ELSEVIER

Alzheimer's & Dementia 9 (2013) S10–S20

Alzheimer's & Dementia

Putting the Alzheimer's cognitive test to the test II: Rasch Measurement Theory

Jeremy Hobart[a,*,†], Stefan Cano[a,†], Holly Posner[b], Ola Selnes[c], Yaakov Stern[d], Ronald Thomas[e], John Zajicek[a]; for the Alzheimer's Disease Neuroimaging Initiative[‡]

[a]*Clinical Neurology Research Group, Plymouth University Peninsula Schools of Medicine and Dentistry, Plymouth, UK*
[b]*Clinical R&D, Pfizer, Inc., New York, NY, USA*
[c]*Department of Neurology, Johns Hopkins University School of Medicine, Baltimore, MD, USA*
[d]*Cognitive Neuroscience Division, Columbia University, New York, NY, USA*
[e]*Department of Family and Preventative Medicine, University of California, San Diego, CA, USA*

Abstract

Background: The Alzheimer's Disease Assessment Scale—Cognitive Behavior section (ADAS-Cog) is the most widely used measure of cognitive performance in AD clinical trials. This key role has rightly brought its performance under increased scrutiny with recent research using traditional psychometric methods, questioning the ADAS-Cog's ability to adequately measure early-stage disease. However, given the limitations of traditional psychometric approaches, herein we use the more sophisticated Rasch Measurement Theory (RMT) methods to fully examine the strengths and weaknesses of the ADAS-Cog, and identify potential paths toward its improvement.

Methods: We analyzed AD Neuroimaging Initiative (ADNI) ADAS-Cog data (675 measurements across four time-points over 2 years) from the AD participants. RMT analysis was undertaken to examine three broad areas: adequacy of scale-to-sample targeting; degree to which, taken together, the ADAS-Cog items adequately perform as a measuring instrument; and how well the scale measured the subjects in the current sample.

Results: The 11 ADAS-Cog components mapped-out a measurement continuum, worked together adequately, and were stable across different time-points and samples. However, the scale did not prove to be a good match to the patient sample supporting previous research. RMT analysis also identified problematic "gaps" and "bunching" of the components across the continuum.

Conclusion: Although the ADAS-Cog has the building blocks of a good measurement instrument, this sophisticated analysis confirms limitations with potentially serious implications for clinical trials. Importantly, and unlike traditional psychometric methods, our RMT analysis has provided important clues aimed at solving the measurement problems of the ADAS-Cog.

Keywords: Alzheimer's disease; Clinical trials; Psychometrics; Reliability; Validity; Rasch Measurement Theory

1. Introduction

Widespread use of the Alzheimer's Disease Assessment Scale—Cognitive Behavioral section (ADAS-Cog) [1,2] in clinical trials has made it a key outcome measure in crucial decisions about patient care, health policy, and the direction of research. Confidence and evidence that it is fit for this responsibility is clearly critical.

The extent to which cognitive tests and scales, such as the ADAS-Cog, are clinically and scientifically robust is judged using psychometric methods [3]. Traditional

The authors have no conflicts of interest to report.

[†]J.H. and S.C. contributed equally to this article and share first author status.

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

*Corresponding author. Tel: +44-1752-315272; Fax: +44-1752-315254.
E-mail address: Jeremy.Hobart@pms.ac.uk

1552-5260/$ - see front matter © 2013 The Alzheimer's Association. All rights reserved.
http://dx.doi.org/10.1016/j.jalz.2012.08.006

psychometric methods involve the reliability and validity testing understood best by most clinicians [3,4]. Our group [5,6] and others [7–12] have examined the ADAS-Cog using these methods and have found limitations in the use of this scale in mild AD. However, these types of analyses, which are based on classical test theory (CTT), have many clinically relevant limitations, including scale and sample dependency. We have detailed these elsewhere [3,13]. However, modern psychometric methods, which are growing in popularity in clinical rating scale research, have the potential to go further than traditional methods and tell us much more about the performance of rating scales [13].

"Modern" psychometric methods refer to two schools of thought: Rasch Measurement Theory (RMT) [14–18] and Item Response Theory (IRT) [19–22]. The methods from these schools of thought provide highly advanced evaluations of scale performance, based on proven mathematical models [13]. However, despite many similarities, RMT and IRT differ fundamentally.

The aim of an IRT analysis is to find a mathematical model that best explains the rating scale response data. Thus, IRT is a statistical modeling psychometric paradigm. In contrast, the aim of an RMT analysis is to examine the extent to which the observed rating scale data satisfies the requirements of the Rasch model—a mathematical model that articulates the conditions that must be satisfied if measurement is to be achieved from rating scales [14]. Thus, RMT is an experimental psychometric paradigm.

We performed an RMT, rather than IRT analysis of ADAS-Cog data. This is because the goal of the ADAS-Cog, when used as an outcome measure in clinical research and trials, is to measure cognitive performance. Therefore, our aim in this study was to determine the extent to which the ADAS-Cog assumed that role.

2. Methods

2.1. Setting and participants

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu2). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the U.S. Food and Drug Administration (FDA), private pharmaceutical companies, and nonprofit organizations, as a $60 million, 5-year public-private partnership. The primary goal of the ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biologic markers, and clinical and neuropsychologic assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early AD. Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians in developing new treatments and monitoring their effectiveness, as well as to lessen the time and cost of clinical trials.

The principal investigator of this initiative is Michael W. Weiner, MD (VA Medical Center and University of California, San Francisco). The ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the USA and Canada. The initial goal of the ADNI was to recruit 800 adults, 55–90 years of age, to participate in the research, including approximately 200 cognitively normal older individuals to be followed for 3 years, 400 with MCI to be followed for 3 years, and 200 with early AD to be followed for 2 years. For up-to-date information, refer to www.adni-info.org. For the present study, anonymized, longitudinal ADAS-Cog data on AD subjects from the ADNI central database were made available for analysis. The data set was downloaded on April 9, 2008.

2.2. The ADAS-Cog: Structure and scoring

The original ADAS-Cog has 11 components[1] with scores combined to give a single figure that counts as the patient's overall measure of cognitive performance.[2] Different ADAS-Cog components have different numbers of response categories and thus different score ranges, including: 0–8 (orientation); 0–10 (word recall); 0–12 (word recognition); and 0–5 (remaining eight components). All scores are combined to give a total ADAS-Cog score (ranging from 0 [best cognitive performance] to 70 [worst performance]).

