comparing change scores on the ADAS-Cog between experimental groups (e.g., Mecocci, Bladström, & Stender, 2009). Other studies use change scores to define “responders to treatment” by a cardinal change in the ADAS-Cog score. For example, several studies define responders as those persons whose ADAS-Cog scores change more than 4 (or sometimes 7) points over a fixed period of time (e.g., Mega et al., 2005; Schrag, Schott, & Alzheimer’s Disease Neuroimaging Initiative, 2012; Winblad et al., 2001).

Despite researchers using change scores in these varied ways, there are no studies that assess the ADAS-Cog change scores, for any use. Although existing psychometric analyses found that scores from a single administration of the ADAS-Cog are reliable (Rosen et al., 1984; Weyer et al., 1997), one cannot infer that change scores from the ADAS-Cog are also reliable. For example, if all participants' scores increased by 5 points from time one to time two, and the scores from each time had high reliability, then participants could not be reliably ranked on their relative change; they would all have changed by exactly the same amount and the reliability would be zero, even if the true change itself were substantially large (see Miller & Kane, 2001). It is important to estimate the reliability of change scores because unreliable scores can increase Type-II error rates (Allen & Yen, 2001). True differences in individual or group change scores might go undetected because of low change score reliability.

Cronbach and Furby (1970) criticized the use of change scores because they often lack reliability for ranking persons, despite evidence of obvious and meaningful change. However, Kane and his colleagues (Kane, 1996; Miller & Kane, 2001) introduced methodology for measuring change that defines change in an absolute sense (i.e., the amount a person's score changes over time), and in a relative sense (i.e., ranking persons based on how much they changed). Kane (1996) showed that even if change scores are not reliable for ranking, they can still be dependable as an absolute measurement. In the example above where all participants' scores increased by 5 points, the relative reliability of change scores would be low, but the absolute reliability would be high, meaning researchers could reliably interpret the absolute change of 5 points.

Miller and Kane's (2001) solution to the change score reliability problem was to use generalizability theory. Generalizability theory (G theory; Brennan, 2010; Cronbach, Gleser, Nanda, & Rajaratnam, 1972) is distinct from other psychometric theories because it decomposes measurement error from complex sources. For example, classical test theory, upon which popular reliability measures such as Cronbach's alpha (Cronbach, 1951) were developed, only considers a lump-sum and unspecified error that is invariant across persons and test conditions. In contrast, G theory identifies different test and scoring conditions (referred to as *facets*), and quantifies their influences on the reliability, usually in the form of variances by sources of error using the factorial ANOVA framework. The analysis of these complex sources of variance is referred to as *the generalizability study* (the g-study). The estimated variance components are then used to model indices such as reliability in the original design where the data were collected, as well as alternative test and scoring designs. Such alternative designs may resemble the original design but with varying levels in facets (e.g., with a shorter or longer test, and/or with more or less raters than the original). The analysis of these modified conditions is referred to as *the decision study* (the d-study), because the results produced by this step (e.g., the reliability for an alternative design) may be used to decide whether the alternative design is worth pursuing (e.g., whether the design may provide satisfactory score precision) and/or outperforms the original. G theory offers the advantage and flexibility to explore optimal test and scoring designs, and to allow different score uses (see Brennan, 2010; Cronbach et al., 1972; for more on multivariate generalizability theory).

When investigating ADAS-Cog change scores using G theory, several types of change score reliability are of interest, because each use of change scores has its own reliability. Change scores can be (1) reliable for ranking persons' change relative to one another, as in a norm-referenced test, (2) reliable for interpreting a person's observed change score as an estimate of her true change score, and (3) useful for comparing an observed change score to another value, as often used with cut scores or measures of clinically meaningful change. The first type of reliability answers the question of how accurately persons can be ranked based on how much they have changed over time. This is referred to as the *relative reliability* of the change scores. The second form of reliability, referred to as the *absolute reliability* of the change scores, informs how accurately one can use an observed change score as an absolute measurement (i.e., interpreting a person's score without reference others' scores). The final form of reliability, referred to as *cut-score dependability*, allows one to compare persons' scores with a pre-designated criterion value (e.g., a 4-point true score change since a previous administration).

In this paper, we estimate the reliability of the ADAS-13 change scores under these three conditions of score use. In addition, we consider a few ways that may improve the reliability. One is to extend the overall length of the scale by including more subtests. Another way to improve reliability is to change the weighting scheme of the subtests in the composite score by, for example, giving more weight to subtests that have less error, and less weight to subtests containing more error, while balancing the subtests so that no particular subtests have too much influence in the total score. We also explore combining these two ways, i.e., both lengthening and re-weighting the total scale.

To summarize, the goals of the current paper are to assess change score reliability of the ADAS-Cog, and evaluate ways to improve reliability of the change scores.

Method

Participants and measures

The data used to estimate the reliability of the ADAS-13 came from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, the Food and Drug Administration, private pharmaceutical companies, and nonprofit organizations. The primary goal of ADNI has been to test whether serial magnetic resonance imaging, positron emission tomography, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD.

