The ADAS-cog in Alzheimer's Disease clinical trials: Psychometric evaluation of the sum and its parts

Stefan J Cano, Holly Posner, Margaret Moline, Stephen Hurt, Jina Schwartz, Tim Hsu, Jeremy Hobart

The ADAS-cog in Alzheimer's Disease clinical trials:

Psychometric evaluation of the sum and its parts.

Stefan J Cano,[1] Holly B Posner,[2] Margaret L Moline,[3] Stephen W Hurt,[5] Jina Swartz,[6] Tim Hsu,[3] Jeremy C Hobart[1]

[1]*Clinical Neurology Research Group, Peninsula College of Medicine and Dentistry, Plymouth, UK*

[2]*Pfizer Inc, New York, NY, USA (formerly of Eisai Medical Research Inc)*

[3]*Eisai Medical Research Inc, Ridgefield Park, NJ, USA*

[5]*Weill Medical College of Cornell University, NY, USA*

[5]*Eisai Global Clinical Development, London, UK*

Address correspondence and reprint requests to:

Dr Jeremy Hobart, Department of Clinical Neuroscience, Peninsula College of Medicine and Dentistry Room N16 ITTC Building, Tamar Science Park, Davy Road, Plymouth, Devon PL6 8BX, UK

T: +44 (0) 1752 315272; F: +44 (0) 1752 315254; E: Jeremy.Hobart@pms.ac.uk

Word count (excluding title page, abstract, references, figures and tables): 3073

Keywords: Alzheimer's disease, clinical trials, rating scales, reliability, validity

ABSTRACT

Background & Aims: The Alzheimer's Disease Assessment Scale Cognitive Behaviour Section (ADAS-cog), a measure of cognitive performance, has been used widely in AD trials. Its key role in clinical trials should be supported by evidence that it is both clinically meaningful and scientifically sound. Its conceptual and neuropsychological underpinnings are well-considered, but its performance as an instrument of measurement has received less attention.

Objective: To examine the traditional psychometric properties of the ADAS-cog in a large sample of people with AD.

Methods: Data from three clinical trials of donepezil (Aricept®) in mild-to-moderate AD (n=1421; MMSE 10-26) were analysed at both the scale- and component-level. Five psychometric properties were examined using traditional psychometric methods. These methods of examination underpin upcoming FDA recommendations for patient rating scale evaluation.

Results: At the scale-level, criteria tested for data completeness, scaling assumptions (eg component total correlations= 0.39-0.67), targeting (no floor or ceiling effects), reliability (eg Cronbach's alpha = 0.84; test retest intraclass correlations = 0.93) and validity (correlation with MMSE = -0.63) were satisfied. At the component-level 7 of 11 ADAS-cog components had substantial ceiling effects (range 40-64%).

Conclusions: Performance was satisfactory at the scale level, but most ADAS-cog components were too easy for many patients in this sample and did not reflect the expected depth and range of cognitive performance. The clinical implication of this finding is that the ADAS-cog's estimate of cognitive ability, and its potential ability to detect differences in cognitive performance under treatment, could be improved. However, because of the limitations of traditional psychometric methods, further evaluations would be desirable using additional rating scale analysis techniques to pinpoint specific improvements.

INTRODUCTION

Alzheimer's disease (AD) is a terminal dementing neurodegenerative disease that impacts on cognition and behaviour.[1] It is the most common form of dementia, affecting approximately 27 million people worldwide,[2] and incidence rates are expected to quadruple by the middle of this century.[3] Considerable interest and resources have been targeted at slowing AD progression as reflected in the growing number of clinical trials in AD.[4]

The most widely used primary outcome measure in these clinical trials has been the AD Assessment Scale - Cognitive Behaviour Section (ADAS-cog).[5,6] It was developed in the early 1980s in response to the then perceived lack of appropriate instruments available to test the efficacy of AD drug treatments,[5,6] to assess the “severity of dysfunction and research in patients with Alzheimer’s Disease”.[6(p1360)] Since its inception, the ADAS-cog has been used in over 127 AD clinical trials, and although developed specifically for AD, it has frequently been used in non-AD populations, including mild cognitive impairment,[7] vascular dementia,[8] and Parkinson’s disease.[9] Of particular relevance to the present study is that clinical trials are increasingly focusing on people earlier in the disease process and with less severe AD. As awareness increases, diagnoses are likely to be made much earlier than they were 25 years ago.

If the ADAS-cog is to be considered fit for future measurement of all severities of AD including milder forms, it should satisfy stringent criteria as a reliable and valid measure of cognitive performance. Awareness of this issue is now widely recognised by international regulatory agencies concerning the use of patient rating scales. The ADAS-cog was developed with sound consideration of relevant neuropsychological consequences of AD, but without being subjected to rigorous psychometric techniques of rating scale construction. Although we are unsure as to the precise reasons the ADAS-Cog was developed in this way, the lack of standard rating scale construction methods may have resulted from a lack of awareness. As such, although these methods have existed for decades, they have been rarely applied to clinical rating scale research.

At its introduction, data on the ADAS-cog were provided on inter-rater and test-retest reliability in small samples of AD (n=27) and normal elderly (n=28).[6] Since then, it has undergone additional scale-level psychometric evaluations,[12-14] with some authors suggesting possible key limitations.[15,16] The reason as to why psychometric evaluations of rating scales before their use are important requires a brief overview of the key issues surrounding the use of rating scales as outcomes measures.

