Gaining precision on the Alzheimer's Disease Assessment Scale-cognitive: A comparison of item response theory-based scores and total scores

Steve Balsis[a,*], Alexis A. Unger[a], Jared F. Benge[b], Lisa Geraci[a], Rachelle S. Doody[c]

[a]Department of Psychology, Texas A&M University, College Station, TX, USA
[b]Department of Physical Medicine and Rehabilitation, Jack C. Montgomery VA Medical Center, Muskogee, OK, USA
[c]Department of Neurology, Baylor College of Medicine Alzheimer's Disease and Memory Disorders Center, Houston, TX, USA

Abstract

Background: The Alzheimer's Disease Assessment Scale-cognitive (ADAS-cog) is a commonly used measure for assessing cognitive dysfunction in patients with Alzheimer's disease (AD). The measure has 11 subscales, each of which captures an important aspect of cognitive dysfunction in AD. Traditional scoring of the ADAS-cog involves adding up the scores from the subscales without regarding their varying difficulty or their strength of relationship to AD-associated cognitive dysfunction. The present article analyzes problems associated with this approach and offers solutions for gaining measurement precision by modeling how the subscales function.

Methods: We analyzed data collected at the Baylor College of Medicine Alzheimer's Disease and Memory Disorders Clinic from 1240 patients diagnosed with varying degrees of dementia. Item response theory was used to determine the relationship between total scores on the ADAS-cog and the underlying level of cognitive dysfunction reflected by the scores.

Results: Results revealed that each total score corresponded to a spectrum of cognitive dysfunction, indicating that total scores were relatively imprecise indicators of underlying cognitive dysfunction. Furthermore, it was common for two individuals with the same total score to have significantly different degrees of cognitive dysfunction.

Conclusions: These findings suggest that item response theory scoring of the ADAS-cog may measure cognitive dysfunction more precisely than a total score method.

© 2012 The Alzheimer's Association. All rights reserved.

Keywords: ADAS; Alzheimer's disease; Clinical trial; Item response theory; Measurement

1. Introduction

The Alzheimer's Disease Assessment Scale-cognitive (ADAS-cog) [1] is a commonly used measure for assessing cognitive change in clinical trials of Alzheimer's disease (AD) medications [2,3]. In fact, it has been highlighted in regulatory guidelines and translated into many languages for use in clinical trials all over the world in places such as the United States, South Africa, Japan, Australia, South America, and Western and Eastern Europe [4–7]. Frequently, in combination with other measures, the ADAS-cog is used to gauge the cognitive effect of a given treatment. A treatment is said to have affected cognitive ability if there is a change in total score on the measure from before treatment to after treatment, relative to a control condition. If the change in the scores is different beyond a statistical threshold, relative to a control condition, and confirmed with an independent measure (such as global or activities of daily living scales), a meaningful change is thought to occur as a result of the treatment.

The purpose of this study is to illustrate a problem with this seemingly solid approach to measuring cognitive dysfunction (and cognitive change) with the ADAS-cog and to offer an alternative method of scoring this measure. The problem