The principal investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California, San Francisco. ADNI is the result of efforts of many co-investigators, and subjects have been recruited from over 50 sites across the United States and Canada. To date, the ADNI, ADNI-GO, and ADNI-2 protocols have recruited over 1500 adults, ages 55–90, to participate in the research, consisting of cognitively normal older individuals, people with early or late MCI, and people with early AD. For up-to-date information, see www.adni-info.org. The data included

Table 1. Demographics of Alzheimer's Disease Neuroimaging Initiative sample.

	AD baseline (<i>n</i> = 153)		MCI baseline (<i>n</i> = 352)	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Age	75.5	7.3	74.9	7.3
Years of education	14.9	3.1	15.7	3.0
	<i>n</i>	%	<i>n</i>	%
Gender				
Male	80	53	226	64
Ethnicity				
Asian	2	1	9	3
Black	6	4	10	3
White	144	94	333	95
Other	1	1	0	0
Marital status				
Married	125	82	284	81
Divorced	6	4	21	6
Never married	7	5	4	1
Widowed	15	10	43	12

AD: Alzheimer's disease sample; MCI: mild cognitive impairment sample.

in this manuscript were obtained in compliance with regulations of the local institutional review board.

From the original ADNI sample of 819 participants, we analyzed all participants between the ages of 55 and 90 years who were diagnosed with amnesic MCI or AD. Participants periodically completed the ADAS-13, as well as additional psychological assessment scales over the course of the study. We used 153 complete records from baseline and 1-year administrations of the ADAS-13 for persons with AD, and 352 complete records from baseline to 1 year for persons with MCI. We analyzed the change scores for a 12-month span because this is a reasonable amount of time to expect measurable and meaningful change in cognitive performance for persons with AD and MCI. After 1 year, the mean ADAS-13 change score for the AD sample was 4.81, $t(152) = 8.76$, $p < .001$ (note that this average exceeds the 4-point change that some researchers set for clinical significance, suggesting that, on average, the AD sample exhibits clinically significant change). The mean ADAS-13 change score over 1 year for the MCI sample was 1.50, $t(351) = 5.31$, $p < .001$.

Table 1 shows the demographics of the sample used in the current study, and Table 2 includes descriptive statistics of the ADAS-13 scores by subtests. Additional measures from the ADNI database were analyzed to assess the validity of the modified change scores. Measures included ADAS-11, Mini Mental State Exam (MMSE; Folstein, Folstein, & McHugh, 1975), and the Rey Auditory Verbal Learning Test (RAVLT; Rey, 1941).

Analysis

Subtests of the ADAS-13 were first divided into three sections, because of the homogeneity of the measured content and similarity of the scoring method (i.e., self-report or rating by clinicians) in each. The composition of the three sections contain: tests measuring verbal memory (word recall, delayed word recall, word recognition),

Table 2. Mean and standard deviations of scores on the ADAS-13 subtests for the Alzheimer’s disease and mild cognitive impairment samples.

	Baseline		One Year		Baseline		One Year	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Word recall	6.06	1.48	6.39	1.61	4.52	1.39	4.84	1.47
Commands	0.37	0.60	0.54	0.85	0.18	0.47	0.22	0.48
Construction	0.80	0.65	0.90	0.80	0.53	0.56	0.55	0.60
Delayed recall	8.56	1.61	8.95	1.53	6.16	2.32	6.68	2.50
Naming	0.46	0.73	0.65	0.88	0.28	0.51	0.28	0.55
Ideational praxis	0.34	0.75	0.68	0.94	0.13	0.39	0.15	0.42
Orientation	2.05	1.69	3.05	2.05	0.64	0.93	0.92	1.25
Word recognition	6.48	2.79	7.53	2.95	4.62	2.77	4.74	3.26
Recall instructions	0.28	0.81	0.51	1.24	0.06	0.34	0.09	0.46
Spoken language	0.36	0.71	0.55	0.93	0.09	0.34	0.14	0.44
Word finding	0.63	0.92	0.95	1.13	0.27	0.58	0.40	0.72
Comprehension	0.30	0.66	0.51	0.90	0.07	0.31	0.12	0.41
Number cancellation	1.75	1.26	2.04	1.53	0.95	0.93	0.87	1.07

Subtests are listed in the order that they are administered.
M: mean; *SD*: standard deviation of subscores.

clinician-rated tasks (word-finding difficulty, spoken language ability, comprehension of spoken language, recall of test instructions), and general cognitive tests (naming: objects and fingers, constructions: drawing, ideational praxis, orientation, digit cancellation, commands). The three sections were then treated as three individual tests, each with its own score (by aggregating the corresponding subtest scores) and error variance, which could be combined into the total ADAS-13 score. The total ADAS-13 score is the same whether it is calculated by summing the three section scores or summing the 13 subtest scores. However, sectioning simplifies the reliability analysis for two reasons: first, when an alternate weighting scheme is considered, we only had to find three section weights to maximize reliability, rather than 13 subtest weights; second, we could reduce the error in the total analysis by attributing error to specific sections. Our division of the test into three sections is not entirely novel. Skinner et al. (2012) found an acceptable fit from a multifactor solution that identified the verbal and rating sections in addition to a general factor. Furthermore, the correlations among the three sections in the current data are weak, ranging from .20 to .38, showing that they are not strongly measuring the same construct (see Table 8).