Measurement requires the construction of an instrument for carrying out the practical process of *measuring*. Some variables, like height, can be measured directly and by relatively straightforward means. Other variables, like cognitive performance, need to be approached ***indirectly*** through quantifying their manifestations. It is important here to note that in its role as a clinical assessment tool the relevance of evaluating the ADAS-Cog using rating scale testing methods is appropriate but less crucial than to do so for its role as a measurement instrument for clinical research. This is because clinical assessment and measurement are different processes that have different requirements. We have previously summarised these,[11] but the key issue is that measurement has a specific meaning with respect to the quantification of attributes. By contrast, clinical assessment is, frequently, a qualitative process. Here instrument development is not straightforward and requires the construction of tools that transform numerically graded manifestations into measurements of underlying variables. Indirectly measured variables are often called latent (hidden) variables to emphasise this fact.

Rating scales are constructed to measure latent variables. It is customary for a rating scale to consist of a set of items, each of which represents a different manifestation. In relation to the ADAS-cog we have referred to these as components, as the eleven questions used are more detailed, time consuming and involved than traditional rating scale items. Every item is scored, and item scores are combined to give a total score for each person. This value is a *measure* of the variable quantified by the set of items.

Whether a rating scale generates clinically meaningful and scientifically sound measurements depends on decisions during its construction and its performance during testing. The decisions concern the components selected to form the set, their clinical grading and numerical scoring, and how components are combined to give a single value. Performance is tested against a number of predefined measurement (psychometric) criteria.

The original ADAS-cog measures cognitive performance by combining ratings of 11 components (Word Recall, Word Recognition, Constructional Praxis, Orientation, Naming Objects and Fingers, Commands, Ideational Praxis, Remembering Test Instruction, Spoken Language, Word Finding, Comprehension) representing six broad areas of cognition: memory, language, ability to orientate oneself to time, place and person, construction of simple designs and planning and performing simple behaviors in pursuit of a basic, predefined goal.[5,6] Seven of the eleven ADAS-cog components are scored as the "number incorrect". For example, in the commands component, the number of five commands performed incorrectly (none, 1, 2, 3, 4, or all 5). The remaining four ADAS-cog components are scored from 0 (no limitations) to 5 (max limitations) as the examining clinician's perception of: remembering test instructions, spoken language ability, word finding, and comprehension. Scores for the 11 components are summed, without weighting, into a total ADAS-cog score. Low total scores indicate better cognitive performance. Appendix 1 shows the component structure of the ADAS-cog. Note that the 11 components have different score ranges.

This process appears clinically appropriate but requires empirical proof that it "works". This means that evidence is needed to support the choice of items forming the set, scoring of the individual items, and appropriateness of combining item scores into a single score. Also evidence should be available demonstrating that the single score is a reliable and valid measure of cognitive performance. Psychometric methods provide formal frameworks for gathering this evidence.

There are two main types of psychometric method; traditional and modern.[11] Traditional methods are the most widely used analytic strategy for determining rating scale reliability and

validity and will be reported here.[17] These are the psychometric methods best understood by clinicians and clinical researchers, and underpin the forthcoming FDA guidelines for rating scales.[10] The aim of this study was to provide clinicians and researchers with a traditional psychometric evaluation of the ADAS-cog, which goes beyond the existing published examinations in type (ie detailed evaluations of data quality, scaling assumptions, targeting, reliability, validity) and kind (ie the inclusion of scale- and importantly component-level analyses).

METHODS

Setting and Participants

Anonymized screening and baseline data from three large clinical trials of donepezil[18-20] in people with AD were pooled for analysis. The inclusion criteria were: healthy ambulatory people aged ≥50 with a diagnosis of probable AD, of mild to moderate severity (Clinical Dementia Rating 1 or 2), with a Mini-Mental State Examination (MMSE) score between 10 and 26 and uncomplicated by stroke.

Data analysis

Many clinicians are familiar with reliability and validity testing, but a more thorough traditional psychometric evaluation involves the assessment of six properties: data quality, scaling assumptions, targeting, reliability, validity and responsiveness. Data completeness concerns the extent to which a scale's components are completed in the target sample, and the percent of people for whom it is possible to report a single score. Tests of scaling assumptions examine whether it is appropriate statistically to sum the 11 components to generate a single scale score. Targeting assesses the match between the range of cognitive performance measured by the ADAS-cog and the range of cognitive performance in the sample. Reliability describes the extent to which scale scores are free from random error. Validity refers to the extent to which the ADAS-cog measures cognitive performance. Responsiveness is the ability to detect accurately true change in cognitive

performance when it has occurred. We examined five of these six psychometric propeties (see Appendix 2) which are extensively documented elsewhere.[21-23]

RESULTS

Sample

1,418 of 1,421 patients tested provided sufficiently complete ADAS-cog component scores. The sample is characterized in (Table 1). The main analyses were undertaken in the total sample. Additional targeting, reliability and validity analyses were conducted in MMSE subgroups (10-14 moderately severe; 15-20 moderate; 21-26 mild) to examine the impact of cognitive impairment on the psychometric properties of the ADAS-cog. The outcomes of the original clinical trials and further specification of the study populations are provided elsewhere.[18-20]

Table 1: Respondent Characteristics (N=1421)

Characteristics	**Mean, SD (range)**
Age	72, 8 (50-94)
Gender	**Percentages**
Female	59
Male	41
Ethnicity	
White	95
Black/Caribbean	2
Hispanic	2
Other	1

Psychometric properties

Data completeness (Tables 2&3)

Data completeness was high. The proportion of component-level missing data was low (≤0.02%). ADAS-cog total scores could be computed for 99.7% of the sample (1418/1421).