we would like to address centers on the use of total scores, which may provide a relatively imprecise estimate of underlying cognitive dysfunction. The possible total scores obtained on the ADAS-cog range from 0 to 70, with 0 being associated with little or no cognitive impairment and 70 being associated with severe cognitive impairment [1]. An individual may miss any combination of items on the subscales to yield a given total score. In fact, there are more than 2.16 billion possible combinations of raw scores across the subscales, with literally millions of ways to obtain many single total scores (There are 2,161,665,792 patterns of raw scores across the subscales. This number should be calculated by the rule of permutations, which states that, "If any one of K mutually exclusive and exhaustive events can occur on each of N trials, there are K^N different sequences that may result from a set of such trials" [p 130] [8]. If two six-sided dice are rolled, there are 6^2 or $6 \times 6 = 36$ possible different patterns of scores that can be obtained. In other words, there are 36 different ways to make the possible raw scores, which range from 2 to 12. A similar calculation can be made with regard to the ADAS-cog. There are different possible integer raw scores for each scale [8 scales with 6 possible scores {0 through 5}, a scale with 9 possible scores, 1 with 11, and 1 with 13]. The total number of possible patterns across the scales $= 6^8 \times 9 \times 11 \times 13 = 2.16$ billion. Thus, there are more than 2 billion patterns that can give rise to the 71 possible whole number raw scores [0 through 70]. Of course, a portion of the subscales is more complex than others. Even individual scores on these subscales can be reached in multiple ways. The number of possible response patterns gets exponentially larger if one begins to consider these complexities. In this way, one can consider 2.16 billion to be a lower-bound estimate of the possible response patterns. Because there are more than 2 billion patterns of raw scores, there are necessarily millions of patterns that could give rise to many of the individual raw scores). Ignoring the pattern of scores and merely summing them may miss important information. This stems from the fact that different parts of the ADAS-cog may be differentially related to the construct of cognitive dysfunction that it measures. Thus, a person can score a 17 by making errors on easy subscales, difficult subscales, or some combination of both easy and difficult subscales, and these subscales may be more or less strongly associated with cognitive dysfunction.

However, whether people make errors on easy, difficult, strongly related, or weakly related subscales has not been considered historically; instead, total scores have been calculated by summing scores from individual subtests. Thus, if a person scored a 17 before and after receiving a drug treatment, he/she would be labeled a nonresponder by current total score interpretation of the ADAS-cog, even if that person had a different pattern of responses across subscales. What is needed is a statistical framework within which we can consider that different patterns of responses can give rise to the same raw score, and the different patterns are not necessarily "created equal." Each pattern may reflect a slightly different degree of cognitive dysfunction, and this is because each subscale is differentially sensitive to measuring cognitive dysfunction. If we could model which scores were received on which subscales, along with the information to arrive at a more precise estimate of dementia severity, then we might be able to determine more accurately whether a patient who received a similar raw score at two time points did in fact experience an increase or decrease in level of cognitive dysfunction, even though an examination of this patient's total scores indicates nonresponse (that there was essentially no change). Using a statistical framework known as item response theory (IRT) [9,10], it is possible to determine whether cognitive dysfunction in AD is measured more precisely by considering information about these differences across the subscales.

2. Methods

2.1. Participants

This study included 1240 patients, who were recruited and assessed longitudinally as published elsewhere [11], from the Baylor College of Medicine Alzheimer's Disease and Memory Disorders Clinic (BCM ADMDC). Consented patients eligible for the current analysis were diagnosed with dementia and completed the ADAS-cog during their most recent clinic visit. Of the participants included, the majority (84%) had probable AD (National Institute of Neurological and Communicative Diseases and Stroke/Alzheimer's Disease and Related Disorders Association [12], Diagnostic and Statistical Manual of Mental Disorders, Third Edition, Revised), whereas the remaining individuals had possible AD (6%) or met the criteria for both probable AD and another disorder, such as dementia with Lewy bodies (5%), or had vascular dementia with or without AD (5%). The average age of the participants was 74.73 years (SD = 8.45). Sixty-three percent were women, 92% were Caucasian, and overall they had attained an average of 13.8 years of education (SD = 3.5). Regarding participants' level of cognitive impairment, the average score on the ADAS-cog was 32.38 (SD = 16.05), 17.03 (SD = 7.37) on the Mini-Mental State Exam, and 1.53 (SD = 0.88) on the Clinical Dementia Rating total (see Table 1 for demographic and sample statistics).

2.2. Measures

The ADAS-cog consists of 11 subscales designed to assess various areas of cognitive function, including word recall, recognition memory, orientation, language comprehension, expressive language, and praxis [1].