2.3. Analysis

Any rating scale can be considered an hypothesis of how a variable might be measured. A number of reasons underpin this statement. First, the aspects of people that rating scales are seeking to measure are complex socially constructed variables, here cognitive performance. As such, their measurement is not simple. Second, there is uncertainty concerning the definitions of these variables. This hampers the construction of rating scales and opens the door for a range of potential measurement methods. Third, socially constructed variables are measured through their manifestations. For example, the cognitive performance of an individual is estimated from their scores on a finite set of tasks. It follows then that the extent to which scores on a set of tasks can be combined is an empirical question. Finally, there is uncertainty of the extent to which the numbers generated by any

[1] We use the term "component" as opposed to the more traditional "item" to reflect the extensive and complex nature of each of the "questions" of the ADAS-Cog.

[2] Word recall, word recognition, constructional praxis, ideational praxis, orientation, naming objects and fingers, commands, remembering test instruction, spoken language ability, word finding difficulties, and comprehension.

rating scales satisfy criteria as reliable and valid measurements. For these reasons, the ADAS-Cog should be viewed as an hypothesis of how cognitive performance might be measured that requires careful testing.

There are three main paradigms for developing, analyzing, and modifying rating scales: CTT; IRT; and RMT. An evaluation of the ADAS-Cog is well suited to the RMT paradigm because the Rasch model, a mathematical equation (model), provides an hypothesis test. This is because it articulates, *a priori*, the requirements of rating scale data for rating scales to satisfy criteria as measurement instruments. The model was derived from theory and is independent of any data set. Therefore, discrepancies detected by the analysis, that is, between the hypothesis (ADAS-Cog data) and the hypothesis test (Rasch model requirements), indicate anomalies. In this way, a Rasch measurement analysis provides diagnostic information informing measurement instrument development by exposing anomalies to be understood and improved empirically.

The information provided by an RMT analysis is both sophisticated and extensive. Information from multiple tests is integrated. These are considered simultaneously and interactively, rather than individually and sequentially. Test result interpretation requires professional judgment, rather than adherence to rigid criteria, because the information needs to be contextualized and most statistical tests depend on sample size. Also, as the analyses compare observed rating scale data against a stringent mathematical model, anomalies are expected. To facilitate interpretation, in this study analyses are grouped under three broad, clinically relevant, simple (but not simplistic) questions [25]: Is the scale to sample targeting adequate for making judgments about the performance of the scale and the measurement of subjects? Has a measurement ruler been constructed successfully? How have the people been measured by the ruler? We used RUMM2030 to conduct the data analysis [26]. Two of the present authors (J.H., S.C.) analyzed the data independently for quality control.

2.3.1. Is the scale-to-sample targeting adequate?

Scale-to-sample targeting concerns the match between the range of cognitive performance measured by the ADAS-Cog, and the range of cognitive performance measured in the study sample. A simple examination of histograms of these two relative distributions provides a frame of reference for interpreting the other results, and informs about the suitability of the sample for evaluating the scale and the suitability of the scale for measuring the sample. Not surprisingly, the better the targeting the better the information.

2.3.2. Has a measurement ruler been constructed successfully?

This question is assessed in five main parts:

2.3.2.1. Do the response categories work as intended?

Each ADAS-Cog component has multiple response categories labeled to imply an ordered continuum of worsening cognitive performance, from less to more. This continuum is implied further by assigning sequential integer scores to the response categories. For example, the response categories for the commands component are scored: 0 = no commands wrong; 1 = one command wrong; 2 = two commands wrong; 3 = three commands wrong; 4 = four commands wrong; and 5 = five commands wrong.

Although this rank ordering is intuitively sound and clinically sensible at the individual component level, it must also work when a component is part of a set. By this we mean that the ADAS-Cog component response categories must have the same logical sequence when a subject moves up and down the variable measured by the whole ADAS-Cog component set (here cognitive performance). For example, as a person's cognitive performance worsens, their scores on all the components should progress sequentially: 0, 1, 2, 3, 4, 5.

RMT analyses test this requirement empirically by estimating the location, on the cognitive performance variable, of the points of transition (thresholds) between adjacent categories. A threshold is the location, on the cognitive performance variable, at which the probability of responding in adjacent categories is 50%. Thus, the commands component has five transition points: 0–1, 1–2, 2–3, 3–4, 4–5. When the categories are working as intended, the thresholds are ordered sequentially along the continuum: 0–1 < 1–2 < 2–3 < 3–4 < 4–5.

When the thresholds are not correctly ordered, that is, they are disordered, the implication is that the response categories for that component are not working as intended. Clinically, for the ADAS-cog, this means that a higher score does not necessarily mean more cognitive impairment, which has major implications for clinical trials. Visually, thresholds are displayed as category probability curves that provide potential diagnostic information.

2.3.2.2. Does the ADAS-Cog map out a continuum?

Before anything can be measured, the variable (or continuum) along which measurements are to be made needs to be marked out [18]. Rating scales, such as the ADAS-Cog, use a set of items (or components) to define the variable they intend to measure. Therefore, for the ADAS-Cog to define a cognitive performance variable along which measures can be interpreted, the components must be located at different points so that the direction and meaning of the variable can be identified. This question is addressed by examining the ADAS-Cog threshold locations,[3] their range, how they are spread, their proximity to each other, and the precision of the estimates (standard error).

2.3.2.3. Do the components work together?

The components of the ADAS-Cog should work together as a conformable set, both clinically and statistically. Otherwise, it makes no sense conceptually, logically, clinically, or empirically to sum component responses to obtain a total

[3]The item location estimate is the mean of all the threshold location estimates for an item.

score and consider using that total score as a measurement of an individual. If the components spread out and work together to define a single continuum then the responses to items should be predictable. Thus, examining the responses to each item for their consistency is important to determine whether the components define a cohesive continuum. More specifically, the responses to components should be in general agreement with the ordering of subjects implied by the majority of components. When this is not the case, the validity of the components and the higher order construct they seek to measure may be questioned.

