STATISTICS IN MEDICINE
*Statist. Med.* 2000; **19**:1631–1644

# Application of item response theory for development of a global functioning measure of dementia with linear measurement properties

Dan Mungas[1,2,*,†] and Bruce R. Reed[1,2]

[1] *UC Davis Alzheimer's Disease Center, Department of Neurology, School of Medicine, University of California, Davis, U.S.A.*
[2] *Veterans Administration, Northern California Health Care System*

## SUMMARY

An ideal measure of global functioning for patients with dementia would discriminate at very high and very low levels of functioning and would have linear measurement properties such that a given change in score corresponds to the same amount of change in underlying ability at any part of the ability continuum. Using item response theory methods, linearity of test measurement can be directly assessed and items can be selected to construct a test with desired measurement characteristics. The purpose of this study was to apply item response theory methods to evaluating and developing global functioning scales. Subjects were 1207 patients who had received comprehensive dementia evaluations. Items were selected from two measures of cognitive functioning (Mini Mental State Examination, MMS; Blessed Information Memory Concentration Test, BIMCT) and one measure of independent functioning (Blessed–Roth Dementia Rating Scale, BRDRS). The MMS and BIMCT showed significant non-linearity of measurement, especially at low and high ability levels. A brief composite measure was created by selecting from the three instruments 25 items that fit a uniform distribution of item difficulty across the entire range of ability measured by the three instruments. This composite measure and the BRDRS showed better linearity of measurement than the other two instruments. Results have implications for development of a psychometrically sophisticated, brief measure of global functioning for clinical and research use in dementia. Copyright © 2000 John Wiley & Sons, Ltd.

## INTRODUCTION

Dementia is a disorder defined by loss of cognitive ability that significantly impairs the patient's capacity for independent functioning. While different patterns of impairment of cognitive abilities can be observed in dementia, currently used diagnostic criteria require that more than one cognitive

*Correspondence to: Dan Mungas, UC Davis Department of Neurology, 4860 Y Street, Suite 3700, Sacramento, CA 95817, U.S.A.
† E-mail: dmmungas@ucdavis.edu

Contract/grant sponsor: NIA; contract/grant number: AG10220, AG10129, AG12435;
Contract/grant sponsor: California Department of Health Services Alzheimer's Disease Program; contract/grant number: 94-20354

domain be affected [1], and frequently, generalized deficits across multiple cognitive domains are observed. The requirement of functional impairment is also a standard part of currently used diagnostic criteria for dementia [1, 2].

A number of tests have been created to assess global functioning in dementia. These measures generally can be administered in a short period of time and provide valuable information about the patient's overall level of functioning. The Mini Mental State Examination (MMS) [3] is undoubtedly the most widely used instrument in both clinical practice and research settings. It has been used frequently as a screening measure to determine the presence of dementia. Diagnostic sensitivity and specificity for dementia in the 80 per cent to 90 per cent range have been demonstrated for population-based [4] and clinical [5] samples, though lower values, especially for sensitivity, have been shown with other populations (For example, References [6] and [7]). A second, important use of global functioning measures is to quantify progression of dementia symptoms over time. This serves important clinical functions including objective documentation of the course of the dementia and pre–post assessment to evaluate effects of interventions, drug treatments for example. Quantitation of course also is important for research and is a prerequisite for studies on the rate and predictors of cognitive decline.

While global functioning measures frequently have been used both for initial diagnosis and for quantitation of course, these are independent functions, and a measure might be good for one but be of limited use for the other. For example, memory measures are highly sensitive indicators of dementia in Alzheimer's disease [8–10], even in very mildly impaired patients. However, high-sensitivity memory tests are less useful as measures of change due to the presence of a floor effect in which performance drops to a near-zero level at a relatively early stage of progression of the dementia [10]. Similarly, measures that might be good for assessing progression through levels of functioning may have limited diagnostic sensitivity for mild cases of dementia.

An ideal measure to quantify the progression of dementia should have several characteristics. First, it should provide sensitive discrimination at levels of ability spanning a broad range of functioning associated with dementia. It should be applicable for mildly impaired, high-functioning patients, and for severely impaired patients. Second, it should broadly measure abilities relevant to dementia. Third, its practicality will be greatly increased if it can be administered in a brief period of time. Finally, it should have interval level [11], linear measurement properties such that a given degree of change in score in one part of the ability continuum represents the same amount of change in ability as the same score change in another part of the ability continuum.

Item response theory (IRT) and its associated statistical methods [12–15] play a very prominent role in modern psychometric test development. While use of IRT has been limited in neuropsychology and in dementia, it has important uses in these areas. One specific potential application of IRT methods is for developing and evaluating measures of global functioning for use in quantifying dementia progression. IRT methods have been used to examine measurement characteristics of the MMS [16, 17], but otherwise have had limited use in this context.

A basic concept in IRT is the item characteristic curve (ICC), which is essentially a non-linear regression on ability of probability of a correct response to a given item. IRT theory and methods also are applicable at the test or scale level. The test characteristic curve (TCC) represents a non-linear regression of overall test score on ability. The TCC can be a very useful tool for evaluating the range of measurement and the degree of discrimination at different points of the ability continuum. In addition, the degree to which the TCC is linear provides an indication of the extent to which the measure provides interval scale or linear measurement.

 

The purpose of this study was to use IRT methodology to evaluate three commonly used global function measures for dementia and to develop a new measure based upon items from these scales. Like the MMS, the Blessed Information Memory Concentration Test (BIMCT) and the Blessed–Roth Dementia Rating Scale (BRDRS) [18] have been widely used for assessment of dementia. IRT methods were applied to results from these three instruments to evaluate measurement properties across the ability continuum relevant to dementia, and IRT methods were used to construct and evaluate a new scale composed of items from these tests.