2.3. Procedure

Patients at the BCM ADMDC receive neurological and neuropsychological evaluations at baseline and annually thereafter. The details of this protocol can be found

Table 1
Demographics and sample statistics

Variable	Mean/%
Age	74.7 (SD = 8.5)
Education	13.8 years (SD = 3.5)
Sex	63% women
Race	92% Caucasian
ADAS-cog raw score	32.4 (SD = 16.1)
MMSE raw score	17.0 (SD = 1.5)
CDR raw score	1.5 (SD = .9)

Abbreviations: ADAS-cog, Alzheimer's Disease Assessment Scale-cognitive; MMSE, Mini-Mental State Exam; CDR, Clinical Dementia Rating.

elsewhere [12]. A diagnosis is assigned based on a consensus conference review of the medical records, psychometric test scores, and the medical evaluation by the staff at BCM ADMDC. Each patient is followed longitudinally and returns annually for neurological and neuropsychological evaluation as well as management of current medications.

The ADAS-cog is administered both at baseline and at each of the annual visits thereafter. For the purpose of this study, the results from the most recent ADAS-cog administration for each patient included have been used. This procedural detail allowed us to gather the full range of dementia severity from the ADAS-cog scores. If we had used only baseline administration of the ADAS-cog, it is likely that mild cases of dementia would have been oversampled.

2.4. Data analysis

An IRT framework was used for the core data analyses in this study. To use this framework, the latent variable measured by the scale being used must be unidimensional [9], that is, it must measure only one latent variable. It is noteworthy that measures commonly used in any domain are rarely perfectly unidimensional. To confirm that the data did indeed meet the assumption of unidimensionality, we performed factor analyses of this same dataset and found that it met the unidimensionality assumption. This analysis yielded a first factor with an eigenvalue of 6.88 (62.48% of the variance), and no other factors came close to accounting for a significant proportion of the variance (none had an eigenvalue >1). In addition, the ratio between the first and second eigenvalues was 7.73, a very large ratio [13]. The exploratory factor analysis was supported by the results of a confirmatory factor analysis. As with the preliminary analysis, a strong fit was found for a one-factor model (Tucker–Lewis index = 0.98, comparative fit index = 0.92). As outlined by Hu and Bentler [14], when the Tucker–Lewis index [15] and the comparative fit index [16] values are "close to 0.95" (larger is better and 1.00 is perfect), one has evidence to conclude that a hypothesized model provides an adequate fit to the observed data. Based on the results of the exploratory factor analysis, the confirmatory factor analysis, and the relative robustness of IRT, we can be confident that these

data are in fact unidimensional enough for these analyses. This is not to say that the ADAS-cog measures only a singular cognitive process; rather, it is to say that there is enough covariation among the subscales to render it adequate for this type of analysis. In other words, these analyses require the grouping of seemingly disparate cognitive functions such as constructional and ideational praxis. In fact, these are distinct neuropsychological constructs. Their grouping in the current study is not related to whether they are distinct neuropsychological constructs but is related to their covariation in this demented sample.

IRT analyses for the current study were run using MULTILOG software [17]. Using Samejima's graded model [18], the probability of obtaining a given score on a subtest at each level of the latent dimension, θ, was determined. Parameters were estimated to define how the subscales functioned, and then that information about the subscales and each patient's pattern of scores across the subscales was used to generate a particular person's score along the latent variable, known as θ. Commonly, a person's place along the latent variable is estimated with a *maximum a posteriori* (MAP) [19,20]. A MAP is a posterior estimate of θ that considers both responses to items and how items of the scale function. The two previous references [19,20] delve into the details of MAP estimation; the interested reader should seek these references. The fact that this is a standard procedure and MULTILOG software [17] carries out the computations instantly is noteworthy.