These ideas are examined formally using indicators of goodness-of-fit of the observed rating scale data to the requirements of the Rasch mathematical model. No one indicator is sufficient to describe fit. We examined two statistical (fit residuals, χ^2 statistics) and one graphical (item characteristic curves, or ICCs) indicator of fit.

2.3.2.4. *Do responses to one ADAS-Cog component bias responses to others?*

The response to one ADAS-Cog component is expected, in general, to be related to another. For example, people who are less cognitively impaired are likely to perform better on all ADAS-Cog components than people who are more cognitively impaired. However, the response to one ADAS-Cog component should not *directly influence* (or be dependent on) the response to another. When this happens, measurement estimates are artificially inflated or deflated (biased), and reliability is artificially elevated. Therefore, it is important to look actively for dependence among ADAS-Cog components. This is done by examining three indicators: correlations among the residuals; fit residuals; and subtest analyses.

A residual is the difference between a person's observed score on an ADAS-Cog component, and their expected value for that component derived from the RMT analysis. Correlations among residuals, derived from the whole sample, reflect the degree of the interrelationships between the residuals of the 11 components. When measurement error shows residuals are randomly distributed, correlations among residuals of components are low (rule of thumb range: -0.30 to $+0.30$). However, when peoples' responses to one component are biased by (dependent on) their responses to another component, the resulting residuals are not randomly distributed and higher correlations among residuals result ($-0.30 < r > +0.30$).

Not surprisingly, residuals also provide a statistical indication of the observed data "fit" to the requirements of the Rasch measurement model. For each component, residuals are combined across individuals and standardized to produce the fit residual summary statistic (see subsection 2.3.2.3). When there is dependency among components, a high score on one component results in an unexpectedly high score on another component. Likewise, a low score on one component results in an unexpectedly low score on another component. When viewed across the range of the measurement continuum, and shown on the item characteristic curve (ICC), this pattern of dependency leads to the curve of observed scores being steeper than the curve of expected scores. This is reflected in the fit residual statistic as a high negative value. As a rule of thumb, fit residual values are recommended to lie in the range -2.50 to $+2.50$, and values lower than -2.50 points to potential dependency. Naturally, as fit residuals are dependent on sample size, they need to be interpreted with this in mind.

In a subtest analysis, potentially dependent components are combined together to form a single component or subtest. This neutralizes the dependency between the components. Dependency is determined by examining the impact of subtesting on the person separation index (PSI), a reliability indicator. The magnitude of the drop in PSI, when the subtest analysis is compared with the "non-subtest" analysis, indicates the extent to which the reliability of the latter is falsely elevated and the degree of dependency between components.

2.3.2.5. *Is performance stable across relevant groups?*

When the ruler mapped-out by the ADAS-Cog's components is stable, the measurements generated by them can be used to make meaningful comparisons. Thus, we need the scale components to perform similarly across relevant groups that we intend to study and compare (e.g., men and women, different age groups, etc.). When item performance is not stable across relevant groups, and displays differential item functioning (DIF), the measurement ruler is not stable across circumstances and measurement is affected to an unknown degree. Herein we have examined the ADAS-Cog for DIF across study and time-point (screening, baseline).

2.3.3. *How have the people been measured?*

We examined three specific questions:

2.3.3.1. *Are people separated by the ADAS-Cog?*

The aim of measurement is to locate people on a continuum and to detect differences between them and changes over time. It is therefore valuable to examine the extent to which a scale can detect differences between people in any study sample. In RMT analyses this is quantified as the PSI, computed as the ratio of error-corrected person variance to the total person variance. In addition, the distribution of person measurements, and percent extremes also provides information on the capacity of the scale to separate the sample. It is important to note that this value is sample-specific.

The PSI of RMT is analogous to a Cronbach's α coefficient in CTT: It is a reliability statistic that can range from 0 to 1, with higher values indicating greater separation of the persons in this specific sample by this specific scale. Values do not generalize directly from sample to sample. Although CTT posits recommended values for α, this is somewhat misleading as it is a finding about the data. However, the PSI has implications for the power of the tests of fit. The greater the separation index the greater the power of the tests of fit to detect fit.

2.3.3.2. How valid are people's measurement?

When a person is measured using the ADAS-Cog it is important to know that the scale has been used in the expected way; that is, consistent with the idea that the items map out a variable along which the items have a unique order. This can be determined by examining the extent to which the responses for an individual person are in general agreement with the ordering of components implied by the majority of persons. If not, the validity of that person's measurement is questionable. This is determined by examining the person fit residual, which is analogous to the item fit residual.

2.3.3.3. What is the implication of ADAS-Cog raw scores?

It is useful to understand the extent to which the raw summed ADAS-Cog total scores, which are, by definition, ordinal in nature (have unequal intervals), are equivalent (or close to) their implied measurements, which by definition are linear in nature (have equal intervals). Typically, clinical trials of AD have analyzed raw ADAS-Cog scores. If raw ADAS-Cog *scores* provide good approximations of ADAS-Cog *measurements*, then there are fewer concerns related to treating one as the other. RMT analyses estimate linear measurements from the ordinal raw scores, and the plot of the linear measurements implied by each raw score can be examined. It is necessary to consider the extent to which the data fit the Rasch model when interpreting this plot. The less the data fit the model, the less confident the estimates of the linear measurements.

3. Results

3.1. Sample characteristics

At the time we accessed the ADNI data set there were a total of 675 measurements from people with AD at five time-points: 0, 6, 12, 18, and 24 months. The mean age of this group was 74 (range 53–90) years, 47% of whom were women. The mean Mini-Mental State Examination (MMSE) score for the group was 23 (SD 8) across all time-points.

3.2. RMT findings

3.2.1. Is the sample-to-scale targeting adequate?

Figure 1 shows the sample-to-scale targeting, based on ADAS-Cog component location estimates. The figure shows suboptimal targeting as the range of cognitive performance measured in the sample (upper histogram; range −3.2 to +0.9 logits; mean −1.6 logits) was not well matched to the range of performance measured by the ADAS-Cog components (lower histograms; range −2.7 to +2.2, mean 0.0).