Since we created three sections of ADAS-13 subtests, each person in our analysis had four ADAS-13 scores: a verbal memory section score, a rated section score, a general section score, and a total score. Change scores were then calculated by taking the differences of the scores at the baseline and in the 12-month follow-up. This resulted in four change scores per person.

The change scores were analyzed separately for the AD and MCI groups using G theory. The g-study design for all sections was $p \times s$, meaning all participants responded to (i.e., were crossed with) all subtests within sections. We did not include a random effect for raters in the rated section because there were no indicators for different raters in the ADNI data file. Variance components were estimated in R statistical software using the base package (R Core Team, 2014).

For each d-study design, three types of change score reliability estimates were computed as outlined by Miller and Kane (2001). The relative reliability and the absolute

reliability were calculated for each of the three sections as well as for the full scale. In addition, the cut-score dependability was evaluated for the full scale only, because there is no established cut score for each section. For the analysis in this paper, we considered a change score of 4, which has been used in the past to identify clinically relevant cognitive change (Winblad et al., 2001). The Appendix includes a more detailed and technical description of all reliability calculations.

It is worth noting that when we considered alternate weighting schemes, we estimated section variance impacts, or effective weights. The effective weights estimate the proportion of the total error that comes from each subtest and the proportion of the participant change score variability that comes from each subtest (Brennan, 2010). The sections' effective weights can be used to adjust the weights of the sections in the composite. For example, if one section is contributing mostly error variance, it can be underweighted to improve change score reliability. As no explicit rationale was offered by the original authors of the ADAS-Cog for the weighting of the subtests, we explored the use of reweighting as an empirical method for setting test and section weights. However, since reweighting the sections changes the calculation of the total score, the reweighted total scores could have different meaning from the unweighted scores. To collect preliminary validity evidence of the alternate composite scores and thus analyze the effects of reweighting on score interpretation, we correlated the new composite scores with other cognitive measures, including RAVLT, RAVLT-I, and MMSE.

Results

The first step in this generalizability analysis of change scores was to calculate the g-study variances for the both the AD and MCI populations, which are listed in Table 3.

The variances were used to calculate subsequent d-study variances and reliability estimates (see the Appendix for details). There are three sources of variance in this g-study design: (1) person universe score (true score) variance, (2) variance due to subtests, and (3) variance due to error. For the AD population, the person universe score variances were 0.06 for the Verbal Memory section, 0.35 for the rated section, and 0.10 for the general section, which fill the diagonal entries in the first three rows. The off-diagonal entries are the sections' universe score covariances. The fourth and fifth rows of

Table 3. G-study variance estimates for the Alzheimer's disease and mild cognitive impairment samples.

Sample	Source of variance		Verbal memory	Rated	General
AD	Persons	Verbal memory	0.06	0.22	0.05
		Rated	0.22	0.35	0.16
		General	0.05	0.16	0.10
	Subtests		0.13	0.00	0.10
		Error	4.06	0.70	1.09
MCI	Persons	Verbal memory	0.65	0.05	0.05
		Rated	0.05	0.05	0.01
		General	0.05	0.01	0.02
	Subtests		0.03	0.00	0.01
		Error	4.30	0.22	0.57

AD: Alzheimer's disease sample; MCI: mild cognitive impairment sample.

the table list the variance due to subtests within sections and error within sections, respectively.

The subtest variability ranged from 0.00 to 0.13 for the three sections, which are included in the fourth row of Table 3. The subtest variability of 0.00 for the rated section suggests that change scores did not vary much within persons. For example, if a participant had a score of 4 on one rated subtest, then the participant likely scored 4 on all of the subtests in the rated section. The error variances are the fifth row of Table 3, and the error variance in the verbal memory section is nearly 400% greater than the error in the other two sections, which is notable. Such a difference is notable because we sectioned the scale.

The bottom five rows of Table 3 contain the g-study variance estimates for the MCI population. The MCI person universe score variance estimates are much lower than the AD universe score variance estimates, meaning that the universe scores did not vary much from person to person for the MCI sample. Only the verbal memory section had non-negligible universe score variance, which means that any differentiation of participants' change scores on the entire scale would be based primarily on their verbal memory scores. The MCI subtest variance was lower than the AD subtest variance, and this was likely due to the floor effect (i.e., scores within subtests were uniformly low). Like the results for the AD sample, the error variance for the MCI sample was wide ranging, with a high error variance for the verbal memory section.

After the g-study variances were estimated, the reliability of the change scores was studied. First we assessed the reliability of change scores for the scale, as it is currently used. The sections of the ADAS-13 are included in Table 4 as three rows. The weight of each section in the test was calculated by dividing the number of subtests in the section by the total number of subtests (e.g., verbal memory's section weight was $3/13 = .23$). The last row of each sub-table (i.e., for AD or MCI group, respectively) includes the composite reliabilities for total change scores.

For the AD group, the verbal memory section had the lowest change score reliabilities ($Ep^2 = .05$ for relative use and $\phi = .21$ for absolute), and the rating section reliability is the highest (.67 and .70 for relative and absolute uses, respectively). The contributions or effective weights of the sections reveal that the verbal memory subtests contributed the least information useful for ranking persons ($ew(p) = .17$), but the most error

Table 4. Weight and reliability summary for the three-section ADAS-13 change scores with original weights.