## METHODS

### *Subjects*

Study participants were patients who received a comprehensive multidisciplinary evaluation through the UC Davis Alzheimer's Disease Center (UCD-ADC) between June 1989 and August 1997. The UCD-ADC is a university dementia programme with two clinical sites, one located in Sacramento and the other in Martinez (Berkeley prior to July 1997). Both sites are California Alzheimer's Disease Diagnostic and Treatment Centers and are clinical sites for a National Institute on Aging Alzheimer's Disease Center. The two sites use a uniform clinical evaluation protocol that includes a common neuropsychological test battery. Patients are referred from a variety of professional and non-professional sources for evaluation of cognitive impairment associated with ageing. The standard clinical evaluation is conducted by a multidisciplinary team and includes a thorough medical and social history, neuropsychological evaluation, psychological/psychiatric evaluation, physical and neurological examination, routine dementia work-up laboratory tests and a brain imaging study. The MMS, BIMCT and BRDRS are administered for all patients, while a more extensive neuropsychological test battery is administered to patients with MMS scores of 15 or greater and to lower scoring patients when there is a specific clinical indication.

Subsequent to completion of all evaluation procedures each patient is presented and discussed in a multidisciplinary case conference. Laboratory test results and neuroimaging films are reviewed. A consensus diagnosis is established which consists of a syndrome diagnosis (demented, amnestic syndrome, age-associated memory impairment, other cognitive impairment not meeting criteria for dementia, delirium, no cognitive impairment, and deferred) and, when significant cognitive impairment is present, an aetiologic diagnosis (for example, Alzheimer's disease, ischaemic vascular disease). DSM-III R criteria are used for diagnosing dementia [2]. NINCDS/ADRDA diagnostic criteria [19] are used for diagnosing possible and probable Alzheimer's disease.

A total of 1207 individuals for whom complete data was available for the MMS, BIMCT and BRDRS were included in this study. Demographic characteristics of this subject sample, frequency of diagnoses, and descriptive statistics for these three global functioning measures are presented in Table I. This sample included a very broad range of patients. While most were demented, 27 individuals determined to have no cognitive impairment were included. There were 30 individuals who had MMS scores of 30 and 13 with MMS scores of 0; there were 16 with BIMCT scores (total correct) of 33 and 14 with scores of 0, and there were 34 with BRDRS scores of 0.0 and 5 with scores of 17.0.

### *Variables*

The MMS and BIMCT were collected using the California ADDTC Combined Blessed–Folstein Mental Status Assessment. This integrated instrument contains items of the MMS and BIMCT and

 

Table I. Demographic, cognitive and functional characteristics of sample.

| | | | |
|---|---|---|---|
| Age: | | Gender; $n$(%) | |
| mean | 76.0 | female | 780 (64.7) |
| SD | 8.9 | male | 426 (35.3) |
| range | 39–100 | Diagnostic syndrome $n$(%): | |
| Education: | | dementia | 1072 (88.8) |
| mean | 12.3 | other cognitive impairment | 81 (6.7) |
| SD | 3.8 | no cognitive impairment | 27 (2.2) |
| range | 0–24 | diagnosis deferred | 27 (2.2) |
| MMS: | | Aetiology; $n$(%): | |
| mean | 17.7 | possible AD | 176 (15.2) |
| SD | 7.3 | probable AD | 592 (51.2) |
| range | 0–30 | possible + probable IVD | 60 (5.2) |
| BIMCT: | | mixed dementia | 107 (9.3) |
| mean | 16.9 | other dementia | 221 (19.1) |
| SD | 8.3 | Ethnicity, $n$(%): | |
| range | 0–33 | White | 924 (76.6) |
| BRDRS: | | Black | 166 (13.8) |
| mean | 5.6 | Hispanic | 66 (5.5) |
| SD | 3.8 | Asian | 22 (1.8) |
| range | 0–17 | other | 29 (2.4) |

includes items unique to each individual test as well as items shared by the two. Separate MMS and BIMCT scores can be derived. This instrument was administered to the patient by a neuropsychologist, neuropsychology trainee, or a trained psychometrist. The BRDRS was completed based upon an interview with an external informant who was an identified caregiver for the patient and accompanied the patient to the evaluation appointment. A social worker or nurse completed this instrument.

Data was recoded prior to data analysis to meet requirements of the IRT methods used and to provide clarity of interpretation. All items were coded as dichotomous variables for which a score of 0 indicated failure and a score of 1 indicated that the item was passed. This is the standard format for the MMS. The BIMCT total score is often reported as the number of items failed, whereas the coding used in this study corresponds to the number of items passed. The BRDRS is composed of 11 items. Eight, generally representing instrumental activities of daily living, are scored as 0 for intact ability, 0.5 for mildly impaired ability, and 1 for impaired ability. Three items assessing more basic activities of daily living are scored $0, 1, 2$ or 3 with higher numbers corresponding to increasing degree of impairment.

Specific coding procedures were as follows. Dichotomous MMS and BIMCT items were simply coded as 1 if the item was passed and as 0 if it was failed. Ordinal scale items (for example, MMS 'world backward', BIMCT 'count forwards') were converted to a number of dichotomous items equal to the maximum score on that item. (A graded response IRT model [20] could also be used for non-dichotomous items). For example, MMS 'world backward' was converted to five dichotomous items: world1; world2; world3; world4; world5. World1 was scored 1 if 'world backward' was 1 or higher and was scored 0 otherwise. World3 was scored as 1 if 'world backward' was 3 or higher, 0 otherwise. The other three 'world' variables were scored similarly. Thus, a 'world backward' score of 3 would correspond to 1's for world1, world2 and world3 and 0's for world4 and world5. BRDRS items were coded similarly. The eight items with values of 0, 0.5, and 1 were each coded into two dichotomous items. One item was scored as 1 if no impairment was present and 0 otherwise, the second was scored as 1 if no or mild impairment was present and 0 otherwise. The three items with a range of 0 to 3 were coded into three dichotomous variables. One was scored as 1 if no impairment was present, 0 otherwise. A second was scored as 1 if no or mild impairment was present, 0 otherwise. A third was scored 1 if no, mild or moderate impairment was present and 0 otherwise. Coding in this manner yielded potential maximum scores of 30 for the MMS, 33 for the BIMCT and 25 for the BRDRS, with high scores corresponding to relatively intact functioning and lower scores corresponding to increasing degrees of impairment.