3.2.2. Has a measurement ruler been constructed successfully?

3.2.2.1. Did the response categories work as intended?

The thresholds for six ADAS-Cog components (commands, constructional praxis, naming objects and fingers, ideational praxis, remembering test instructions, spoken language) were not ordered sequentially. This means that, in this data set, the response categories did not work as intended. Clinically, this means that, for 6 of the 11 ADAS-Cog components, a higher score did not confirm more cognitive impairment. This has substantial implications for interpreting clinical trial and longitudinal monitoring data.

To demonstrate and explain this critically important issue further, Figure 2 shows two response category probability curves (CPCs). Figure 2A shows the CPC for word recall, one of the five components for an ADAS-Cog component in which the response categories work as intended. Figure 2B shows the CPC for naming objects and fingers,

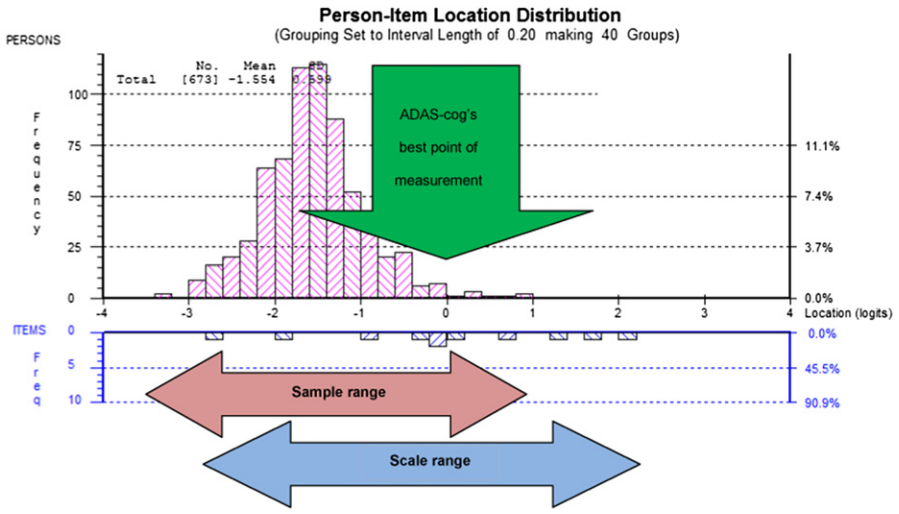

Fig. 1. Targeting of components to person distributions of the 11 components of ADAS-Cog. The suboptimal targeting between the distribution of person measurements (upper pink histogram) and the distribution of item locations (lower blue histogram) is presented. A rating scale measures at its best in the middle of its range of item locations (as indicated). In the RMT analysis, the mean of the item locations is set to 0 unit (logits). The mean of the person locations is −1.554 logits, and very few in the sample are around the 0-logit area.

one of the six components in which the response categories do not work as intended.

For CPCs the x-axis is the continuum of cognitive performance represented by the ADAS-Cog as a whole (the combination of the 11 components), which goes from better performance (left) to worse performance on (right). The y-axis is the probability of responding to each of the response categories (the curves numbered 1, 2, 3, 4…). Logically, as individuals become more cognitively impaired, they move from left to right along the x-axis. In doing so we would expect their probability of responding to the response categories to move through the sequence 1–2–3–4….

Figure 2A shows the CPC that, for word recall, this is the case. However, Figure 2B shows that, with naming objects and fingers, this is not the case. At approximately +0.25 logit on the x-axis there is a "mesh" of curves for five response categories. It seems that the most likely that a change of score is from 1 to 5. At no point on the continuum is a response to categories 2, 3, or 4 the most likely.

The implication of this finding is that the scoring function for the naming objects and fingers component (and the other five which have similar plots) is not working as intended when considered within the context of the ADAS-Cog component set of 11. This means that a higher score does not