		No. of subtests	Weight	Ep^2	ϕ	$\phi_C(\lambda)$	$ew(p)$	$ew(\Delta)$
AD	Verbal memory	3	.23	.05	.21	–	.17	.57
	Rating	4	.31	.67	.70	–	.49	.13
	General	6	.46	.36	.51	–	.34	.30
	Composite	13	–	.53	.59	.64	–	–
MCI	Verbal memory	3	.23	.31	.34	–	.68	.75
	Rating	4	.31	.48	.50	–	.15	.05
	General	6	.46	.19	.18	–	.17	.20
	Composite	13	–	.39	.40	.61	–	–

AD: Alzheimer's disease sample; MCI: mild cognitive impairment sample; Ep^2 is the relative change score reliability; ϕ is the absolute change score reliability; $\phi_C(\lambda)$ is the cut-score dependability; $ew(p)$ is the effective weight of section v on participant change score variance (proportion of score variance contributed by each section); and $ew(\Delta)$ is the effective weight of section v on error variance (proportion of error variance contributed by each section).

($ew(\Delta) = .57$), while the rating subtests contribute both the most useful information and the least error (.49 and .13, respectively).

The change scores were less reliable for the MCI group than the scores for the AD group, ranging from .39 to .61 for the MCI sample. The change scores for the verbal memory subtest heavily influenced these results, as 68% of useful information and 75% of error for the scale came from the verbal memory subtests. The other subtests contributed negatively to the scale, suggesting that the ADAS-13 is mostly an unreliable ($E\rho^2 = .39$) measure of change in verbal memory for persons with MCI when the scale is used to assess cognitive change.

The reliability of the change scores was low, and so we explored ways of improving it by hypothetically adding seven additional subtests in a d-study analysis. The choice of seven subtests was arbitrarily high, as it is approximately a 50% increase in scale length. For this analysis, we used the original section weights. Since the seven additional subtests would have to be assigned to one of the three sections, we used an optimization formula that assigns each of the additional subtests to the sections such that the composite scale score will have the lowest error variance possible (see the computational appendix for details about section length optimization). The optimal lengths for the sections were 8, 4, and 8 tests, respectively. Table 6 lists the weighting and reliability estimates for change scores on a test that has 20 items with optimized numbers of subtests in each section. The change score reliabilities under these conditions range from .67 to .76 for the AD group. For the MCI group, the reliabilities were still low, ranging from .57 to .76, despite the scale being lengthened by nearly 50%.

Because of the impracticality of making such a long scale, we used the information from the results from Table 4 (especially the effective weights) to reweight the sections in an effort to improve change score reliability. The alternate weights were chosen based on a few criteria: the weights should be nonzero and positive to allow for all sections to contribute to the score, the weights should be balanced such that no section dominates the score variance or the error variance, and the resulting composite total score should have a high correlation with ADAS-11 and ADAS-13 total scores (which we assess later).

As shown in Table 6, reliability for all three uses for the AD group improved to a more acceptable range of .68 to .76. This is a result of underweighting the verbal memory section to .10, and increasing the weights of the other two to .45 each. The reweighting decreased the error in the change scores; the effective weight of the verbal memory subtest to the total relative error variance decreased from 0.57 in Table 4 to 0.16 in Table 5. However, because the contribution of the verbal memory section was decreased, the amount of useful information it provided also decreased from 0.17 to 0.07. As a result of the decrease in information from the verbal memory section, rating subtests contributed almost two-thirds of the useful variance.

In the MCI sample, only the cut-score dependability entered the acceptable range. There is also a more even balance of useful variance across the three sections, such that the verbal memory section no longer dominates the scores.

The last option for analysis was to assess change score reliability when the scale is both reweighted and lengthened. We applied the same new weighting schema that was used in Table 6, and increased the scale length to 20 items, as in Table 5. Table 7 reports the results of these combined modifications. The improvement in reliability for both the AD and MCI scores are modest, compared to the reliability of the AD and MCI scores that

Table 5. Weight and reliability summary for the three-section ADAS-13 change scores with original weights and additional subtests.

		No. of subtests	Weight	Ep^2	ϕ	$\phi_c(\lambda)$	$ew(p)$	$ew(\Delta)$
AD	Verbal memory	8	.23	.11	.41	–	.17	.37
	Rating	4	.31	.67	.70	–	.49	.23
	General	8	.46	.43	.58	–	.34	.40
	Composite	20	–	.67	.72	.76	–	–
MCI	Verbal memory	5	.23	.55	.58	–	.68	.59
	Rating	6	.31	.48	.50	–	.15	.11
	General	9	.46	.24	.23	–	.17	.31
	Composite	20	–	.57	.58	.76	–	–

AD: Alzheimer’s disease sample; MCI: mild cognitive impairment sample; Ep^2 is the relative change score reliability; ϕ is the absolute change score reliability; $\phi_c(\lambda)$ is the cut-score dependability; $ew(p)$ is the effective weight of section v on participant change score variance (proportion of score variance contributed by each section); and $ew(\Delta)$ is the effective weight of section v on error variance (proportion of error variance contributed by each section).

Table 6. Weight and reliability summary for the three-section ADAS-13 change scores with new weights.