### *Statistical methods*

IRT methods [12–15] were used to evaluate item characteristic curves (ICCs) and test characteristic curves (TCCs) for scales used in this study. IRT analysis is generally applied to an $n \times k$ data matrix consisting of the dichotomous responses of n participants to $k$ items. Iterative maximum likelihood computational methods are used to arrive at a joint solution which provides an ability estimate for each of the $n$ participants and item parameter estimates for each of the $k$ items. The item parameters define the ICC, with any given point on the curve representing the probability of passing the item given the corresponding ability value. A two-parameter model was used in this study, which generated item parameters representing item difficulty and item discrimination. Item difficulty refers to the point on the ability scale that corresponds to a probability of passing the item of 0.50. Item discrimination corresponds to the slope of the ICC at this point.

The TCC can be directly derived from the ICCs of the items comprising the test or scale. The TCC relates ability to overall test performance, which can be scaled as either total correct score or proportion of correct items to total items. The TCC value at any given point of the ability scale is essentially the sum of the ICC probabilities of passing for all items corresponding to that ability value (for a TCC with total score as the $y$-axis; the average of ICCs when proportion correct is used for the $y$-axis).

Data analysis for this study was accomplished using XCALIBRE [21], a marginal maximal likelihood IRT application for dichotomous data. Data from all three instruments were initially combined into a data matrix consisting of $n = 1207$ rows and $k = 82$ columns. The 82 columns corresponded to 82 individual items: 24 unique MMS items; 27 unique BIMCT items; 6 combined MMS/BIMCT items; and 25 BRDRS items. Item and ability parameters were estimated from this combined data set. These parameters were then used to derive TCCs for specific scales.

## RESULTS

The TCC for the MMS (30 items) is presented in Figure 1 and the TCC for the BIMCT (33 items) is presented in Figure 2. These curves relate estimated ability to total number of items correct for each scale. The ability metric essentially refers to standard deviation units for the overall subject sample. For both of these measures, non-linearity is apparent at the two ends of the ability distribution. This indicates decreased discrimination for low and high functioning individuals, which provides evidence of significant floor and ceiling effects. Change in total score is less for a specified change in ability at the two ends of ability distribution than it is in the

 

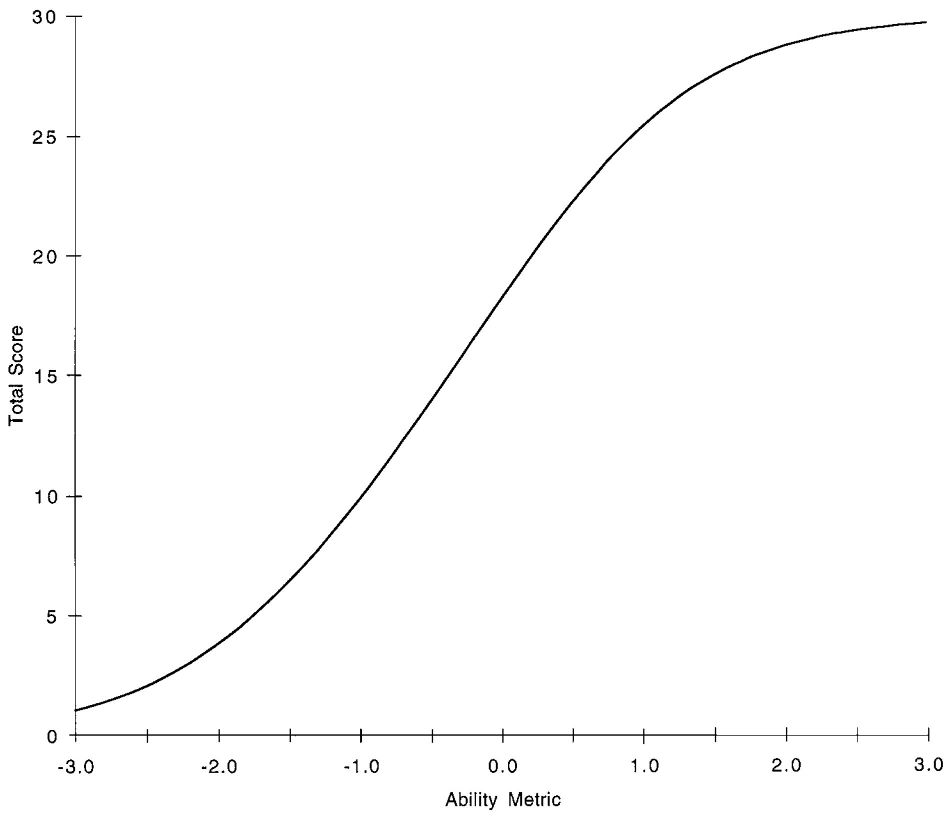


Figure 1. Test characteristic curve (TCC) for the Mini Mental State Examination (MMS). The TCC relates predicted number of total correct items ($n = 30$) to ability.

middle of the ability distribution. For example, a two standard deviation change in ability from 3.0 to 1.0 corresponds to approximately a 5 point MMS loss, whereas a two standard deviation change from 1.0 to $-1.0$ corresponds to a 15 point MMS loss. A similar pattern is apparent for the BIMCT.