Table 2: ADAS-cog Scale level analyses* - Data Completeness, Scaling Assumptions, Targeting, Reliability, Validity (N=1421)

	ADAS-cog TOTAL
Psychometric property	
Data completeness**	
Computable scale scores (%)	100
Scaling Assumptions	
Corrected ITC	0.39 – 0.67^
Targeting	
Possible range	0-70
Range midpoint	35
Score range	3-61
Mean score	24.0
SD	10.7
F/C effect (%)	0/0
Skewness	0.7
Reliability	
Internal consistency	
Cronbach's alpha (n=1418)	0.84
SEM	4.3
95% CI	+/-8.4
Mean IIC (n=1418)	0.39
Range IIC	0.18 – 0.70
Test Retest reproducibility	
ICC Consistency***	0.93
ICC Absolute***	0.93
Correlation ***	0.93
Validity****	
Correlation with MMSE	0.63

* The analyses and interpretation of the statistics presented in this table relating to data completeness, scaling assumptions, targeting, reliability, validity are further described in Appendix 2 and also presented in Tables 3, 4 and Appendix 3. In brief, data completeness includes percentage of missing data and computable scale scores; scaling assumptions involved tests of the legitimacy of summing components based on component means, standard deviations and item total correlations; targeting involved analyses of sccale score distributions; reliability included tests of random error including internal consistency and test retest reproducibility; validity invloved within and between scale correlational analyses and known groups analyses focussing on MMSE sub-samples.

**<0.5% MD rounded to 0

^Range ITC

***TRT between screening and baseline

**** expanded in Table 4

Table 3: ADAS-cog Component level analyses - Data Completeness, Scaling Assumptions, Targeting, Reliability* (N=1421)

Psychometric property	Word recall (BASED 3)	Naming objects and fingers	Comm-ands	Construct-ional praxis	Ideational praxis	Orientation	Word recognition	Remembering test instruction	Spoken language	Word finding	Comp-rehension
Data completeness											
Component MD (%)**	0	0	0	0	0	0	0	0	0	0	0
Scaling Assumptions											
Possible range	0-10	0-5	0-5	0-5	0-5	0-8	0-12	0-5	0-5	0-5	0-5
Component range midpoint	5	2.5	2.5	2.5	2.5	4	6	2.5	2.5	2.5	2.5
Component score range	1-10	0-5	0-5	0-5	0-5	0-8	0-12	0-5	0-5	0-5	0-4
Component mean score	6.9	0.9	0.8	1.3	0.7	3.2	6.3	1.4	0.6	1.0	0.7
Component standard deviation	1.6	1.0	1.1	1.0	1.2	2.1	3.2	1.6	1.0	1.1	1.0
Corrected item total correlations	0.66	0.55	0.56	0.39	0.55	0.59	0.57	0.67	0.58	0.55	0.62
Targeting											
Possible range	0-10	0-5	0-5	0-5	0-5	0-8	0-12	0-5	0-5	0-5	0-5
Range midpoint	5	2.5	2.5	2.5	2.5	4	6	2.5	2.5	2.5	2.5
Score range	1-10	0-5	0-5	0-5	0-5	0-8	0-12	0-5	0-5	0-5	0-4
Mean score	6.9	0.9	0.8	1.3	0.7	3.2	6.3	1.4	0.6	1.0	0.7
Standard deviation	1.6	1.0	1.1	1.0	1.2	2.1	3.2	1.6	1.0	1.1	1.0
Floor/Ceiling effect (%)	0/1	43/1	53/1	20/1	60/3	11/0	0/5	42/10	64/0	40/0	59/0
Skewness	-0.4	1.3	1.3	0.8	2.0	0.1	0.1	1.0	1.7	1.0	1.2
Test Retest Reproducibility											
Intraclass correlation coefficient (absolute agreement)***	0.79	0.81	0.69	0.77	0.75	0.77	0.79	0.82	0.83	0.81	0.79

*The analyses and interpretation of the statistics presented in this table relating to data completeness, scaling assumptions, targeting, and test retest reproducibility are further described in Appendix 2 (see also Table 2 legend).

**<0.5% MD (missing data) rounded to 0

^Range ITC

***Test retest between screening and baseline

Scaling Assumptions (Tables 2&3)

The ADAS-cog satisfied most criteria for scaling assumptions. For example, component-total correlations (corrected for overlap) for the eleven ADAS-cog components ranged from 0.39-0.67 satisfying the recommended criteria. This supported the scale components as measures of a common underlying construct, and indicated that components contained a similar proportion of information about that construct.

However, Table 3 shows that ADAS-cog component mean scores and variances were not especially similar. Whilst this implies some criteria for scaling assumptions were not satisfied, it is important to note that ADAS-cog components have different numbers of response categories. Thus, mean scores and variances were similar for components with the same/similar numbers of response categories providing evidence that these criteria were fulfilled.

Targeting (Tables 2, 3; Appendix 3)

The ADAS-cog total scores spanned approximately 83% of the entire scale range, with no significant floor and ceiling effects, and were not notably skewed. This was also found for the Word Recall, Word Recognition, and Orientation components. However, 8/11 components (Naming Objects and Fingers, Commands, Constructional Praxis, Ideational Praxis, Remembering Test Instruction, Spoken Language, Word Finding, Comprehension) had significant floor/ceiling effects (40%-64%) and were notably skewed (+1.0 to +2.0). These findings indicate adequate scale-to-sample targeting but potentially poor component-to-sample targeting. They indicate that the range of cognitive performance measured by these 8 components is poorly matched to the ranges of cognitive performance in this sample.

Reliability (Tables 2&3)

Cronbach's alpha, and test-retest ICCs for the ADAS-cog scale were high (0.84 and 0.94), supporting their reliability. Component level ICCs (range 0.75 – 0.83) were also well above the suggested minimum of 0.50.

Validity (Table 2)

Correlations between the ADAS-cog and MMSE were near our prediction at both screening (-0.63) and baseline (-0.74). Correlations between the ADAS-cog at baseline and sociodemographic variables (age and sex) were -0.01 and -0.07, respectively, indicating that ADAS-cog scores were not biased by these variables. These findings provided evidence for convergent and discriminant construct validity.