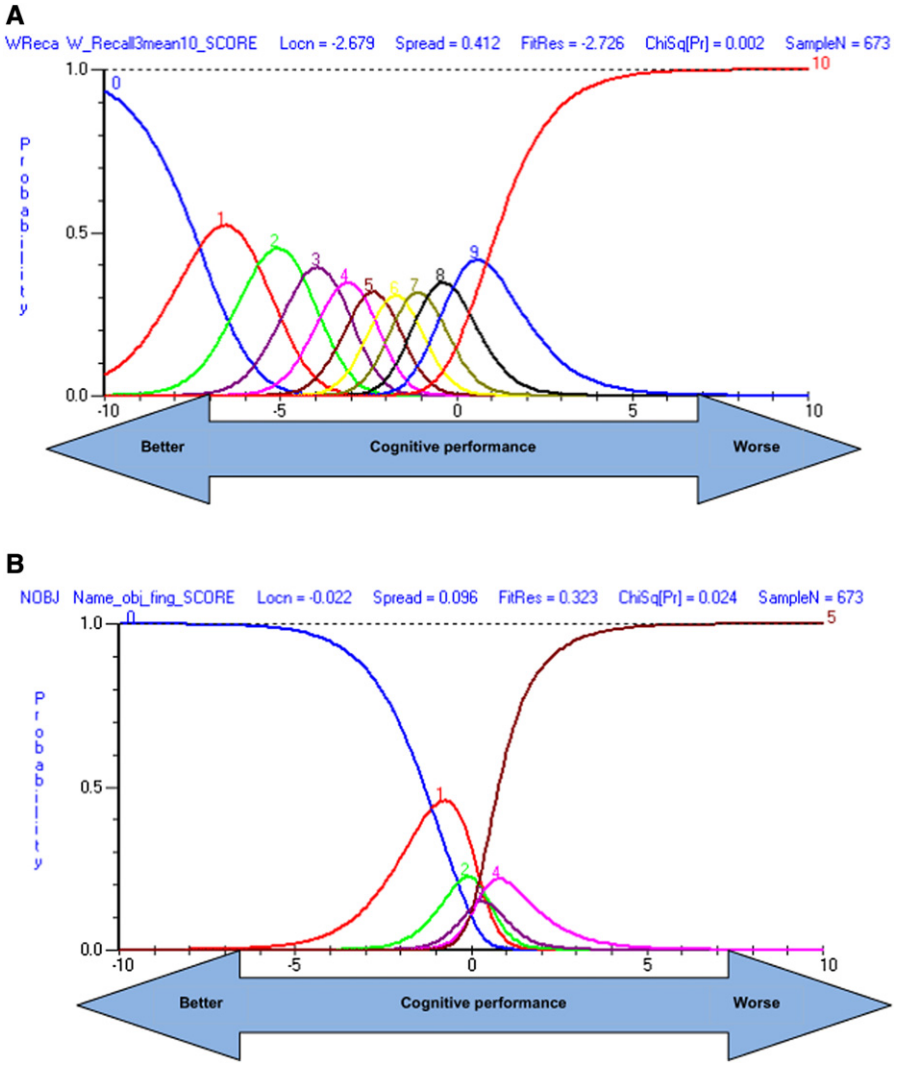

Fig. 2. Two category probability curves (CPCs) are presented: (A) is for word recall, a component in which the response options *do* work as intended; (B) is for naming objects and fingers, a component in which the response options *do not* work as intended. CPCs show the probability (y-axis) of being scored in each response category (colored lines), for each level of cognitive performance measured by the ADAS-Cog as a whole (x-axis). In (A) it can also be seen that an individual's cognitive performance worsens, and in moving along the x-axis from left to right, the probability of scoring 0–10 follows the intended sequence (0–1–2–3–4…10). This means that any score on this component has an interpretation in reference to the cognitive performance metric (x-axis). For example, a person who scores 3 on the word recall component has a cognitive performance measurement of between −4.339 and −3.361 logits. That is, the categories are working as intended. In contrast, (B) shows that as an individual's cognitive performance worsens, and in moving along the x-axis from left to right, the probability of scoring 0–5 does not follow the intended sequence. Specifically, the response categories for 2 (green), 3 (purple), and 4 (pink) curves are *never* the most likely categories to be scored at any level of cognitive performance. As such, there is no clear interpretation of scores 2, 3, or 4 for the naming objects and fingers component. Thus, the sequential ordering of the six categories is not working as intended for naming object and fingers. In other words, the response scale for that component is not valid.

mean more cognitive impairment. More specifically, we cannot interpret scores for those who scored 2, 3, or 4 on this component. This finding points to a measurement problem that requires investigation and rectification.

3.2.2.2. Does the ADAS-Cog map out a continuum?

Figure 1 shows that the ADAS-Cog components spread out to map out a continuum, rather than define a point on a line. The location estimate of each ADAS-Cog component in Figure 1 is the mean of multiple thresholds for that ADAS-Cog component. The gaps in this continuum imply areas where measurement precision is limited.

3.2.2.3. Do the components work together?

Table 1 shows the values for the two statistical indicators of fit (fit residuals and χ^2). Bold values indicate misfitting ADAS-Cog components—ADAS-Cog components for which the predicted scores differ from the observed scores by more than statistical reason. For the fit residuals, seven ADAS-Cog components lie within the "rule-of-thumb" range −2.5 to +2.5, three ADAS-Cog components lie just outside that range (word recall, remembering test instructions, comprehension), and one ADAS-Cog components lies well outside the range (word recognition). For the χ^2 values, nine ADAS-Cog components have similar values, with one ADAS-Cog component (word recognition) lying well away from the pack.

Figure 3A and B shows the ICC for the two ADAS-Cog components with the best (naming objects and finger) and worst (word recognition) statistical results of fit. As the black dots (observed scores) closely map the S-shaped lines (values predicted by the Rasch model), these graphic indicators of fit imply good fit to the model despite the statistical values. These findings support the 11 ADAS-Cog components as a statistically conformable set.

3.2.2.4. Do responses to one component bias responses to others?

Correlations among residuals ranged from −0.321 to +0.328, and only 3 of 55 correlations exceeded the rule-of-thumb recommended range (−0.30 to +0.30). This finding implies no notable relationships among the residuals and therefore no notable dependency among ADAS-Cog components.

Three components had negative residuals exceeding, marginally, the recommended value of −2.50: word recall (−2.73); remembering test instructions (−2.56); and comprehension (−2.59). The sample is quite large (n = 673), and the ICCs for these items do not demonstrate overdiscrimination. This finding implies no notable dependency among ADAS-Cog components. On the basis of these findings no subtest analyses were undertaken.

3.2.2.5. Is performance stable across relevant groups?

Table 2 shows the results for DIF by time-points. One ADAS-Cog component, word recognition, showed DIF by time-point (baseline, 6, 12, and 24 months).

3.2.3. How have the people been measured?

3.2.3.1. How are people separated by the ADAS-Cog?

The PSI was 0.77, indicating that the persons in this sample were reasonably separated by the ADAS-Cog, and providing evidence that the tests of fit are reasonably able to detect misfit if present. Person measurements were spread over a wide range of cognitive performance (>4 logits).

3.2.3.2. How valid are people's measurements?

Fit residuals were within the "rule-of-thumb" range of −2.5 to + 2.5 for 99% of people. A total of 8 people had fit residuals >+2.5 (maximum +4.2; data available from authors).

3.2.3.3. What is the implication of using ADAS-Cog raw scores?

Figure 4 shows the relationship between ADAS-Cog raw (ordinal) scores and the linear measurements they imply. The relationship is S-shaped, meaning that the change (or difference) in ADAS-Cog *measurement* implied by a 1-point change (or difference) in ADAS-Cog *score* varies across the range of the scale. It is highest at the extremes, and lowest at the center of the scale range. For example, a 10-point change in ADAS-Cog total score from 1 to 10 implies a change of 5.8 logits, whereas a change in ADAS-Cog total score from 25 to 35 implies a change of 0.7 logits. Thus, the implications of changes at the extremes are much (eightfold) greater than extremes than toward the center of the scale.