		No. of subtests	Weight	Ep^2	ϕ	$\phi_c(\lambda)$	$ew(p)$	$ew(\Delta)$
AD	Verbal memory	3	.10	.05	.21	–	.07	.16
	Rating	4	.45	.67	.70	–	.63	.41
	General	6	.45	.36	.51	–	.31	.43
	Composite	13	–	.68	.71	.76	–	–
MCI	Verbal memory	3	.10	.05	.21	–	.34	.32
	Rating	4	.45	.67	.70	–	.42	.25
	General	6	.45	.36	.51	–	.24	.43
	Composite	13	–	.42	.43	.74	–	–

AD: Alzheimer’s disease sample; MCI: mild cognitive impairment sample; Ep^2 is the relative change score reliability; ϕ is the absolute change score reliability; $\phi_c(\lambda)$ is the cut-score dependability; $ew(p)$ is the effective weight of section v on participant change score variance (proportion of score variance contributed by each section); and $ew(\Delta)$ is the effective weight of section v on error variance (proportion of error variance contributed by each section).

Table 7. Weight and reliability summary for the three-section ADAS-13 change scores with new weights and additional subtests.

		No. of subtests	Weight	Ep^2	ϕ	$\phi_c(\lambda)$	$ew(p)$	$ew(\Delta)$
AD	Verbal memory	3	.10	.05	.21	–	.07	.23
	Rating	7	.45	.78	.80	–	.63	.35
	General	9	.45	.43	.61	–	.31	.42
	Composite	20	–	.76	.78	.82	–	–
MCI	Verbal memory	5	.10	.45	.48	–	.34	.28
	Rating	6	.45	.57	.58	–	.42	.28
	General	9	.45	.26	.25	–	.24	.44
	Composite	20	–	.53	.54	.81	–	–

AD: Alzheimer’s disease sample; MCI: mild cognitive impairment sample; Ep^2 is the relative change score reliability; ϕ is the absolute change score reliability; $\phi_c(\lambda)$ is the cut-score dependability; $ew(p)$ is the effective weight of section v on participant change score variance (proportion of score variance contributed by each section); and $ew(\Delta)$ is the effective weight of section v on error variance (proportion of error variance contributed by each section).

are only reweighted (in Table 6) or only lengthened (in Table 5). The results from Tables 5 and 7 suggest that lengthening the test, whether reweighted or not, does not provide a practical improvement in score reliability.

The reweighting schema in Table 6 provided the largest and most practical improvement in change score reliability. We refer to these reweighted scores as ADAS-13RW, as they are simply composite change scores of the reweighted section scores from the

Table 8. Correlation matrix of the re-weighted ADAS-13 with existing forms and related cognitive measures.

	ADAS-13RW	ADAS-11	ADAS-13	MMSE	RAVLT	RAVLT-I	ADAS- Verb Mem	ADAS-Rating
ADAS-11	.94	–						
ADAS-13	.95	.98	–					
MMSE	–.43	–.47	–.44	–				
RAVLT	–.14	–.18	–.22	.08	–			
RAVLT-I	–.58	–.61	–.66	.24	.44	–		
ADAS-Verb Mem	.61	.80	.83	–.32	–.30	–.60	–	
ADAS-Rating	.75	.60	.56	–.22	.04	–.26	.20	–
ADAS-General	.82	.74	.77	–.42	–.12	–.50	.38	.34

ADAS-13RW: Alzheimer’s Disease Assessment Scale, 13-item version, reweighted; ADAS-11: Alzheimer’s Disease Assessment Scale, 11-item version; ADAS-13: Alzheimer’s Disease Assessment Scale, 13-item version; MMSE: Mini Mental State Exam; RAVLT: Rey Auditory Verbal Learning Test; RAVLT-I: Rey Auditory Verbal Learning Test, Immediate; ADAS-Verb Mem: Alzheimer’s Disease Assessment Scale verbal memory section; ADAS-Rating: Performance tasks on the Alzheimer’s Disease Assessment Scale that are rated by a clinician; ADAS-General: General tasks on the Alzheimer’s Disease Assessment Scale.

ADAS-13. Although the reweighted scores from the ADAS-13RW had improved reliability, reweighting the sections of the scale opens the possibility of substantially altering the meaning of the scores. To assess the impact of the new weights on score meaning, we correlated the ADAS-13RW scores with scale scores from related scales, including MMSE, RAVLT, RAVLT-I, the verbal memory section of the ADAS-13, the rating section of the ADAS-13, and the general subtest section of the ADAS13. Table 8 contains the correlations.

The original ADAS-11 and ADAS-13 correlate .98 with each other, and .95 and .94 with the ADAS-13RW, respectively. The correlation between the ADAS-13 and the verbal memory section was .83, and the correlation between the ADAS-13RW and the verbal memory section was .61, reflecting the down weighting of the verbal memory section in the new version. As a result, the rating and general sections had greater contribution to reliability than in the ADAS-11 and ADAS-13. Reweighting caused a negligible change in the relationship with the MMSE, as ADAS-13RW correlates with MMSE at –.43, which is comparable to ADAS-11 and ADAS-13 (–.47, –.44, respectively). ADAS-13RW has a correlation of –.14 with RAVLT (verbal learning test), which is lower than the ADAS-11 and ADAS-13 at –.18 and –.22, also a result of down weighting of the verbal memory section.