The TCC for the BRDRS is presented in Figure 3. The total score on the y-axis has been transformed to the standard 0–17 range for the BRDRS (the axis is inverted; a score of 17 indicates completely intact functioning, whereas a score of 17 in standard usage indicates severe impairment). The BRDRS shows greater overall linearity of measurement across the ability continuum than the MMS or BIMCT.

A new scale, labelled 'Global Fcn', was created from items of the MMS, BIMCT and BRDRS in an attempt to optimize linear measurement across the ability continuum. The principle guiding creation of this scale was to select 25 items spanning a uniform distribution of item difficulty over the $-3.0$ to 3.0 ability range. A broad range of item difficulty was desirable because this tends to result in a linear TCC over the range of ability spanned by the included items [22, 23]. The 25 items that provided the best approximation to this model and their corresponding item parameter values are shown in Table II. Figure 4 shows distribution of item difficulty. The goal of a uniform

 

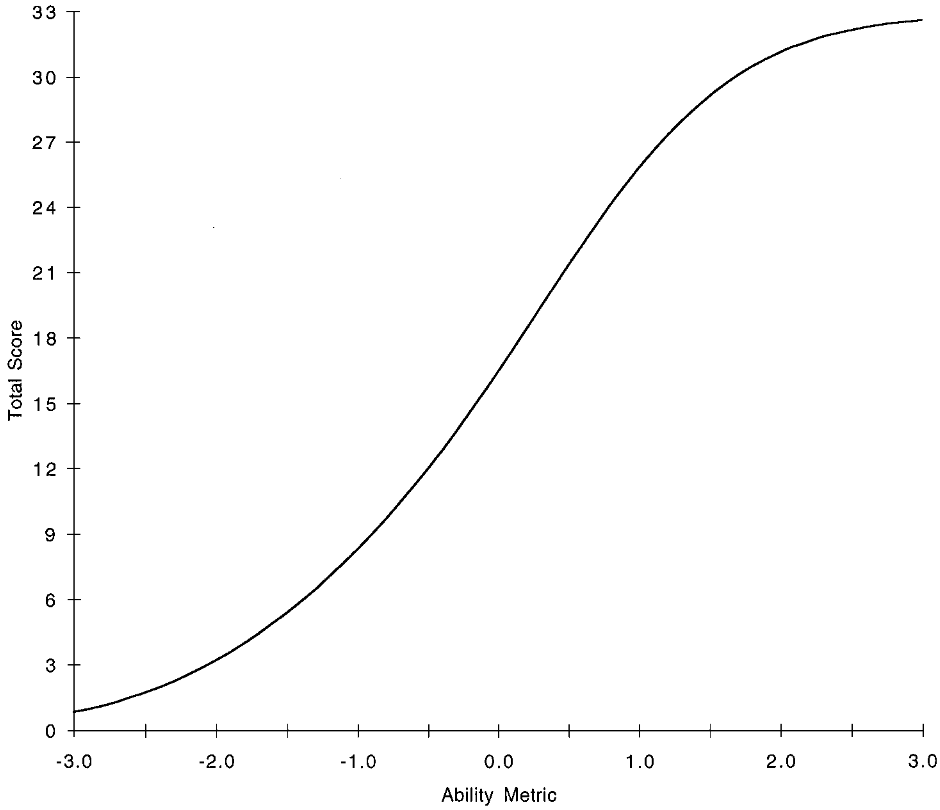


Figure 2. Test characteristic curve (TCC) for the Blessed Information Memory Concentration Test (BIMCT). The TCC relates predicted number of total correct items ($n = 33$) to ability.

distribution of difficulty was achieved for most of the ability continuum, but there were few items available representing the 1.5 to 3.0 range. The TCC for this scale is presented in Figure 5. It shows relatively linear measurement up to an ability level of 1.50, with increased non-linearity above 1.50. The decreased discrimination above the 1.50 ability level is a reflection of the relative absence of items with difficulty greater than 1.50.

TCCs from the MMS, BIMCT BRDRS and Global Fcn scales are presented in Figure 6, in which the $y$-axis for all scales has been transformed to proportion of correct items. The BRDRS and Global Fcn scales have very similar measurement properties and show greater linearity of measurement than the other two scales. All scales show decreased discrimination at high ability levels, and the MMS and BIMCT show decreased discrimination at low ability levels. All scales show linear measurement at mid-range ability levels.

Coefficient alpha values for these four scales were MMS $-0.90$, BIMCT $-0.92$, BRDRS $-0.88$, Global Fcn $-0.87$. Intercorrelations of these four scales are presented in Table III. Global Fcn showed strong correlation with all three of the other scales. It is noteworthy that Global Fcn was more highly correlated with MMS and BIMCT than was BRDRS.

 