MMSE subgroups (10-14 moderately severe; 15-20 moderate; 21-26 mild; Table 4)

Targeting analyses revealed that ADAS-cog component-level ceiling effects progressively increased as the severity of AD, measured by the MMSE, decreased (range: moderately severe - 0-32%; moderate - 0-59%; mild 0-82%; Table 4). Reliability, as assessed by Cronbach's alpha and test-retest ICCs were low (range 0.62 to 0.75 and 0.71-0.77, respectively). Finally, the examination of group differences validity revealed a stepwise decrease in ADAS-cog score as the MMSE score increases. The mean scores for the three groups are significantly different, in line with prediction ($F=404.22$, $p<.0001$). However, correlations between ADAS-cog and MMSE scores within each group were low to moderate (0.17 – 0.49) and much lower than the predicted association between these two measures of cognitive performance, and that found in the total sample.

Table 4: ADAS-cog psychometric analyses by MMSE subgroups (score ranges 10-14, 15-20 and 21-26) *

	Subgroup by MMSE score								
Sample (n)	207			507			692		
MMSE range	10-14			15-20			21-26		
MMSE mean score (SD)	12.5, (1.4)			17.9 (1.7)			23.4 (1.7)		
ADAS-cog mean score (SD)	38.8 (8.4)			26.8 (8.9)			19.0 (7.2)		
ADAS-cog range	18-61			7-53			3-52		
	CITC	Correlation with MMSE (r)	Ceiling effect (% scoring zero)	CITC	Correlation with MMSE (r)	Ceiling effect (% scoring zero)	CITC	Correlation with MMSE (r)	Ceiling effect (% scoring zero)
Word recall	0.31	-	0	0.51	-	0	0.47	-	0
Naming objects and fingers	0.33	-	10	0.45	-	38	0.29	-	64
Commands	0.40	-	14	0.45	-	49	0.31	-	70
Constructional praxis	0.18	-	6	0.19	-	15	0.08	-	34
Ideational praxis	0.33	-	12	0.38	-	53	0.24	-	82
Orientation	0.17	-	0	0.31	-	8	0.34	-	17
Word recognition	0.31	-	0	0.46	-	0	0.38	-	0
Remembering test instruction	0.45	-	14	0.59	-	34	0.45	-	59
Spoken language	0.40	-	32	0.48	-	59	0.32	-	79
Word finding	0.35	-	16	0.45	-	34	0.30	-	55
Comprehension	0.56	-	23	0.53	-	54	0.35	-	73
ADAS-cog TOTAL	0.67** 0.77^	0.25[s] 0.27[b]	0	0.75** 0.71^	0.48[s] 0.49[b]	0	0.62** 0.76^	0.17[s] 0.30[b]	0

*Tests included scaling assumptions (corrected-item-total correlations; CITC), reliability (alphas, test-retest intraclass correlation coefficients), and validity (correlations between the ADAS-cog and MMSE)

** This table shows selected psychometric analyses of sub-samples as defined by the MMSE: 10-14 (middle left column); 15-20 (middle column); 21-26 (middle left column) The analyses and interpretation of the statistics presented in this table relating to scaling assumptions, reliability and validity are further described in Appendix 1.

**reliability (Cronbach's alpha), ^ test-retest reproducibility (intraclass correlation), [s]screening, [b]baseline

DISCUSSION

At the scale level, the ADAS-cog met most traditional psychometric criteria in this large dataset of people with mild and mild-to-moderate AD, supporting the findings of previous research.[5 6 12 13] However, a closer examination of the component level findings, a form of analysis rarely undertaken in previous ADAS-cog research,[14] revealed suboptimal scale-to-sample targeting. The key issue here is that we would expect patients in this study to have a range of cognitive abilities. Despite this, over half the ADAS-cog components have substantial percentages of people (often >75%) scoring either 0 or 1, implying few or no problems in cognitive performance. As there is likely to be more clinical heterogeneity in patients' abilities than these components imply, this indicates a targeting problem, or mismatch, between the components' difficulties and patients' abilities in this sample. This is important because the limited component level targeting will impact on the overall ability of the ADAS-cog to detect cognitive differences between people and groups and potentially be less sensitive to the effects of interventions, as reflected in the findings of others. [14]

Our findings demonstrate the importance of targeting rating scales to the individuals within a study sample. Specifically, the range of cognitive performance measured by the ADAS-cog should be well-matched to the range of cognitive performance present in the study sample so that the scale has the ability to detect variability among and within individuals. Poorly targeted scales most likely underestimate changes over time and differences between groups, which is particularly relevant for future AD clinical trials that are tending to recruit people with milder AD. The issue becomes more evident when targeting was examined in AD severity subgroups, demonstrating that the component-level ceiling effects progressively increased as the severity of AD decreased. This underscores the importance of examining component level targeting and demonstrates a misleading aspect of scale-level results.

The problem of targeting could be improved by developing the components of the ADAS-cog so that they span a wider and more appropriate range of measurement. Although component-level floor and ceiling effects will almost always exist to some extent, they should be minimized if the potential of the ADAS-cog to detect change is to be maximized. However, although demonstrating these issues, the information provided by the traditional psychometric analyses used here does not provide specific guidance on how the ADAS-cog items might be improved. Alternative approaches are needed to elaborate upon these findings and propose an evidence-based strategy for restructuring and expanding the existing ADAS-cog components.