3. Results

In general, results confirmed our hypothesis. Individuals with the same total score tended to have somewhat different levels of corresponding cognitive dysfunction. For example, one participant who had a total score of 28 on the measure had the following response pattern for the 11 subscales on the ADAS-cog: 7/0/1/0/11/0/0/0/2/0/7. Here, each number corresponds to the score on a given subscale of the test. For example, this participant made 7 errors on the first subscale, 0 on the second subscale, 1 on the third, and so on. This particular participant's response pattern corresponded to a θ of −0.49 SDs. In contrast, another participant in the sample who also had a total score of 28 on the ADAS-cog had the following pattern of errors: 1/3/1/1/8/0/3/3/2/3/3, a pattern that corresponds to a θ of 0.15 SDs.

For complete results, refer to Figure 1. Notice that total scores for this sample ranged from 4 to 70 with scores at each and every point within this range on the ADAS-cog. Each of the graphs in Figure 1 displays unit increments in total ADAS-cog scores from 0 to 70. Each point represents the total score of one participant and the corresponding level of θ. As can be seen, for any given total score, there is a range or spectrum of MAPs, which are estimates of cognitive dysfunction. That is, people who obtained the same total score had various corresponding levels of θ. For example, for a total score of 15 on this measure, individuals ranged
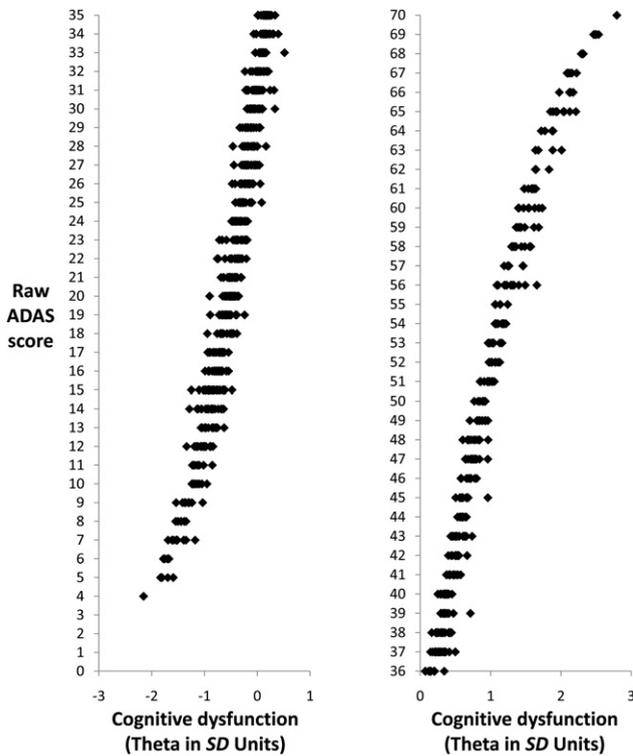
Fig. 1. Raw scores and estimates of cognitive dysfunction on the Alzheimer's Disease Assessment Scale-cognitive (n = 921).

from −1.29 SDs to −0.51 SDs of cognitive impairment. Thus, the width of the band for a total score of 15 is 0.78 SDs. Across all scores, 4 to 70, the width of the bands varied substantially. For example, the band for the total score of 44 was only 0.13 SDs.

Additionally, there was a substantial overlap between the spectra of scores identified for different total scores. For example, individuals with a total score of 22 on the ADAS-cog had levels of cognitive dysfunction that ranged from −0.79 to −0.23 SDs, whereas individuals with a total score of 23 had cognitive dysfunction that ranged from −0.75 to −0.21 SDs. Thus, there was an overlap of 0.52 SDs between these two abutting total scores; the values of θ overlapped from −0.75 to −0.23 SDs. With the exception of total scores 4 and 70, there were no instances of a case in which two abutting total scores did not have overlapping estimates of θ to some degree. The fact that estimates of θ overlapped suggests that total scores may be relatively imprecise indicators of cognitive dysfunction.