Discussion

We analyzed the change scores from the ADNI administration of the ADAS-13, estimated reliability for change scores with modified weights and different section lengths, and assessed whether the modified scores changed the construct measured by the ADAS-Cog. We found that the total change scores on the ADAS-13, measured over a 12-month period, were not adequately reliable for each of the three reliability analyses and uses or interpretations that we examined. Our analyses indicate that ADAS-13 change scores may not be accurate estimates of true change, and analysis of the scores for ranking or absolute interpretation may not be appropriate. When researchers do analyze ADAS-13 change scores, we recommend that they estimate and report the change score reliability as we did in this study.

The low reliability of ADAS-13 change scores was mostly due to the influence of the verbal memory section. The original weight of the verbal memory section was 0.23, and the section's relative reliability was only 0.05. As a result, nearly a quarter of the total change score information was unreliable.

We improved the total change score reliability by underweighting the verbal memory scores (adjusting the section's weight from .23 to .10). The estimates of score reliability improved, ranging from .68 to .76; although these are not high estimates of change score reliability, they are a meaningful improvement over the original estimates that ranged from .53 to .64.

We repeated the analysis for the MCI sample, but the change score estimates were not reliable, regardless of the modifications we chose. When we compared the reliabilities of the samples' ADAS-13 section scores, the MCI sample had more reliable change scores in the verbal memory section, but less reliable scores in the rated section. The MCI sample's lower reliability of change scores for the rated section was possibly due to a combination of the ceiling effect and consistent ratings over time. Since the MCI and AD samples had different rating and verbal memory score reliabilities, the reweighting that improved change score reliability for the AD sample did not improve reliability for the MCI sample.

We also assessed the effect of lengthening the ADAS-13 from 13 to 20 subtests, as lengthening scales is a common method for increasing score reliability. Adding seven hypothetical subtests improved change score reliability for the AD sample. However, this total scale length might cause fatigue. Several past modifications to the original ADAS-11 lengthened the scale, and although they reportedly increased the sensitivity of the scale at higher levels of cognitive functioning, they did not likely improve the reliability of change scores. Despite the addition of seven subtests, the MCI sample's change score reliability did not meaningfully increase. For the AD population, reweighting the sections is a more practical approach to increasing reliability of the ADAS-13 change scores.

Overall, the change scores for the MCI sample had poor reliability, despite the modifications we applied, suggesting that the existing scores from the ADAS-13 may not be reliable for measuring change in persons with MCI.

Finally, we considered whether the sectioning and reweighting of the scores changed the overall meaning of the ADAS-Cog scores. We divided the ADAS-13 scores into verbal memory, rating, and general subtests sections, and found the sections were weakly related, suggesting they measure relatively unrelated content, justifying our sectioning. When we correlated the total and section scores with other scales, we found negligible differences in the relationship between the reweighted scores, the original scores, and scores from scales that measure similar constructs. Furthermore, the intercorrelations of the ADAS-13RW, ADAS-13, and ADAS-11 ranged from .94 to .98, suggesting that they all measure the same construct.

To measure clinically meaningful change, we measured the reliability of change scores between baseline and 1 year. However, the reliability estimates we reported do not generalize to shorter or longer time spans (e.g., 6 months or 4 years). We conducted additional analyses for different time spans, and we found that change score reliability for the ADAS-13 decreases for periods shorter than 1 year and increases to an acceptable level for periods greater than 2 years. [Table 9](#) lists the reliabilities of the raw ADAS-13 change scores over various periods of measurement, which were calculated using the

Table 9. Reliability of ADAS-13 change scores over varying periods of time.

Group	Period (months)	<i>N</i>	<i>Ep</i> ²	<i>φ</i>	<i>φ_c</i> (<i>λ</i>)
AD	6	170	.20	.26	.52
	12	153	.53	.59	.64
	24	121	.69	.76	.71
	36	10	.68	.80	.74
MCI	6	372	.45	.46	.66
	12	352	.39	.40	.61
	18	320	.61	.60	.72
	24	294	.69	.72	.76
	36	225	.76	.78	.79

AD: Alzheimer’s disease sample; MCI: mild cognitive impairment sample; *Ep*² is the relative change score reliability; *φ* is the absolute change score reliability; and *φ_c*(*λ*) is the cut-score dependability for a score change of 4 points.

same method that was used to produce Tables 3 and 4 (note that the values for 12 months in Table 9 are the same composite reliabilities as those reported in Table 4).

Limitations

Several issues limit our analysis of the ADAS-Cog scores. We could not assess rater variability, the section weights were not canonically estimated, and our sectioning of the ADAS-Cog subtests is somewhat arbitrary. The section reweighting improved reliability and retained score meaning, but we did not estimate weights that have the maximum possible reliability using canonical methods (Joe & Woodward, 1976), as these methods could result in negative section weights. Thus, a more exhaustive analysis of weighting schemas could result in improved section weights. Our analysis was also limited because the ADNI data did not include a variable that identified the raters, so we could not assess the variability of ADAS-Cog raters, which could change the estimates of absolute reliability and cut-score dependability if raters are highly variable in scoring.