Results from this study may have important implications for clinical research. Developments in our understanding of AD have led to attempts to produce treatments aimed at slowing or altering disease progression. Appropriate evaluation of these treatments is dependent on rigorous measurement of clinically meaningful outcomes. Although the ADAS-cog offers clinicians a method of quantifying cognitive performance in people with AD, our findings highlight important limitations. This research emphasizes the importance of fully testing measures before clinicians and researchers apply them in clinical practice and treatment trials. In particular, it highlights the value of the component-level analyses, not typically undertaken, that identified problems with the ADAS-cog that were not detected by standard tests of scale reliability and validity.

Our study has three key limitations. First, the dataset was formed from baseline and screening data from proprietary clinical trial data. It would be valuable to repeat these analyses in non-proprietary data, in other large datasets, to ensure generalisability of our findings to the wider mild to moderate AD population. Second, the current dataset did not allow for analyses of responsiveness to clinical change of cognitive performance over time. Although examinations of responsiveness will be useful to elaborate on and substantiate our present findings, they should not detract from addressing the component level targeting problems identified. Validity testing was also limited. In particular, we were

restricted in the extent to which we could examine aspects of construct validity. Essentially, we were limited to using the MMSE as an external measure; a less detailed, and comprehensive measure of cognitive performance. Thus, further examinations would be beneficial, including head-to-head comparisons with other more comprehensive neuropsychological measures of cognitive performance.

A third limitation is, although the current dominant paradigm for rating scale testing procedures, traditional psychometric analyses have many clinically important limitations, which we have outlined in detail elsewhere.[11,24] In relation to the current study there are two key issues.

First, these methods are sample and scale dependent. This is clearly seen when we compare the performance of the ADAS-cog in terms of scaling assumptions (range of item-total correlations), reliability (alphas, test-retest ICCs), and validity (correlations between the ADAS-cog and MMSE) in the three AD severity subgroups (as described above in the results above and presented in Table 4). These results, if taken at face value, imply that the measurement performance of the ADAS-cog is AD severity dependent. However, the variability in results can be explained by the limited variance of the estimates in each subgroup. This is because traditional psychometric methods are largely based on correlational analyses, and correlations are strongly influenced by variability in the entities correlated. Unfortunately, traditional psychometric methods do not enable us to determine if the differences detected are real (ie scale performance is dependent on AD severity) or simply an artefact of the data distributions. More sophisticated psychometric approaches, for example an analysis of differential item functioning using Rasch analysis, are required to make that distinction.

The second key issue relating to traditional psychometric analyses is that they provide limited information at the component level; particularly about the adequacy of the response options. Importantly, there are concerns over the use of traditional analyses in scales (for discussion see reference 11), such as the ADAS-cog, that combine components with differing number and type of response categories. Therefore, once again, further examinations are required utilising newer

sophisticated rating scale analysis techniques that overcome these limitations, such as Rasch measurement methods,[25,26] to better *diagnose* the specific issues surrounding the performance of the ADAS-cog.[11]

In this study, the ADAS-cog showed the potential to be a scientifically strong measurement instrument. However, our study also suggests that the ADAS-cog has limited ability to detect cognitive performance differences between people, changes over time, and the impact of treatment mild AD. Our analyses of the ADAS-cog by MMSE sub-group (Table 4) indicate that these limitations are more pronounced the milder forms of AD. The natural extrapolation of these findings is that the situation may be more problematic in people with mild cognitive impairment. Thus, in order for this scale to be a valuable cognitive performance measure in these patient groups, these limitations may need to be addressed.

Overall, although the ADAS-cog's psychometric performance was found to be satisfactory, more than half of its components may underestimate differences in cognitive performance in people with mild and moderate AD. The limited distributions indicate widespread targeting issues which may lead to problems in detecting clinical change when it occurs. This has important implications for the inferences of present and future clinical trials of AD using the ADAS-cog. Given the limitations of traditional psychometric methods, further evaluations would be desirable using more sophisticated modern rating scale analysis techniques to pinpoint the specific improvements that are required to maximise the ADAS-cog as a measure of cognitive performance in people with AD.

ACKNOWLEDGMENTS

None

COMPETING INTERESTS

TH and MM are employees of Eisai Medical Research. JS is an employee of Eisai Global Clinical Development. HP is an employee of Pfizer (previously as employee of Eisai Medical Research). SH was retained as a consultant to Eisai Medical Research. JH and SC were supported in part through a grant from Eisai Medical Research.

FUNDING

Eisai Medical Research, Inc.

REFERENCES

1. Blennow K, de Leon M, Zetterberg H. Alzheimer's disease. Lancet 2006;368:387-403.
2. Alzheimer's Association. Alzheimer’s Disease Facts and Figures. Chicago: Alzheimers Association, 2008.
3. Brookmeyer R, Johnson E. Forecasting the global burden of Alzheimer's disease. Alzheimer Dementia 2007;3:186-191.
4. Aisen P, Schafer K, Grundman M, et al. Effects of rofecoxib or naproxen vs placebo on Alzheimer disease progression: a randomized controlled trial. JAMA 2003;289(21):2819-2826.

5. Mohs K, Rosen W, Davis K. The Alzheimer's Disease Assessment Scale: An instrument for assessing treatment efficacy. Psychopharmacol Bull 1983;19:448-450.

6. Rosen W, Mohs R, Davis K. A new rating scale for Alzheimer'sdisease. Am J Psychiatry 1984;141:1356-1364.

7. Farlow M, He Y, Tekin S, et al. Impact of APOE in mild cognitive impairment. Neurology 2004;63:1898-1901.

8. Malouf R, Birks J. Donepezil for vascular cognitive impairment. Cochrane database of systematic reviews. London: St George's Hospital Medical School, 2004.

9. Emre M, Aarsland D, Albanese A, et al. Rivastigmine for dementia associated with Parkinson's disease. N Engl J Med 2004;351:2509-2518.