To confirm that the θ values at each total score did in fact represent different degrees of cognitive dysfunction, we calculated their statistical significance. Statistical differences can be established by creating confidence intervals around two specific MAP scores by using the standard errors (SEs) associated with each MAP. The SE for the first data point described earlier in the text associated with total score 28 and MAP score −0.49 is 0.23, and the SE of the second data point with the same total score of 28 and MAP score of 0.15 is 0.22. Using these two key values (0.23 and 0.22), the

SE of the difference (SE_{DIFF}) was calculated to be 0.32. Recall that the $SE_{DIFF} = (SE_1^2 + SE_2^2)^{1/2}$. The most relevant confidence interval can be calculated by multiplying the SE_{DIFF} by 1.96, which here yields the 95% confidence interval. The number 1.96 multiplied by the SE_{DIFF} (0.32) equals 0.63. If 0.15 is more than 0.63 units from −0.49, then it follows that these two data points (the MAPs from these two individuals with the same total score) differ beyond this confidence interval and thus are statistically different from one another. Not only are these two data points statistically different from one another, they also represent more than two-thirds of an SD difference, which is considered to be a large difference in both social and medical science research. Clearly, if the data from these two people were compared using only total scores, a statistical (or meaningful) difference between these points could not be found. In fact, using this approach one would have to conclude that there was no difference between these two individuals because both received the same exact total score, and the difference between these total scores would be zero if they were used as repeated measures in a clinical trial.

Results showed that individuals with the same total scores could have different degrees of cognitive dysfunction. The converse was also true; individuals with the same amount of cognitive dysfunction could have very different total scores. For example, two individuals in this sample had MAPs = −0.36 SDs. However, one of these individuals had the following response pattern on the test: 6/2/1/0/7/0/2/0/1/2/1, which corresponds to a total score of 22, whereas the other individual had the following response pattern: 6/0/3/5/9/2/0/0/0/1/3 (see Figure 2 for clearer explication), which corresponds to a total score of 29. Using classical methods of determining change, simply comparing total scores, there is a statically significant difference between these two scores. However, IRT scoring, which models not only the scores on items but also how those items function, shows no difference between these two individuals, because the difference between their θ values is zero.
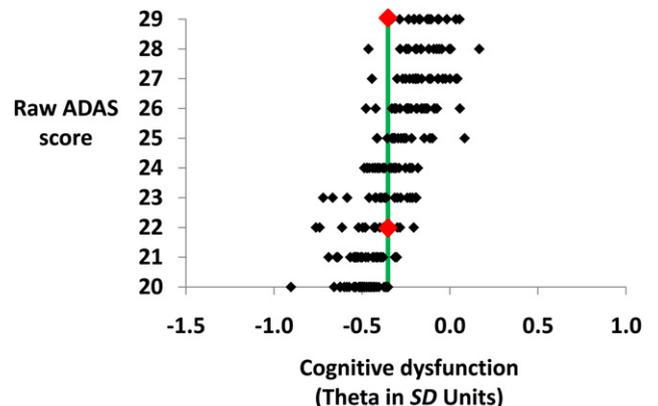

Fig. 2. Two patients with the same degree of cognitive dysfunction but significantly different raw scores.

An important caveat to note, we also ran the analyses using only the consensus "pure AD" group and found remarkably similar results. The data continued to be adequately unidimensional for IRT analysis and continued to display the same main effect (see Figure 1 for main effect). The original analysis yielded a first factor with an eigenvalue of 6.88 (62% of the variance), and no other factors came close to accounting for a significant proportion of the variance (none had an eigenvalue >1). In addition, the ratio between the first and second eigenvalue was very large, 7.73. In this second analysis, we found nearly identical results. The first factor had an eigenvalue of 6.84 (62%) of the variance, and no other factor was of much relevance (here again none was >1). The ratio of first to second eigenvalue remained very large, 7.70. Given the fact that the results hardly varied, we are comfortable reporting the original analysis, an analysis that generalizes to a slightly larger group. The fact that one factor dominated both analyses indicates that although particular diagnoses are associated with constellations of clinical symptoms, they do indeed share a large common component. This is consistent with current research and prevailing theory, which propose that amnestic mild cognitive impairment likely represents prodromal AD for most patients and that other forms of dementia may share similar presentations.