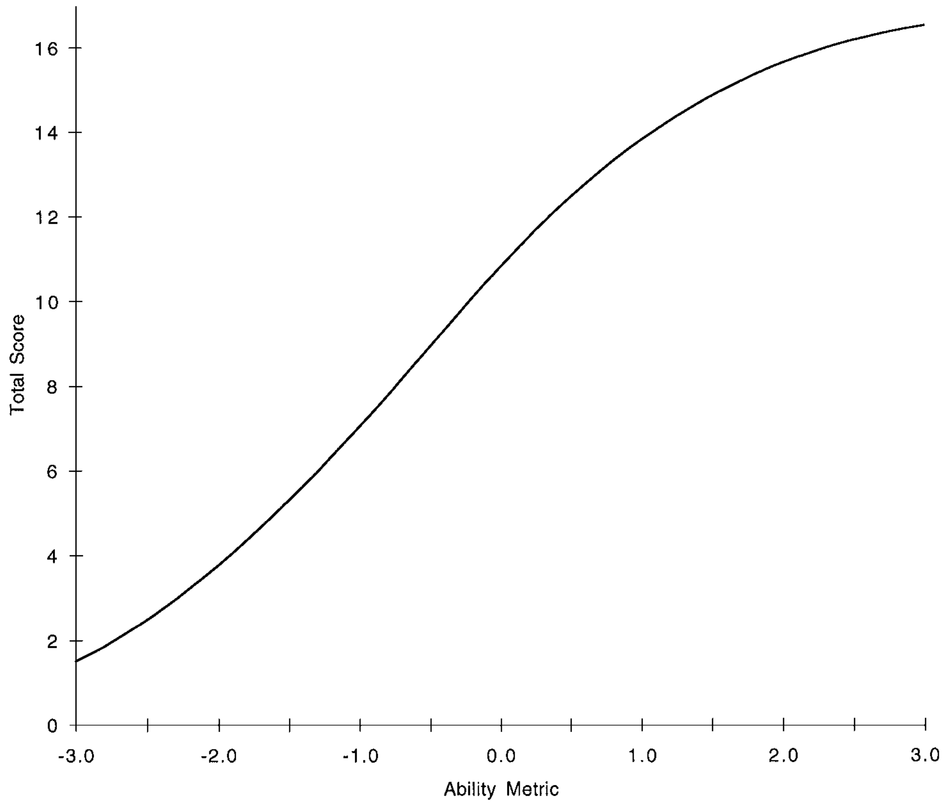


Figure 3. Test characteristic curve (TCC) for the Blessed–Roth Dementia Rating Scale (BRDRS). The TCC relates predicted total score (17.0 = no significant impairment) to ability.

## DISCUSSION

This study provides a demonstration of the applicability of item response theory (IRT) to assessment of global functioning in patients with dementia. The test characteristic curve (TCC) derived from IRT provides a direct method for assessing the measurement characteristics of a given scale across the ability continuum. The TCCs for the MMS and BIMCT were similar and distinctly non-linear. Significant ceiling and floor effects were observed for both instruments, such that test scores are relatively insensitive to changes in ability at the low and high ends of the ability continuum. The BRDRS and Global Fcn Scale appear to have better overall measurement properties in comparison with the MMS and BIMCT. Specifically, their range of linear measurement is greater and low-end discrimination is better. Non-linear high-end measurement is still present to some degree, though not to the extent of that observed for the MMS and BIMCT.

The IRT-based concept of ability is central to measurement issues addressed by this study. In IRT, ability (or a preferably, perhaps, proficiency level) is considered an unknown parameter ($\theta$) that determines each subject's item and test performance. Ability for a given subject is independent of the specific items used and of the abilities of other subjects in the sample [12, 14]. Since ability

 

Table II. Global Fcn items and item parameters.

| Item | Test | Discrimination | Difficulty | Standardized residual |
|---|---|---|---|---|
| Name | BIMCT | 1.16 | −2.39 | 0.44 |
| Age | BIMCT | 1.10 | −0.02 | 0.75 |
| Place of birth | BIMCT | 1.20 | −1.97 | 0.71 |
| Type of work | BIMCT | 1.25 | −1.39 | 0.45 |
| Current year | BIMCT, MMS | 1.34 | 0.15 | 0.57 |
| Current month | BIMCT, MMS | 1.26 | 0.07 | 0.34 |
| State | MMS | 1.27 | −1.20 | 0.53 |
| Count forward (error, corrected) | BIMCT | 0.98 | −1.81 | 0.98 |
| Recall 'John' | BIMCT | 1.32 | 0.63 | 0.61 |
| Recall 'Brown' | BIMCT | 1.33 | 0.53 | 0.52 |
| Recall '42' | BIMCT | 1.41 | 1.24 | 0.43 |
| Recall 'Market Street' | BIMCT | 1.44 | 1.24 | 0.53 |
| Recall 'Chicago' | BIMCT | 1.43 | 0.62 | 0.46 |
| Watch | MMS | 1.18 | −1.83 | 0.82 |
| Pencil | MMS | 1.29 | −2.10 | 0.50 |
| Find way, familiar streets (normal or some trouble) | BRDRS | 1.10 | −0.44 | 0.92 |
| Perform household tasks (normal) | BRDRS | 0.95 | 1.20 | 1.23 |
| Perform household tasks (normal or some trouble) | BRDRS | 1.01 | −0.66 | 0.97 |
| Handle money (normal or some trouble) | BRDRS | 1.24 | −0.23 | 1.15 |
| Remember short lists (normal) | BRDRS | 0.99 | 2.06 | 0.80 |
| Recall recent events (normal) | BRDRS | 0.84 | 1.70 | 0.74 |
| Eating (moderate problem or less) | BRDRS | 1.23 | −3.00 | 0.67 |
| Dressing (mild problem or less) | BRDRS | 1.13 | −0.91 | 0.84 |
| Dressing (moderate problem or less) | BRDRS | 1.27 | −1.81 | 0.44 |
| Incontinence (moderate problem or less) | BRDRS | 0.78 | −2.24 | 1.33 |

Items passed (at the level indicated) were scored 1, and otherwise were scored 0. Standardized residual is an index of goodness-of-fit distributes as standard normal. Values in excess of 2.0 roughly correspond to a significance test ($p = 0.05$) for lack of fit of the model for the item.