10. Food and Drug Administration. Patient reported outcome measures: use in medical product development to support labelling claims, 2009 http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM193282.pdf; site accessed 20th December 2009.

11. Hobart J, Cano S, Zajicek J, et al. Rating scales as outcome measures for clinical trials in neurology: problems, solutions, and recommendations. Lancet Neurol 2007;6:1094-1105.

12. Kim Y, Nibbelink D, Overall J. Factor structure and reliability of the Alzheimer's Disease Assessment Scale in a multicenter trial with linopirdine. J Geriatr Psychiatry Neurol 1994;7:74-83.

13. Weyer G, Erzigkeit H, Kanowski S, et al. Alzheimer's Disease Assessment Scale: reliability and validity in a multicenter clinical trial. Int Psychogeriatr 1997;9:123-38.

14. Doraiswamy P, Kaiser L, Bieber F, et al. The Alzheimer's Disease Assessment Scale: Evaluation of Psychometric Properties and Patterns of Cognitive Decline in Multicenter Clinical Trials of Mild to Moderate Alzheimer's Disease. Alzheimer Dis Assoc Disord 2001;15:174–183.

15. Wesnes K, Satek S, Ferguson J, et al. P4-401: Identifying cognitive enhancement in man: Identifying efficacy in the dementias [abstract]. Alzheimer Dementia 2008;4(4): T792.

16. Dichgans M, Markus H, Salloway S, et al. Donepezil in patients with subcortical vascular cognitive impairment: a randomised double-blind trial in CADASIL. Lancet Neurol 2008;7:310-318.

17. Novick MR. The axioms and principal results of classical test theory. J Math Psychol 1966;3:1-18.

18. Rogers S, Farlow M, Doody R, et al. A 24-week, double-blind, placebo-controlled trial of donepezil inpatients with Alzheimer's disease. Neurology 1998;50:136-145.

19. Burns A, Rossor M, Gauthier S, et al. The effects of Donepezil in Alzheimer's disease - Results from a multinational trial. Dementia Geriatr Disord 1999;10:237-244.

20. Seltzer B, Zolnouni P, Nunez M, et al. Efficacy of Dopenzil in early-stage Alzheimer disease: a randmoized-controlled trial. Arch Neurol 2004;61:1852-1856.

21. Hobart JC, Freeman JA, Lamping DL, et al. The SF-36 in multiple sclerosis (MS): why basic assumptions must be tested. J Neurol Neurosurg Psychiatry 2001;71:363-370.

22. Scientific Advisory Committee of the Medical Outcomes Trust. Assessing health status and quality of life instruments: attributes and review criteria. Qual life Res 2002;11:193-205.

23. Cano SJ, Hobart JC, Hart P, et al. The International Co-operative Ataxia Rating Scale (ICARS): An appropriate rating scale for Friedreich's Ataxia? Mov Disord 2005;20:1585-1591.

24. Hobart J, Cano S. Improving the evaluation of therapeutic intervention in MS: the role of new psychometric methods. Health Tech Assess 2009;13:1-200.

25. Andrich D. Rasch models for measurement. Beverley Hills, CA: Sage Publications, 1988.

26. Rasch G. Probabilistic models for some intelligence and attainment tests. Copenhagen Chicago: Danish Institute for Education Research, 1960.

27. McHorney CA, Ware JEJ, Lu JFR, et al. The MOS 36-Item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions and reliability across diverse patient groups. Med Care 1994;32:40-66.

28. World Health Organisation Quality of Life Assessment Group. The World Health Organisation Quality of Life Assessment (WHOQOL): development and general psychometric properties. Soc Sci Med 1998;46:1569-1585.

29. Ware JE, Snow KK, Kosinski M, et al. SF-36 Health Survey manual and interpretation guide. Boston, MA: Nimrod Press, 1993.

30. Hays RD, Hayashi T. Beyond internal consistency reliability: rationale and user's guide for Multi-Trait Analysis Program on the microcomputer. Behav Res Methods Instr Comput 1990;22:167-175.

31. DeVellis RF. Scale development: theory and applications. London: Sage publications, 1991.

32. Guttman LA. Some necessary conditions for common-factor analysis. Psychometrika 1954;19:149-161.

33. Likert RA. A technique for the measurement of attitudes. Arch Psychol 1932;140:5-55.

34. Ware JE, Harris WJ, Gandek B, et al. MAP-R for windows: multitrait / multi-item analysis program - revised user's guide. Boston, MA: Health Assessment Lab., 1997.

35. McHorney CA, Tarlov AR. Individual-patient monitoring in clinical practice: are available health status surveys adequate? Qual Life Res 1995;4:293-307.

36. Hays RD, Anderson R, Revicki DA. Psychometric considerations in evaluating health-related quality of life measures. Qual Life Res 1993;2:441-449.

37. O'Connor RJ, Cano SJ, Thompson AJ, et al. Exploring rating scale responsiveness: does the total score reflect the sum of its parts? Neurology 2004;62:1842-1844.

38. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951;16:297-334.

39. Nunnally JC, Bernstein IH. Psychometric theory. 3rd ed. New York: McGraw-Hill, 1994.

40. Duruoz MT, Poiraudeau S, Fermanian J, et al. Development and validation of a rheumatoid hand functional disability scale that assesses functional handicap. J Rheumatol 1996;23:1167-1172.

41. Kaplan RM, Bush JW, Barry CC. Health status: types of validity and the index of well-being. Health Serv Res 1976;11:478-507.

42. Folstein MF, Folstein SE, McHugh PR. "Mini-Mental State": a practical method for grading the cognitive state of patients for the clinician. Journal of Psychiatric Research 1975;12:189-198.