In conclusion, we assessed the reliability of ADAS-Cog change scores over a 12-month period in samples of individuals with AD and MCI. Our results show that the change scores lack adequate reliability for the typical uses of ADAS-Cog change scores. Evaluations of ways to improve change score reliability indicated that only reweighting the subsections of the ADAS-13 improved the reliability of the change scores, and only for the AD sample. Our findings suggest that use of the ADAS-Cog change scores is not recommended for assessment of meaningful clinical change.

Disclosure statement

The authors report no conflict of interest.

Funding

Data collection and sharing for this project was funded by the Alzheimer’s Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer’s Association; Alzheimer’s Drug Discovery Foundation; Araclon

Biotech; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; 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 Southern California.

References

- Allen, M. J., & Yen, W. M. (2001). *Introduction to measurement theory*. Long Grove, IL: Waveland Press.
- Brennan, R. L. (2010). *Generalizability theory*. New York, NY: Springer.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334. doi:10.1007/BF02310555
- Cronbach, L. J., & Furby, L. (1970). How we should measure "change": Or should we? *Psychological Bulletin*, 74, 68–80. doi:10.1037/h0029382
- Cronbach, L. J., Gleser, G., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York, NY: Wiley.
- 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. doi:10.1016/0022-3956(75)90026-6
- Joe, G. W., & Woodward, J. A. (1976). Some developments in multivariate generalizability. *Psychometrika*, 41, 205–217. doi:10.1007/BF02291839
- Kane, M. (1996). The precision of measurements. *Applied Measurement in Education*, 9, 355–379. doi:10.1207/s15324818ame0904_4
- Mecocci, P., Bladström, A., & Stender, K. (2009). Effects of memantine on cognition in patients with moderate to severe Alzheimer's disease: Post-hoc analyses of ADAS-Cog and SIB total and single-item scores from six randomized, double-blind, placebo-controlled studies. *International Journal of Geriatric Psychiatry*, 24, 532–538. doi:10.1002/gps.2226
- Mega, M. S., Dinov, I. D., Porter, V., Chow, G., Reback, E., Davoodi, P., ... Cummings, J. L. (2005). Metabolic patterns associated with the clinical response to galantamine therapy: A fludeoxyglucose f 18 positron emission tomographic study. *Archives of Neurology*, 62, 721–728. doi:10.1001/archneur.62.5.721
- Miller, T. B., & Kane, M. (2001). The precision of change scores under absolute and relative interpretations. *Applied Measurement in Education*, 14, 307–327. doi:10.1207/S15324818AME1404_1
- Mohs, R. C., Knopman, D., Petersen, R. C., Ferris, S. H., Ernesto, C., Grundman, M., ... Thai, L. J. (1997). Development of cognitive instruments for use in clinical trials of antidementia drugs: Additions to the Alzheimer's disease assessment scale that broaden its scope. *Alzheimer Disease & Associated Disorders*, 11, 13–21. doi:10.1097/00002093-199700112-00003
- Pyo, G., Elble, R. J., Ala, T., & Markwell, S. J. (2006). The characteristics of patients with uncertain/mild cognitive impairment on the Alzheimer disease assessment scale-cognitive subscale. *Alzheimer Disease & Associated Disorders*, 20, 16–22. doi:10.1097/01.wad.0000201846.22213.76

- R Core Team. (2014). *R: A language and environment for statistical computing [Computer Software Manual]*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <http://www.R-project.org/>
- Rey, A. (1941). L'examen psychologique dans les cas d'encéphalopathie traumatique. *Archives de Psychologie*, 28, 215–285.
- Rosen, W. G., Mohs, R. C., & Davis, K. L. (1984). A new rating scale for Alzheimer's disease. *The American Journal of Psychiatry*, 141, 1356–1364. doi:10.1176/ajp.141.11.1356
- Schrag, A., Schott, J., & Alzheimer's Disease Neuroimaging Initiative. (2012). What is the clinically relevant change on the ADAS-Cog? *Journal of Neurology, Neurosurgery & Psychiatry*, 83, 171–173. doi:10.1136/jnnp-2011-300881
- Skinner, J., Carvalho, J. O., Potter, G. G., Thames, A., Zelinski, E., Crane, P. K., ... Gibbons, L. E. (2012). The Alzheimer's disease assessment scale-cognitive-plus (ADAS-Cog-plus): An expansion of the adas-cog to improve responsiveness in mci. *Brain Imaging and Behavior*, 6, 489–501. doi:10.1007/s11682-012-9166-3
- Weyer, G., Erzigkeit, H., Kanowski, S., Ihl, R., & Hadler, D. (1997). Alzheimer's disease assessment scale: Reliability and validity in a multicenter clinical trial. *International Psychogeriatrics*, 9, 123–138. doi:10.1017/S1041610297004298
- Winblad, B., Brodaty, H., Gauthier, S., Morris, J. C., Orgogozo, J.-M., Rockwood, K., ... Wilkinson, D. (2001). Pharmacotherapy of Alzheimer's disease: Is there a need to redefine treatment success? *International Journal of Geriatric Psychiatry*, 16, 653–666. doi:10.1002/gps.496

Appendix

The generalizability change score analysis in this paper follows these steps:

- Step 1. Calculate the change scores for each person on each subtest.
- Step 2. Estimate g-study variances for the sections.
- Step 3. Estimate d-study variances for any desired modifications to the scale, including reliability coefficients.
- Step 4. Combine the d-study variances to estimate the reliability coefficients for the full change scores. The equations and procedures for the change score analysis are discussed below.