Table 1
ADAS-Cog component fit statistics and χ^2 probability ordered by item location

Item	Location	SE	Fit residual	*df*	χ^2	*P*-value
Word recall	−2.679	0.031	**−2.726**	607.8	24.01	.004
Word recognition	−1.905	0.018	**4.056**	606.0	74.12	.000
Orientation	−0.839	0.026	0.490	607.8	13.24	.152
Remembering test instruction	−0.219	0.049	**−2.559**	606.0	13.07	.160
Ideational praxis	−0.138	0.053	−0.376	606.9	24.94	.003
Naming objects and fingers	−0.022	0.054	0.323	607.8	19.05	.025
Word finding	0.079	0.044	1.078	606.9	11.69	.231
Comprehension	0.675	0.056	**−2.586**	606.9	27.60	.001
Spoken language	1.201	0.061	−0.426	606.9	5.137	.822
Commands	1.662	0.057	0.408	607.8	18.76	.027
Constructional praxis	2.184	0.056	1.075	607.8	9.336	.407

NOTE. Bold values indicate misfitting AnAS-Cog components

4. Discussion

The aim of this study was to evaluate the ADAS-Cog in a large sample of people with AD using a sophisticated method of rating scale analysis. The analyses, which build on our previous evaluation using traditional psychometric methods [5,23], further demonstrate and detail both the strengths and weaknesses of the ADAS-Cog as a cognitive measure. Most importantly, they provide the vehicle for its evidence-based improvement as an instrument of measurement.

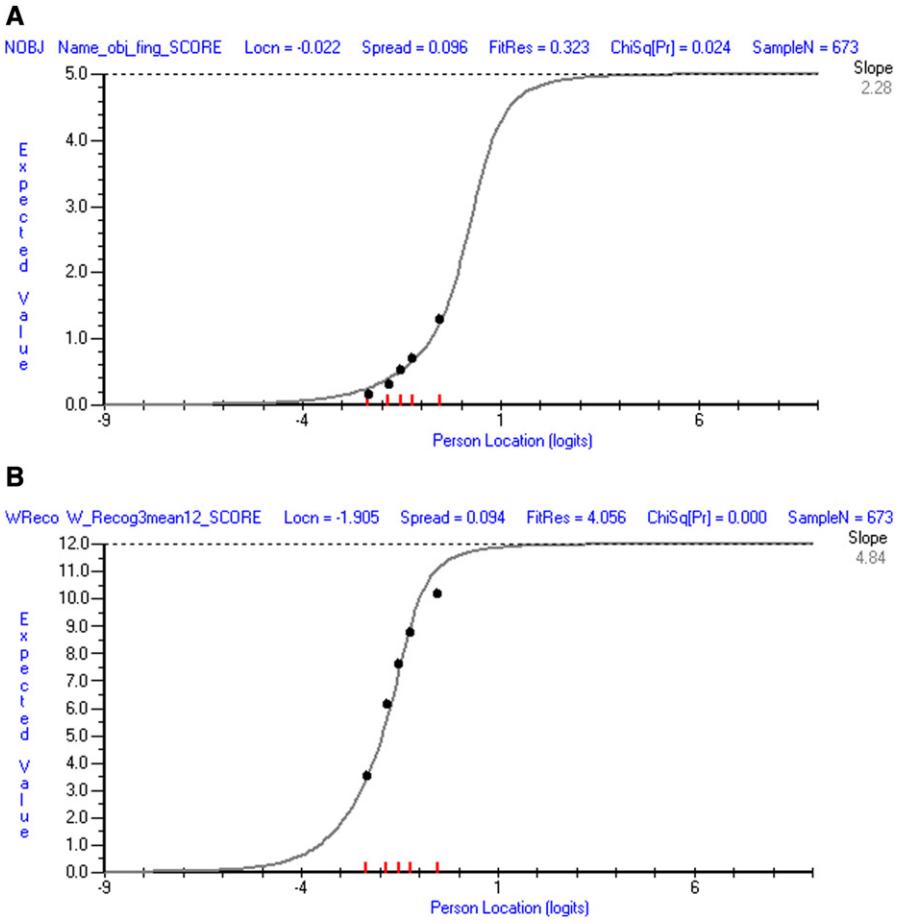

Fig. 3. (A) Item characteristic curves (ICC) for the naming objects and fingers component. The ICC is a graph for an individual component. It plots the expected response (predicted from the model) to a component at each and every level of the measurement continuum. (B) Item characteristic curves for the word recognition component. (A) and (B) also show the ICCs for the ADAS-Cog components with the best (naming objects and finger) and worst (word recognition).

The application of sophisticated methods of rating scale evaluation is important for multiple reasons. Particularly pertinent, but rarely discussed, is the fact that any rating scale (here the ADAS-Cog) is merely an hypothesis of how a complex variable (here cognitive performance) might be measured. The process of developing scales to measure complex variables is complicated by uncertainty about the nature of the variable to be measured (i.e., what is cognitive performance?), and uncertainty about the most valid method of measurement (i.e., how best to articulate, capture, and combine the important elements of cognitive performance).

By necessity, the construction of high-quality rating scales that generate the rigorous measurements required by state-of-the-art clinical trials is a circular iterative process of hypothesis generation, testing, and revision. This process needs methods that can identify measurement problems and guide how they are solved. Traditional psychometric methods, with their many scientific limitations, do not achieve these goals.

Our evaluation of the ADAS-Cog has demonstrated three key important strengths. The implication of these findings is that the ADAS-Cog has the foundation for a valuable, scientifically rigorous instrument for measuring cognitive performance in clinical studies of AD.

First, the 11 components of the ADAS-Cog map out a variable on which cognitive performance can be measured. This is important because it is a requirement that a scale maps out a continuum, rather than defines a point on a line. Second,

Table 2
ADAS-Cog component statistical significance of differential item functioning (DIF) by time based on analysis of variance (ANOVA)

Commands	Time *F*	*P*
Word recall	1.133	.340
Commands	0.124	.974
Constructional praxis	1.094	.359
Naming objects and fingers	0.734	.569
Ideational praxis	0.909	.458
Orientation	1.139	.337
Word recognition	0.945	.438
Remembering test instruction	2.888	.022
Comprehension	0.423	.792
Word finding	1.043	.384
Spoken language	0.931	.445