4. Discussion

Results from our IRT analyses indicated that individuals with the same total score on the ADAS-cog can have different levels of cognitive dysfunction. The severity spectra associated with each total score tended to overlap substantially with the severity spectra of abutting total scores (Figure 1). Conversely, individuals with the same degree of cognitive dysfunction could have very different total scores. Both of these findings have important implications.

From a clinical perspective, total scores are used to characterize the severity of cognitive dysfunction on the ADAS-cog. More precise estimates of cognitive dysfunction can be obtained by not only considering scores on subscales but also by considering how those subscales function, and by weighting the scores appropriately. From a research perspective, cognitive differences between patients or within patients (longitudinally in a clinical trial) can be more readily and precisely estimated than they are at present. For example, a patient could obtain the same total score at time 1 and time 2 during a clinical trial, but have a significant increase or decrease in their level of cognitive dysfunction. Traditionally, the possibility that there could be underlying cognitive changes occurring given a constant total score or a small difference in total scores (as occurs frequently in both treated and untreated patients) has not been considered.

Our results suggest that focusing on total scores alone may provide misleading information about the underlying level of cognitive dysfunction, which can be particularly problematic for interpreting differences in cross-sectional studies and/or change in longitudinal research and clinical trials. An IRT analysis approach may provide a more precise indicator of the actual level of an individual's cognitive change over time. At present, this is only speculation as the results from this study are cross-sectional.

Although the current study has important clinical implications, it is limited by several factors. First, the study is limited by the particular sample used. The individuals who participated in this study were mostly white and highly educated and thus the results may not generalize to other populations, especially those of greater ethnic diversity or lower educational levels. Whether the results generalize is an empirical question that could be examined within an IRT framework rather easily using techniques such as differential item functioning. Additionally, we used a diagnostically heterogeneous sample. Despite this fact, the majority of the patients had probable AD, and their diagnoses were established with clinical consensus meetings, which are known for their accuracy in correctly identifying cases of AD. In addition, this study assessed the particular psychometric properties related to change in cognitive dysfunction in only one instrument, the ADAS-cog. Thus, the results may not generalize to other measures of cognitive dysfunction and do not take into account all aspects of the more comprehensive evaluative batteries commonly used in clinical trials. To increase the generalizability of the current findings, future studies should include other important neuropsychological tests that are regularly used to assess cognitive dysfunction in Alzheimer's research.

Another limitation of the study is that the number of raw score response patterns on the ADAS-cog (approximately 2.16 billion) far outstrips the sample size of this study. The sample used for this study was relatively small, and so conceivably, the spectra at each total score could have been even larger with the inclusion of more participants. It is noteworthy that although this study only has a portion of the possible response patterns on the ADAS-cog subscales represented here, many of the possible combinations of scores are not probable (e.g., profound impairments in praxis and naming would not be expected in an individual with perfect delayed memory performance). One strength of IRT is that the a and b parameters found in this study could be used to estimate the latent score on the underlying dimension of cognitive dysfunction for any individual given his/her unique response pattern [21]. Although the value of θ (cognitive dysfunction) has only been calculated for a small portion of the nearly 2 billion different response patterns to the ADAS-cog, θ can be calculated for any response pattern of raw scores on the measure. In fact, once the item parameters for a particular measure are known (see Appendix for item parameters used in this study), any possible pattern of responses can be input into a program such as MULTILOG, and the resultant estimation will be output in (typically) less than a second. Thus, in practice, one could estimate the θ value for any real or hypothetical response pattern in a very simple and straightforward manner. This possibility is important

because it allows for these findings to have a large effect in how we measure cognitive dysfunction with the ADAS-cog both in basic research and clinical trials.