Appendix 1: Figure - The structure of the ADAS-cog

Figure Legend:

The figure shows the component structure and score ranges of the ADAS-cog.

Appendix 2: Psychometric properties, their definitions and criteria for satisfactory performance

Psychometric property	Definition	Test(s)	Criteria for Acceptability
Data Completeness	The extent to which ADAS-cog components are scored and ADAS-cog total scores can be computed.	Computing the percent of missing data for each component, and the percent of people for whom a scale score can be computed.[27]	• Component-level missing data is <10%[28] • Computable scale scores >50% completed components.[29] • A total score can only be computed if all components are scored as they have substantially different ranges.
Scaling Assumptions	The extent to which it is legitimate to sum a set of component scores, without weighting or standardisation, to produce a single total score.[30 31]	Summing ADAS-cog component socres is considered legitimate, when the components: 1. Are approximately parallell (ie they measure at the same point on the scale). 2. Contribute similarly to the variation of the total score (ie they have similar variances), otherwise these should be standardised. 3. Measure a common underlying construct (ie cognitive performance)[32] otherwise combining them to produce a single score is not appropriate. 4. Contain a similar proportion of information concerning the construct	1. Satisfied when components have similar mean scores.[33] 2. Satisfied when components have similar standard deviations.[27] 3. Satisfied when components have adequate corrected component-total correlation (ITC ≥0.30).[34] 4. Satisfied when components have similar ITCs.[34]

		being measured. Otherwise components should be given different weights.[33]	
Targeting	The extent to which the range of the variable measured by the scale (here cognitive performance) matches the range of that variable in the study sample.	Score distributions were examined at both the ADAS-cog component and scale level. This was conducted in the whole sample and in AD severity subgroups defined by three MMSE ranges: 10-14 (marked); 15-20 (moderate); 21-26 (mild).	Scale scores should span the entire range; floor (proportion of the sample at the maximum scale score for the ADAS-cog) and ceiling (proportion of the sample at the minimum scale score) effects should be low (<15%);[35] and skewness statistics should range from –1 to +1.[36] There is no published criteria for component level targeting. Therefore, we applied the scale-level criteria. This is frequently overlooked but important.[37]
Reliability	Reliability is the extent to which scale scores are associated with random error. High reliability indicates that scores are associated with little random error, i.e. are consistent.	Two types of reliability were examined at both scale and component level. Each quantifies a different source of random error: 1. Internal consistency reliability estimates the random error associated with total scores from the intercorrelations among the components.[38] *2.* Test retest (TRT) reproducibility, based on the agreement between people scores at screening and baseline, estimates the ability of components and scales to produce stable scores.[36]	1. Recommended for adequate scale internal consistency is Cronbach's alpha coefficient ≥0.80,[38] and item internal consistency is item total correlations >0.40. *2.* Recommended for adequate TRT reproducibility are scale-level intraclass correlation coefficients (ICC) ≥0.80[39] and item level ICC ≥0.50.[40]
Validity	The extent to which a	Three aspects of construct validity were	

	scale measures what it intends to measure and is essential for the accurate and meaningful interpretation of scores.[41]	tested: 1. Convergent construct validity was examined by computing correlations between ADAS-cog and the mini mental state examination (MMSE[42]).[a] 2. Discriminant construct validity was examined by computing correlations between the ADAS-cog and sociodemographic variables (age and sex) to determine the extent to which they were biased by these variables. 3. Group difference construct validity was examined by comparing ADAS-cog mean scores for the three MMSE defined groups.	1. We hypothesised that the ADAS-cog and MMSE would be highly negatively correlated (r>-0.70) as the two scales measure cognitive performance but are scored in opposite directions. 2. We predicted ADAS-cog scores should not be notably biased by these variables and, therefore, correlations would be low <0.30. 3. We predicted a stepwise change in ADAS-cog scores across the groups, and that the means scores would be signifincatly different.