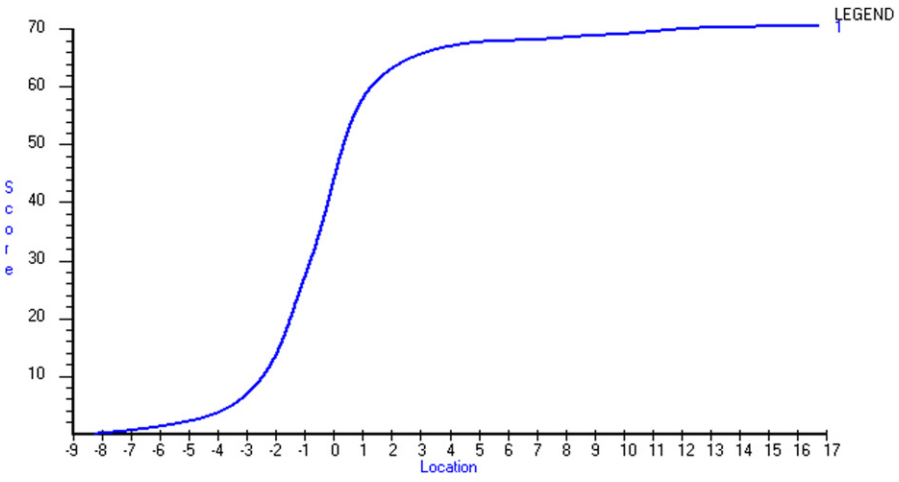

Fig. 4. ADAS-Cog raw score to interval level across the sample. The figure shows the relationship between ADAS-Cog total scores (which are ordinal and therefore have an unequal interval), and the linear measurement they imply (which are equal intervals) is S-shaped. The change in cognitive performance implied by a change of 1 point in ADAS-Cog total score varies eightfold across the subscale range.

the 11 components of the ADAS-Cog work reasonably well together to define a conformable set. This is important because the processes of using an ADAS-Cog total score, achieved by summing the 11 ADAS-Cog component scores, and seeking to measure overall cognitive performance rather than a specific domain of cognition, requires that summation is both clinically meaningful and statistically sound. Third, the performance of the ADAS-Cog was stable across timepoints and studies. This is important because "ruler" stability is required for measuring people over time and across different clinical situations.

In spite the strengths just described, our evaluation of the ADAS-Cog has also demonstrated three key important issues that undermine measurement rigor. First, the match between the range of cognitive performance measured by the ADAS-Cog and the range of cognitive performance measured in this sample is suboptimal. This means that measurements of individuals are associated with large standard errors, and the scale will be less able to discriminate accurately between people in terms of their cognitive performance. This extrapolates to a limited ability to detect change over time as AD progresses, and in association with treatments.

The second issue of the ADAS-Cog we identified is that the response categories for 6 of the 11 core components did not work as intended when considered within the frame of reference of the ADAS-Cog components as a set. The means that the proposed integer scoring of these ADAS-Cog components is not, or only weakly, supported. Further study is required to investigate, understand, and correct this shortcoming.

A third issue of the ADAS-Cog concerns the extent to which the 11 components work together to define a single variable. We indicated previously that these hang together reasonably well, and that the graphical indicator (ICCs) implies better fit than the statistical estimates. However, this highlights the fact that there is no simple binary answer to the interpretation of fit indicators, and that their interpretation depends on the purpose for measurement and the nature of the variable under consideration. The findings are, in fact, in keeping with expectations. Cognitive performance is a broad variable and, as such, we expect this to be reflected in the fit statistics. However, the findings should raise two questions. First, should cognitive performance be measured as two or more subvariables? Second, what is the explicit definition of cognitive performance? It is notable that, like many scales, the ADAS-Cog was not constructed on the basis of an explicit definition of cognitive performance. This hinders scale construction and validation [3].

The findings from this study imply the ADAS-Cog may have reached a crossroads in its history: either it is modified (improved) or replaced. If the choice is to modify, then the current RMT analysis acts as a vehicle for evidence-based scale improvement by implying three main changes are required. First, a number of the components should be made more difficult. This can be achieved either by adding more parts to the components or replacing existing components with more difficult versions of the same. Second, the scoring function of six components requires attention. The first stage would be to investigate and determine the reasons why the existing scoring functions for these components were not working as intended. This understanding will undoubtedly lead to options for resolution. The third modification requires a careful consideration of the definition of cognitive performance, which is discussed further in what follows. Although it may be argued that others have addressed some of these issues by modifying existing components and adding new components (e.g., delayed word recall, number cancellation test, maze), our provisional evaluations have shown that these have not achieved the desired degrees of improvement in measurement performance [23,24]. This may be because existing modifications of the ADAS-Cog were not evidence-based from a sophisticated psychometric evaluation.

One of this article's reviewers requested that we expand upon the issue of whether cognitive performance should be measured as two or more components. The answer depends

on the purpose for measurement. For example, there will be circumstances in which cognitive performance, in its broad sense, will be the most appropriate variable for measurement, such as a study of interventions that have the potential to affect multiple cognitive domains. There will be other circumstances in which subcomponents of cognitive performance are the most appropriate variables for measurement. For example, if an intervention is targeted at improving language function, then it is most appropriate to measure its impact on language (or even specific language subcomponents) rather than the impact on the broader cognitive performance variable. Without knowing the specifics of the purpose for measurement it is impossible to answer this question more completely.

The answer to how cognition might be measured (one or more than one subcomponent) best depends on the explicit definition of cognitive performance. This is because we cannot determine the trade-offs associated with measuring cognitive performance as a single variable, or using two or more subcomponents, until we have a consensus definition of cognitive performance, an agreement of the components that should be included in a cognitive performance measure, and empirical evidence that the instruments used to quantify the individual components satisfy stringent criteria as measurement instruments.

This study has two main limitations. First, the data were not taken from randomized clinical trials. It would therefore be important to replicate the analyses in other data sets. Second, we did not study responsiveness. However, although these analyses are important to undertake they cannot overcome the implications of the targeting limitations of the ADAS-Cog to those people with mild AD (and, by implication, MCI).

In this study we have taken a very specific approach to the problems of the ADAS-Cog identified by RMT analyses. There are, however, other approaches to dealing with the problems exposed in the analyses and achieving better data-to-model fit. One approach is to change the mathematical model. Another approach is to change the data *post hoc* to get better fit to the model. The problem with both of these approaches, which have been used in other evaluations of the ADAS-Cog using new psychometric methods [25], is that they do not address the primary issues—the reasons why discrepancies occur. The apparent measurement problems are solved by manipulation. In contrast, the approach we have taken, which follows the approach to rating scale evaluation proposed by Rasch [14] and developed further by Andrich [16,26], and the role of measurement in science articulated by Kuhn [27], means that we are expecting to identify limitations and aiming to understand why they occur and solving them empirically and experimentally.