The overall findings in this article fit with what is perhaps common knowledge but may not have been so adequately demonstrated. The θ values provide much more precise estimates of cognitive dysfunction than the raw scores and could be harnessed to more precisely estimate dementia severity in future studies. In other words, scoring the ADAS-cog while taking into consideration how the subscales function, one can arrive at a more precise estimation of cognitive dysfunction, designated here by θ.

In the field of AD clinical trials, the impact of the current results is potentially far-reaching. Our ability to predict and detect both differences between people and changes within individuals may be able to be greatly improved using the IRT approach outlined in the current study. In drug trials, the IRT approach may be able to increase our ability to correctly identify group treatment differences and to identify responders and nonresponders to treatment. The results of this study also have the potential to inform our theory of AD as a clinical entity. Many of the current questions revolve around the ability to correctly identify those who will develop dementia and those who will not. The ability to focus on patterns of performance and latent scores rather than on total scores alone has the potential to add more detail to our understanding of cognitive impairment in dementia. The psychometric framework of IRT might allow us to more precisely identify individuals who are at risk for cognitive decline. Future work, for example, may examine whether patterns of scores across items or subtests (and their IRT aggregation) on cognitive screening measures are more predictive of eventual conversion from mild cognitive impairment to AD than total scores alone. As such, IRT scoring has the potential to significantly reduce the number of patients needed in a clinical trial and to reduce heterogeneity in studies, two important goals in this field [22].

References

[1] Rosen WG, Mohs RC, Davis KL. A new rating scale for Alzheimer's disease. Am J Psychiatry 1984;141:1356–64.

[2] Rockwood K, Fay S, Gorman M, Carver D, Graham J. The clinical meaningfulness of ADAS-cog changes in Alzheimer's disease patients treated with donepezil in an open-label trial. BMC Neurol 2007;7:26.

[3] Birks J. Cholinesterase inhibitors for Alzheimer's disease. Cochrane Database Syst Rev 2006:CD005593.

[4] Rosler M, Anand R, Cicin-Sain A, Gauthier S, Agid Y, Dal-Bianco P, Stahelin HB, Hartman R, Gharabawi M. Efficacy and safety of rivastigmine in patients with Alzheimer's disease: international randomised controlled trial. Br Med J 1999;318:633–8.

[5] Wilcock GK, Lilienfeld S, Gaens E. Efficacy and safety of galantamine in patients with mild to moderate Alzheimer's disease: multicentre randomised controlled trial. Br Med J 2000; 321:1445–9.

[6] Raina P, Pasqualina S, Ismaila A, Patterson C, Cowan D, Levine M, Booker L, Oremus M. Effectiveness of cholinesterase inhibitors and memantine for treating dementia: evidence review for a clinical practice guideline. Ann Int Med 2008;148:379–97.

[7] Wilkinson D, Murray J. Galantamine: a randomized, double-blind, dose comparison in patients with Alzheimer's disease. Int J Geriatr Psychiatry 2001;16:852–7.

[8] Hays WL. Statistics. 5th ed. New York, NY: Harcourt College Publishers; 1994.

[9] Hambleton RK, Swaminathan H, Rogers HJ. Fundamentals of item response theory. Newbury Park, CA: Sage Publications; 1991.

[10] Baker FB. The Basics of Item Response Theory. ERIC Clearinghouse on Assessment and Evaluation. College Park, MD: University of Maryland; 2001.

[11] Doody RS, Pavlik V, Massman P, Rountree SD, Darby E, Chan W. Changing patient characteristics and survival experience in an Alzheimer's disease center patient cohort. Dement Geriatr Cogn Disord 2005;20:198–208.

[12] McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM. Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services task force on Alzheimer's disease. Neurology 1984;34:939–44.