G-study

For each section of the ADAS-13, calculate the g-study variance due to persons, $\sigma_v^2(p) = [MS_v(p) - MS_v(pi)]/n_{iv}$, where $MS_v(\cdot)$ indicates the mean square of the facet, as estimated in analysis of variance (ANOVA), and n_{iv} are the number of items in section v . Similarly, calculate the g-study variance due to items, $\sigma_v^2(i) = [MS_v(i) - MS_v(pi)]/n_p$, where n_p is the number of persons, and the g-study interaction variance, $\sigma_v^2(pi) = MS_v(pi)$.

D-study

For each section v , determine the number of desired subtests, n_{iv} . For the analysis in this paper, we analyzed the original number of items, and also optimized numbers of subtest items for a total scale length of 20 subtests (explained in a later step). First, calculate the d-study universe score variance $\sigma_v^2(\tau) = \sigma_v^2(p)$, the relative error variance, $\sigma_v^2(\delta) =$

$\sigma_v^2(pi)/n_{iv}$, and the absolute error variance, $\sigma_v^2(\Delta) = \sigma_v^2(pi)/n_{iv} + \sigma_v^2(i)/n_{iv}$. From these variance estimates, the relative reliability for section v can be estimated:

$$E\rho_v^2 = \frac{\sigma_v^2(\tau)}{\sigma_v^2(\tau) + \sigma_v^2(\delta)}$$

and the absolute reliability:

$$\phi_v = \frac{\sigma_v^2(\tau)}{\sigma_v^2(\tau) + \sigma_v^2(\Delta)}$$

Combined d-study variances

To calculate the reliability of the change scores for the full scale, the d-study variances from the previous step must be combined. First, however, section covariances are estimated:

$$\sigma_{vv'}(\tau) = \frac{n_p}{n_p - 1} \left(\frac{\sum_p \bar{X}_{pv} \bar{X}_{pv'}}{n_p} - \bar{X}_v \bar{X}_{v'} \right)$$

where \bar{X}_{pv} is the mean score of person p for section v , and \bar{X}_v is the mean of section v scores. Then, the d-study universe score variances from the previous step are combined with the covariances:

$$\sigma_c^2(\tau) = \sum_v \sum_{v'} w_v w_{v'} \sigma_{vv'}(\tau)$$

where w_v is the weight of section v in the composite. Similarly, the composite relative error is estimated:

$$\sigma_c^2(\delta) = \sum_v w_v^2 \sigma_v^2(\delta)$$

and the composite absolute error is estimated:

$$\sigma_c^2(\Delta) = \sum_v w_v^2 \sigma_v^2(\Delta)$$

These d-study composite variances are then combined to estimate the composite relative reliability estimate:

$$\rho_c^2 = \frac{\sigma_c^2(\tau)}{\sigma_c^2(\tau) + \sigma_c^2(\delta)}$$

and the composite absolute reliability estimate:

$$E\rho_c^2 = \frac{\sigma_c^2(\tau)}{\sigma_c^2(\tau) + \sigma_c^2(\Delta)}$$

The last form of composite reliability, the cut-score dependability, is similar to the composite absolute reliability estimate, except it requires designation of a cut score λ for the full-scale change score. The cut-score dependability estimate takes the form

$$\phi_c = \frac{\sigma_c^2(\tau) + \hat{\delta}_c^2}{\left[\sigma_c^2(\tau) + \hat{\delta}_c^2 \right] + \sigma_c^2(\Delta)}$$

where $\hat{\delta}_c^2 = \sum_v w_v (\mu_v - \lambda_v)^2$, with

$$(\mu_v - \lambda_v)^2 = (\bar{X}_v - \lambda_v)^2 - \frac{\sigma_v^2(p)}{n_p} - \frac{\sigma_v^2(i)}{n_{iv}} - \frac{\sigma_v^2(pi)}{n_p n_{iv}}$$

where \bar{X}_v is the mean of the difference scores for the administration 1 year after baseline, and λ_v is the designated cut score weighted for the section: $\lambda_v = w_v \lambda$.

The other equations used in this analysis were the optimization formula, which determines the optimal number of subtests in each section for a hypothetical full scale length, and the effective weight formal, which provide information about a section's contribution to universe score variance and error variance. The optimization formula is

$$n'_{vi} = \frac{n'_{i+} w_v \sqrt{\sigma_v^2(i) + \sigma_v^2(pi)}}{\sum_v w_v \sqrt{\sigma_v^2(i) + \sigma_v^2(pi)}}$$

where n'_{i+} is the new hypothetical test length for the full scale, and n'_{vi} is the optimal number of subtests i in section v to minimize error. The effective weight of section v on the universe score variance is estimated:

$$ew_v(\tau) = \frac{w_v \sum_{v'=1}^{n_v} w_{v'} \sigma_{v'}(\tau)}{\sigma_c^2(\tau)}$$

And the effective weight of section v on the error variance can be calculated by substituting Δ for τ in the equation above.