Table III. Intercorrelations of measures of global functioning.

| | BIMCT* | MMS | BRDRS |
|---|---|---|---|
| MMS | 0.91 | | |
| BRDRS | −0.61 | −0.63 | |
| Global Fcn | 0.90 | 0.86 | −0.79 |

*Total correct score.

cannot be directly measured, it must be estimated as part of the IRT analysis. IRT applications like XCALIBRE use iterative numerical methods to jointly estimate $\theta$ and item parameters. The TCC depicts the relationship between $\theta$ and the subject's 'true score' on a test composed of the items included in the scale of interest. The metric on which $\theta$ is measured is not unique, and the distribution of the derived estimates of $\theta$ will be determined by the inherent variability in the subject sample [12, 14]. While it is often assumed that $\theta$ is measured on an interval-level scale, interval-level measurement is not an inherent property of $\theta$ [14].

The IRT derivation of ability has important implications for the methods demonstrated in this study. First, the scale and range of the ability metric in this study corresponds to the range of

 

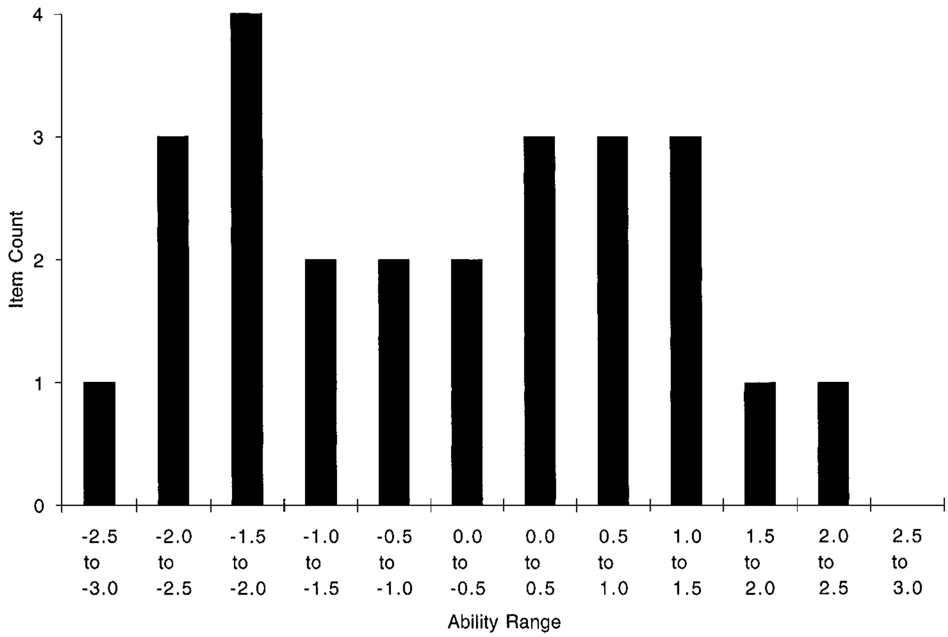


Figure 4. Distribution of item difficulty of 25 items selected for the Global Fcn scale. Item difficulty refers to the value on the validity continuum that corresponds to a probability of 0.50 of a correct response to the item.

variability in global functioning in the subject sample, and would not necessarily be the same in a different sample. In this study, the ability metric is derived from a cognitively diverse sample of individuals seeking evaluation for perceived cognitive deficits. Global functioning in this sample ranged from severe dementia to normal cognitive functioning established by a comprehensive dementia evaluation that included formal neuropsychological testing and interview with an external informant.

Second, ability, as defined in this study, simply refers to proficiency in responding to test items and does not specify what factors determine this proficiency. Premorbid ability, demographic variables like age and education, specific dementing illnesses like Alzheimer's disease and ischaemic vascular disease, and other factors including depression and medication effects are all potential determinants of ability. An advantage of a highly general measure of ability is that it allows for the same metric to be used for assessment in very diverse situations. Thus, for example, the same measure could be used to assess global functioning changes associated with depression and dementia, and relative strengths of the two effects could be directly compared.

The IRT definition of ability also has important implications regarding the concept of linearity of measurement. As demonstrated in this study, the TCC can be a very useful tool for examining linearity of measurement. However, $\theta$ is not necessarily measured on an interval-level scale. Consequently, the TCC should not be interpreted in isolation, and other data is needed to support hypotheses generated from the TCC. In general, empirical construct validity studies are required to demonstrate that the ability being measured, as well as the scale score used to estimate that ability,

 