[13] Kline RB. Principles and practice of structural equation modeling. New York, NY: The Guilford Press; 2004.

[14] Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. Struct Equ Modeling 1999;6:1–55.

[15] Tucker LR, Lewis C. A reliability coefficient for maximum likelihood factor analysis. Psychometrika 1973;38:1–10.

[16] Bentler PM. Comparative fit indexes in structural models. Psychol Bull 1990;107:238–46.

[17] Thissen D, Chen WH, Bock R. MULTILOG 7 for Windows: multiple-category item analysis and test scoring using item response theory [computer software]. Lincolnwood, IL: Scientific Software International, Inc; 2003.

[18] Samejima F. Estimation of Latent Ability Using a Response Pattern of Graded Scores (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society; 1969. Retrieved from: http://www.psychometrika.org/journal/online/MN17.pdf.

[19] Thissen D, Orlando M. Item response theory for items scored in two categories. In: Thissen D, Wainer H, eds. Test Scoring. Hillsdale, NJ: Lawrence Erlbaum Associates; 2001. p. 73–140.

[20] Thissen D, Nelson L, Rosa K, McLeod LD. Item response theory for items scored in more than two categories. In: Thissen D, Wainer D, eds. Test Scoring. London: Erlbaum; 2001. p. 141–86.

[21] Gershon RC, Cella D, Fox NA, Havlik RJ, Hendrie HC, Wagster MV. Assessment of neurological and behavioural function: the NIH Toolbox. Lancet Neurol 2010;9:138–9.

[22] Hachinski V. Shifts in thinking about dementia. J Am Med Assoc 2008; 300:2172–3.

Appendix: Subtest parameters for ADAS-cog

1. Recall error: $a = 2.87$, $b1 = -3.24$, $b2 = -2.99$, $b3 = -2.65$, $b4 = -2.25$, $b5 = -1.65$, $b6 = -1.10$, $b7 = -0.65$, $b8 = -0.11$, $b9 = 0.46$, $b10 = 1.15$.

2. Naming: $a = 2.44$, $b1 = -0.65$, $b2 = 0.22$, $b3 = 0.95$, $b4 = 1.49$, $b5 = 2.07$.

3. Construction: $a = 1.71$, $b1 = -1.47$, $b2 = 0.12$, $b3 = 0.92$, $b4 = 1.84$, $b5 = 2.25$.

4. Ideational praxis: $a = 2.45$, $b1 = -0.36$, $b2 = 0.45$, $b3 = 0.92$, $b4 = 1.24$, $b5 = 1.52$.

5. Word recognition: $a = 2.59$, $b1 = -2.63$, $b2 = -1.84$, $b3 = -1.34$, $b4 = -1.01$, $b5 = -0.76$,

b6 = −0.51, b7 = −0.28, b8 = −0.07, b9 = 0.15, b10 = 0.31, b11 = 0.55, b12 = 0.82.

6. Speech: a = 2.58, b1 = 0.18, b2 = 0.53, b3 = 0.91, b4 = 1.32, b5 = 2.04.

7. Comprehension: a = 3.02, b1 = −0.26, b2 = 0.09, b3 = 0.54, b4 = 1.56, b5 = 2.28.

8. Recall instructions: a = 3.13, b1 = −0.51, b2 = −0.18, b3 = 0.09, b4 = 0.42, b5 = 0.61.

9. Word finding: a = 2.57, b1 = −0.60, b2 = −0.03, b3 = 0.45, b4 = 1.05, b5 = 1.89.

10. Commands: a = 2.76, b1 = −0.46, b2 = 0.27, b3 = 0.77, b4 = 1.41, b5 = 2.25.

11. Orientation: a = 2.18, b1 = −1.70, b2 = −1.12, b3 = −0.81, b4 = −0.53, b5 = −0.18, b6 = 0.25, b7 = 0.72, b8 = 2.11.