Acknowledgments

This study was supported by grants from an anonymous foundation (to J.H.). Some of J.H., S.C., and J.Z.'s research time was funded by the National Institute for Health Research (NIHR) under its Programme Grants for Applied Research Programme (Grant Reference Number RP-PG-0707-10124). The views expressed in this article are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health. The authors thank Teresa Driscoll for editorial assistance with the manuscript.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI; National Institutes of Health Grant U01 AG024904). The ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from: Abbott Laboratories; the Alzheimer's Association; the Alzheimer's Drug Discovery Foundation; Amorfix Life Sciences, Ltd.; AstraZeneca; Bayer HealthCare; BioClinica, Inc.; Biogen Idec, Inc.; Bristol-Myers Squibb Co.; Eisai, Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Co.; F. Hoffmann-La Roche, Ltd., and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; Janssen Alzheimer Immunotherapy Research & Development, LLC; Johnson & Johnson Pharmaceutical Research & Development LLC; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC; Novartis Pharmaceuticals Corporation; Pfizer, Inc.; Servier; Synarc, Inc.; and Takeda Pharmaceutical Co. 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 Neuroimaging at the University of California, Los Angeles. This research was also supported by the National Institutes of Health (P30 AG010129 and K01 AG030514).

References

[1] Mohs K, Rosen W, Davis K. The Alzheimer's Disease Assessment Scale: An instrument for assessing treatment efficacy. Psychopharm Bull 1983;19:448–50.

[2] Rosen W, Mohs R, Davis K. A new rating scale for Alzheimer's disease. Am J Psychiatry 1984;141:1356–64.

[3] Hobart J, Cano S, Zajicek J, Thompson A. Rating scales as outcome measures for clinical trials in neurology: Problems, solutions, and recommendations. Lancet Neurol 2007;6:1094–105.

[4] Hobart J, Lamping D, Thompson A. Evaluating neurological outcome measures: The bare essentials. J Neurol Neurosurg Psychiatry 1996; 60:127–30.

[5] Cano S, Posner H, Moline M, Hurt S, Swartz J, Hsu T, Hobart J. The ADAS-Cog in Alzheimer's disease clinical trials: Psychometric evaluation of the sum and its parts. J Neurol Neurosurg Psychiatry 2010;81:1363–8.

[6] Hobart J, Cano S, Posner H, Selnes O, Stern Y, Thomas R, Zajicek J. Putting the Alzheimer's cognitive test to the test I: Traditional psychometric methods. Alzheimer Dementia.

[7] Weyer G, Erzigkeit H, Kanowski S, Ihl R, Hadler D. Alzheimer's Disease Assessment Scale: reliability and validity in a multicenter clinical trial. Int Psychogeriatr 1997;9:123–38.

[8] 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.
[9] Doraiswamy P, Kaiser L, Bieber F, Garman R. 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;15:174–83.
[10] Ihl R, Frolich L, Dierks T, Martin E, Maurer K. Differential validity of psychometric tests in dementia of the Alzheimer type. Psychiatry Res 1992;44:93–106.
[11] Kincaid M, Harvey P, Parrella M, et al. Validity and utility of the ADAS-L for measurement of cognitive and functional impairment in geriatric schizophrenic inpatients. J Neuropsychiatry Clin Neurosci 1995;7:76–81.
[12] Talwalker S, Overall J, Srirama M, Gracon S. Cardinal features of cognitive dysfunction in Alzheimer's disease: A factor-analytic study of the Alzheimer's Disease Assessment Scale. J Geriatr Psychiatry Neurol 1996;9:39–46.
[13] Hobart J, Cano S. Improving the evaluation of therapeutic intervention in MS: the role of new psychometric methods. Health Technol Assess 2009;13:1–200.
[14] Rasch G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Education Research. Expanded edition 1980. Chicago: University of Chicago Press; 1980 (reprinted 1993, available from www.rasch.org/books.htm).
[15] Wright B. Solving measurement problems with the Rasch model. J Educ Measurement 1977;14:97–116.
[16] Andrich D. A rating formulation for ordered response categories. Psychometrika 1978;43:561–73.
[17] Wright B, Stone M. Best test design: Rasch measurement. Chicago: Mesa Press; 1979.
[18] Wright B, Masters G. Rating scale analysis: Rasch measurement. Chicago: Mesa Press; 1982.
[19] Lord F. A theory of test scores. Psychometric Monographs 1952;7.
[20] Lord F. The relation of the reliability of multiple-choice tests to the distribution of item difficulties. Psychometrika 1952; 17:181–94.
[21] Lord F, Novick M. Statistical theories of mental test scores. Reading, MA: Addison-Wesley; 1968.
[22] Hambleton R, Swaminathan H. Item response theory: Principles and applications. Boston, MA: Kluwer-Nijhoff; 1985.
[23] Hobart J, Posner H, Aisen P, Selnes O, Stern Y, Thomas R, Weiner M, Zajicek J, Zeger S, Cano S. The ADAS-cog's performance as a measure—lessons from the ADNI study. Part 3—do the scale modifications add value? Neurology 2009;72:A92.
[24] Cano S, Posner H, Moline M, Hurt, S, Swartz J, Hsu T, Hobart J. The Alzheimer's Disease Assessment Scalea—Cognitive Behavior Section (ADAS-Cog) as an outcome measure for clinical trials of AD: Do existing modifications for mild cognitive impairment, mild AD and vascular dementia improve its measurement: Alzheimer's Association International Conference on Alzheimer's Disease, 2008, Chicago, IL.
[25] Wouters H, van Gool W, Schmand B, Lindeboom R. Revising the ADAS-Cog for a more accurate assessment of cognitive impairment. Alzheimer Dis Assoc Disord 2008;22:236–44.
[26] Andrich D. Rating scales and Rasch measurement. Expert Rev Pharmacoecon Outcomes Res 2011;11:571–85.
[27] Kuhn TS. The function of measurement in modern physical science. Isis 1961;52:161–90.