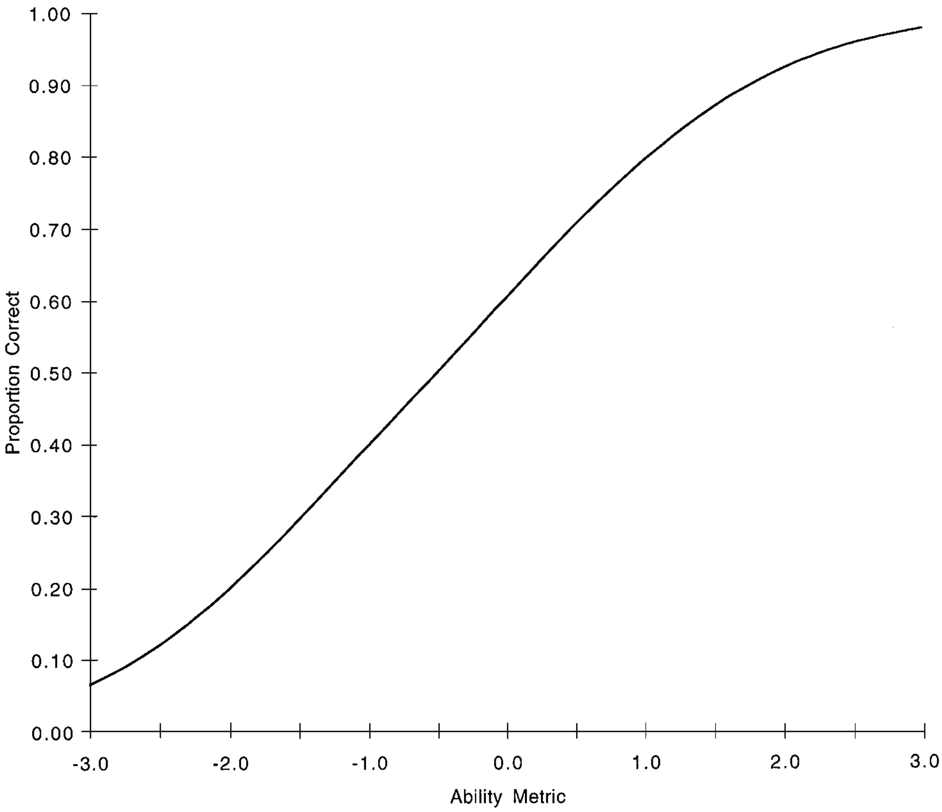


Figure 5. Test characteristic curve (TCC) for the Global Fcn. The TCC relates predicted number of total correct items ($n = 25$) to ability.

behave according to theoretical predictions [12]. Linearity of measurement is one specific aspect of construct validity that would need to be demonstrated with hypothesis-driven empirical studies. A linear TCC would be one piece of evidence supportive of linear measurement, but other characteristics should also be present. For example, a linear scale used to quantify impairment in Alzheimer's disease should show relative impairment in comparison with estimated premorbid ability and an average annual rate of decline which are independent of baseline score after controlling for duration of symptoms. Further, one might predict that longitudinal decline would be linearly related to longitudinal decline on other global functioning measures with seemingly linear measurement.

There is convergent evidence from this study and from other literature to support the conclusion from the TCC of non-linear measurement of the MMS. The IRT analysis of the MMS clearly demonstrated a ceiling effect, with reduced high ability discrimination. The presence of a ceiling effect is further exemplified by the fact that 150 of the 1072 demented patients in this study (10.3 per cent) had MMS raw scores of 25 or greater, and 78 (7.3 per cent) had scores of 27 or greater. Clearly, demented patients can obtain high scores on the MMS, and conversely, the MMS does not adequately detect clinically significant decline in ability in some high functioning individuals.

 

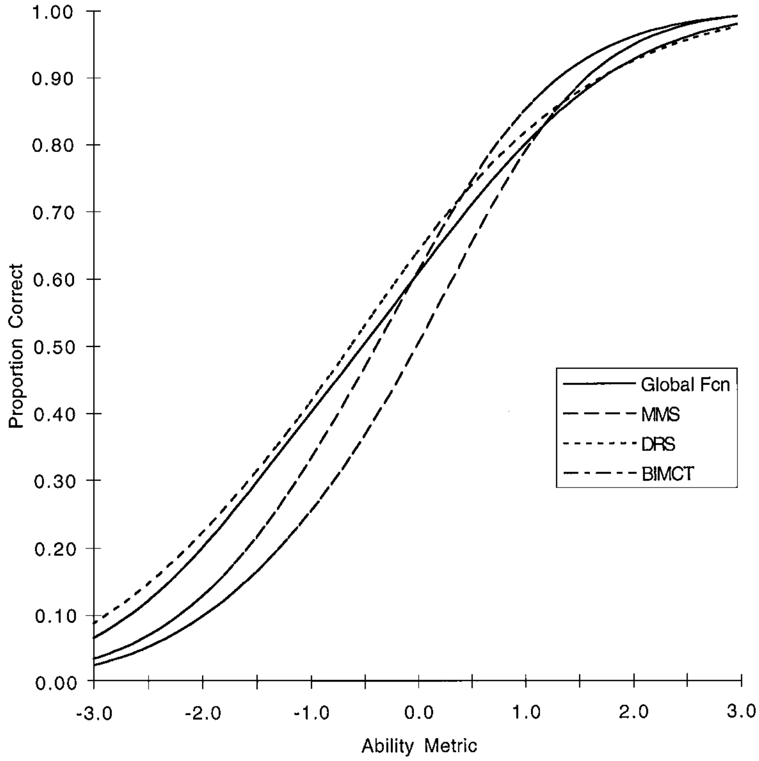


Figure 6. Test characteristic curves (TCCs) for the Mini Mental State Examination (MMS), the Blessed Information Memory Concentration Test (BIMCT), the Blessed–Roth Dementia Rating Scale (BRDRS) and Global Fcn. The TCCs relate proportion of correct items to number of total items to ability.

The non-linear measurement apparent in the TCC for the MMS has been demonstrated in a different manner in previous research. Studies correlating amount of decline on the MMS with initial score have shown an inverted U function, with greater decline associated with initial MMS scores in the mid-ranges (for example, Reference [24]). Results of the present study would suggest that at least part of the observed non-linear rate of decline is due to inherent measurement properties of the MMS. While disease progression may, in reality, occur in a non-linear manner, a linear assessment instrument would be required to demonstrate non-linear progression.

The BRDRS and Global Fcn scale had essentially interchangeable TCCs, suggesting very similar psychometric measurement properties. Substantively, the Global Fcn scale has an advantage in that it incorporates items measuring cognitive functioning as well as independent functioning and hence assesses a broader domain of abilities affected by dementia. This greater breadth of assessment is reflected in the higher correlation of Global Fcn with the MMS and BIMCT.

IRT methods offer significant advantages for psychometric test development. An important feature of IRT is that results are independent of the specific sample of subjects and the specific group of items used if model assumptions are met and the range of subjects and of items is reasonably broad [12–15]. A major limitation of IRT is that large samples are required to achieve stable and reliable solutions, so that 300 to 500 subjects are often required to adequately estimate a two parameter model [15]. The large sample size employed in this study and the broad range of ability of participants are factors that should promote highly reliable results that should be generalizable to other samples of dementia patients.