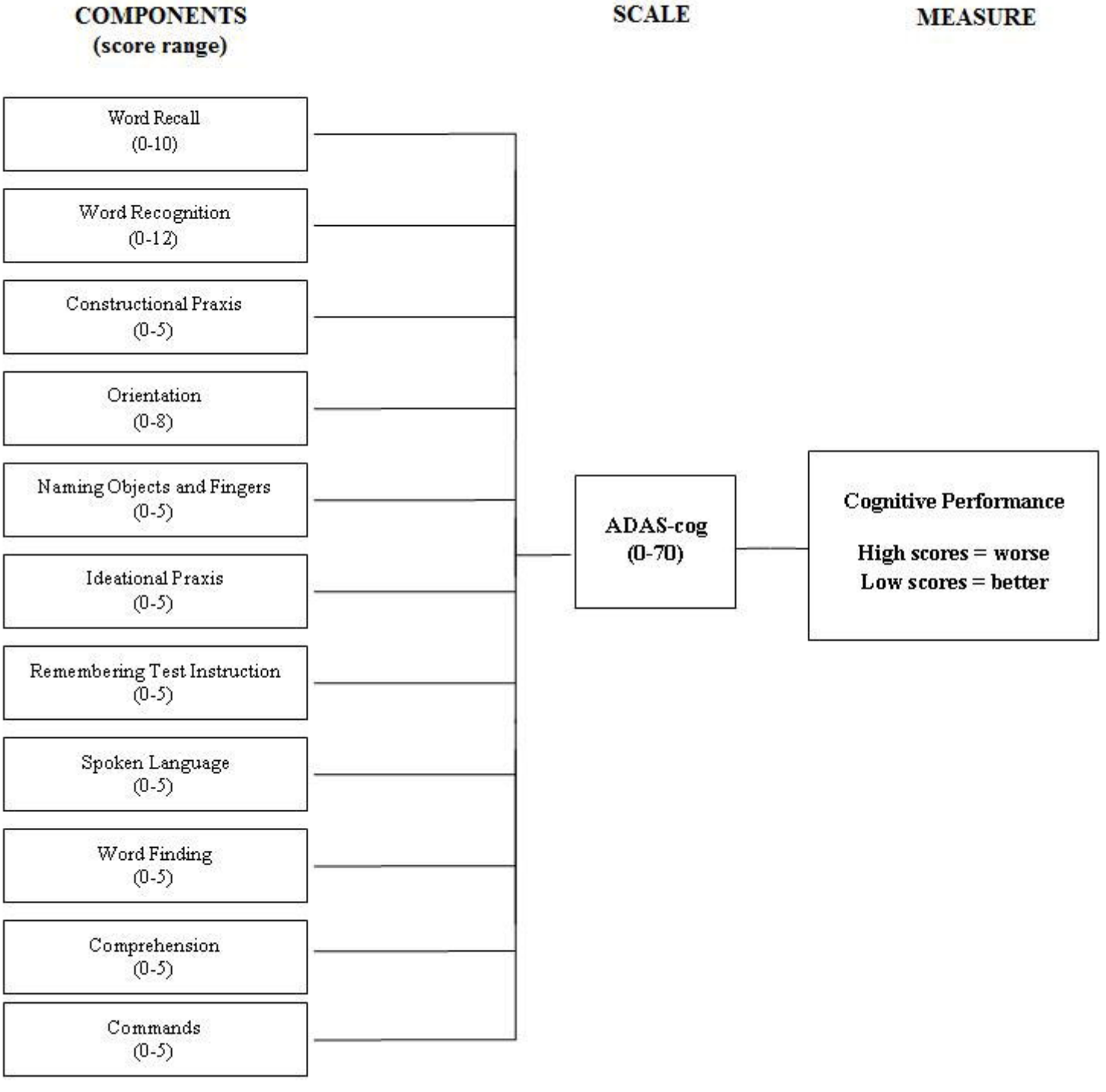
[a] The MMSE is a 30-item rating scale used to assess aspects of cognitive performance (including arithmetic, memory and orientation), and is commonly used in screening for dementia, and also to classify AD as mild (MMSE 21-26), moderate (MMSE 15-20), or servere (MMSE 10-14)

Appendix 3: Component frequency distributions*

Component	Response Options	Screen (N=1421)	Base (N=1421)
Word recall	0	0	0.1
	1	0.1	0.1
	2	0.5	0.6
	3	1.5	1.6
	4	4.6	3.5
	5	10.6	10.6
	6	21.8	19.8
	7	22.2	23.5
	8	21.6	22.9
	9	14.4	14.4
	10	2.9	2.9
Naming objects and fingers	0	42.9	45.9
	1	35.6	34.9
	2	14.4	11.5
	3	4.5	5.2
	4	1.8	1.5
	5	0.8	1.0
Commands	0	52.9	55.9
	1	24.5	22.4
	2	10.8	10.5
	3	9.6	9.3
	4	1.8	1.4
	5	0.5	0.5
Constructional praxis	0	20.2	22.9
	1	46.7	45.6
	2	17.9	15.3
	3	12.2	13.7
	4	2.5	1.8
	5	0.6	0.6
Ideational praxis	0	59.5	61.3
	1	24.8	23.3
	2	7.2	8.0
	3	3.2	3.7
	4	2.7	1.9
	5	2.5	1.8
Orientation	0	11.1	12.4
	1	14.2	14.6
	2	14.1	14.9
	3	14.9	13.0
	4	15.8	16.8
	5	13.2	12.0
	6	10.2	10.4
	7	6.3	5.7
	8	0.2	0.4
Word recognition	0	0.8	0.9
	1	5.1	3.0
	2	7.7	6.0
	3	8.3	10.0
	4	10.5	8.7
	5	11.3	10.1
	6	9.6	9.6
	7	9.4	9.9
	8	9.9	9.9
	9	7.7	9.0
	10	6.8	7.7
	11	5.8	8.2

	12	6.9	7.0
Remembering test instruction	0	42.0	42.8
	1	19.8	19.8
	2	14.6	14.0
	3	10.1	9.7
	4	3.9	3.7
	5	9.5	10.0
Spoken language	0	64.4	64.8
	1	18.6	19.0
	2	9.6	8.7
	3	5.7	5.3
	4	1.5	1.9
	5	0.2	0.3
Word finding	0	40.0	41.2
	1	32.8	31.5
	2	15.8	16.3
	3	8.1	7.2
	4	3.0	3.4
	5	0.2	0.4
Comprehension	0	58.6	58.6
	1	21.5	22.0
	2	13.3	12.9
	3	6.3	5.8
	4	0.4	0.6
	5	0	0.2

*This table presents the ADAS-cog components (column 1) and their response options represented by their scores (eg Word Recall component includes ten words, all words recalled correctly = 0, nine words recalled correctly = 1 ... none of the words recalled = 10; column 2) proportion (percentage) of patients endorsing each of the response options (column 2, as rated by the assessing clinician) for each of the eleven components of the ADAS-cog from the screening (column 3) and baseline (column 4) data.


COMPONENTS
(score range)
SCALE
MEASURE
Word Recall
(0-10)
Word Recognition
(0-12)
Constructional Praxis
(0-5)
Orientation
(0-8)
Naming Objects and Fingers
(0-5)
Ideational Praxis
(0-5)
Remembering Test Instruction
(0-5)
Spoken Language
(0-5)
Word Finding
(0-5)
Comprehension
(0-5)
Commands
(0-5)
ADAS-cog
(0-70)
Cognitive Performance
High scores = worse
Low scores = better