The Global Fcn scale from this study meets many of the criteria of a good screening measure of global functioning. This measure is brief, applies to a very broad range of ability, assesses both cognitive and functional impairment, and shows promise of linear measurement throughout most of the ability continuum. Its major limitation is with respect to high-end discrimination. This may result in less than optimal sensitivity to mild cognitive changes in high functioning individuals. This limitation of high-end sensitivity is a function of the item pool from which the measure was created, and could be improved by adding other high ability items to the item pool. For example, items from neuropsychological tests with demonstrated sensitivity to early-stage dementia might be incorporated into this measure.

The Global Fcn measure was created using data from three widely used scales that were not constructed to be psychometrically sophisticated measures. A more optimal measure of global functioning would be achievable using IRT methods applied to a new item pool specifically designed to discriminate at low and high extremes of ability as well as in mid-ranges. Development of such a measure would probably require a sample of 500 subjects, but the practical utility of a sophisticated, brief, and easily administered measure of global cognitive functioning would be compelling. Such a measure would have a major impact in the areas of screening assessment for diagnostic purposes, quantitation of the course of the dementia for clinical and research purposes, and measurement of intervention effects.

## ACKNOWLEDGEMENTS

This study was supported in part by grants AG10220, AG10129 and AG12435 from the National Institute on Aging, Bethesda, MD, and by the California Department of Health Services Alzheimer's Disease Program, contract 94-20354.

## REFERENCES

1. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*, *Fourth Edition*. American Psychiatric Association: Washington, DC, 1994.
2. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders*, *Third Edition*, *Revised*. American Psychiatric Association: Washington, DC, 1987.
3. Folstein M, Folstein S, 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.
4. O'Connor DW, Pollitt PA, Hyde JB, Fellows JL, Miller ND, Brook CPB, Reiss BB. The reliability and validity of the Mini-Mental State in a British community survey. *Journal of Psychiatric Research* 1989; **23**:87–96.
5. Mungas D, Marshall SC, Weldon M, Haan M, Reed B. Age and education correction of Mini Mental State Examination for English and Spanish speaking elderly. *Neurology* 1996; **46**:700–706.
6. Ganguli M, Belle S, Ratcliff G, Seaberg E, Hoff FJ, von der Porten K, Kuller L. Sensitivity and specificity for dementia of population-based criteria for cognitive impairment: The MoVIES project. *Journals of Gerontology. Medical Sciences* 1993; **48**:M152–161.
7. Wind AW, Schellevis FG, Van Staveren G, Scholten RP, Jonker C, Van Eijk JT. Limitations of the Mini-Mental State Examination in diagnosing dementia in general practice. *International Journal of Geriatric Psychiatry* 1997; **12**:101–108.
8. Albert MS, Moss MB. The assessment of memory disorders in patients with Alzheimer's disease. In: The Neuropsychology of Memory. 2nd edn Squires LR, Butters N (Eds). The Guilford Press: New York 1992; 211–219.
9. Cullum CM, Filley CM, Kozora E. Episodic memory function in advanced aging and early Alzheimer's disease. *Journal of the International Neuropsychological Society* 1995; **1**:100–103.
10. Mungas D, Wallace R, Reed BR. Dimensions of cognitive ability in dementia: sensitivity to impairment and diagnostic specificity. *Clinical Neuropsychologist* 1998; **12**:129–142.

 

11. Stevens SS. On the theory of scales of measurement. *Science* 1946; **103**:667–680.
12. Hambleton RK, Swaminathan H. *Item Response Theory. Principles and applications.* Kluwer-Nijhoff Publishing: Boston, 1985.
13. Baker FB. The Basics of Item Response Theory. Heineman Publishing: Portsmouth, NH, 1985.
14. Hambleton RK, Swaminathan H, Rogers HJ. *Fundamentals of Item Response Theory.* Sage Publications: Newbury Park, CA, 1991.
15. Hulin CL, Drasgow F, Parsons CK. *Item Response Theory. Application to Psychological Measurement.* Dow Jones-Irwin: Homewood, Ill., 1983.
16. Marshall SC, Mungas D, Weldon M, Reed BR, Haan M. Differential item functioning in the Mini-Mental State Exam in English and Spanish speaking elderly. *Psychology and Aging* 1997; **12**:718–725.
17. Ashford WJ, Kolm P, Colliver JA, Bekian C, Hsu L. Alzheimer patient evaluation and the Mini-Mental State: Item characteristic curve analysis. *Journal of Gerontology*: *Psychological Sciences* 1989; **44**:139–146.
18. Blessed GT, Roth BE, Tomlinson M. The association between quantitative measures of dementia and of senile changes in the cerebral grey matter of elderly subjects. *British Journal of Psychiatry* 1968; **114**:797–811.
19. McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM. Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology* 1984; **34**:939–944.
20. Samejima F. Estimation of latent trait ability using a response pattern of graded scores. *Psychometric Monograph* 1969; **17**.
21. Assessment Systems Corporation. *Users Manual for the XCALIBRE Marginal Maximum-likelihood Estimation Program.* Assessment Systems Corporation: St Paul, MN, 1995.
22. Lord FM, Novick MR. Statistical Theories of Mental Test Scores. Addison-Wesley Publishing Company: Reading, MA, 1968.
23. Lord FM. *Applications of Item Response Theory to Practical Testing Problems.* Lawrence Earlbaum Associates: Hillsdale, NJ, 1980.
24. Butler SM, Ashford JW, Snowdon DA. Age, education, and changes in the Mini-Mental State Exam scores of older women: findings from the Nun Study. *Journal of the American Geriatrics Society* 1996; **44**:675–681.

 