STATISTICS IN MEDICINE
Statist. Med. 2004; **23**:241–256 (DOI: 10.1002/sim.1713)

Test bias in a cognitive test: differential item functioning in the CASI

Paul K. Crane[1,*,†], Gerald van Belle[2] and Eric B. Larson[3]

[1] *Medicine and Public Health and Community Medicine, University of Washington, Seattle, WA, U.S.A.*
[2] *Biostatistics and Environmental and Occupational Health Sciences, University of Washington, Seattle, WA, U.S.A.*
[3] *Center for Health Studies, Group Health Cooperative, and Medicine, University of Washington, Seattle, WA, U.S.A.*

SUMMARY

Assessment of test bias is important to establish the construct validity of tests. Assessment of differential item functioning (DIF) is an important first step in this process. DIF is present when examinees from different groups have differing probabilities of success on an item, after controlling for overall ability level. Here, we present analysis of DIF in the Cognitive Assessment Screening Instrument (CASI) using data from a large cohort study of elderly adults. We developed an ordinal logistic regression modelling technique to assess test items for DIF. Estimates of cognitive ability were obtained in two ways based on responses to CASI items: using traditional CASI scoring according to the original test instructions as well as using item response theory (IRT) scoring. Several demographic characteristics were examined for potential DIF, including ethnicity and gender (entered into the model as dichotomous variables), and years of education and age (entered as continuous variables). We found that a disappointingly large number of items had DIF with respect to at least one of these demographic variables. More items were found to have DIF with traditional CASI scoring than with IRT scoring. This study demonstrates a powerful technique for the evaluation of DIF in psychometric tests. The finding that so many CASI items had DIF suggests that previous findings of differences between groups in cognitive functioning as measured by the CASI may be due to biased test items rather than true differences between groups. The finding that IRT scoring diminished the impact of DIF is discussed. Some preliminary suggestions for how to deal with items found to have DIF in cognitive tests are made. The advantages of the DIF detection techniques we developed are discussed in relation to other techniques for the evaluation of DIF. Copyright © 2004 John Wiley & Sons, Ltd.

*Correspondence to: Paul K. Crane, MD MPH, Division of General Internal Medicine University of Washington, Box 359780, Harborview Medical Center, 325 9th Avenue, Seattle, WA 98104, U.S.A.
†E-mail: pcrane@u.washington.edu

Contract/grant sponsor: National Alzheimer's Coordinating Center; contract/grant number: NIA AG 16976
Contract/grant sponsor: University of Washington's Alzheimer's Disease Research Center; contract/grant number: NIA AG 05136
Contract/grant sponsor: University of Washington's Alzheimer's Disease Patient Registry; contract/grant number: NIA AG 06781

INTRODUCTION

Bias in cognitive screening scales

Bias is a serious problem in psychometric tests. Differential item functioning (DIF) is said to be present when examinees from different groups have differing probabilities of success on an item, after controlling for overall ability [1]. If an item is free of bias, responses to that item will be related only to the level of the underlying trait that the item is trying to measure. If item bias is present, responses to the item will be related to some other factor as well as the level of the underlying trait [2]. The tight relationship between the probability of correct responses and ability or trait levels is an explicit assumption of item response theory (IRT) [3] and an implicit assumption of classical test theory [4]. The presence of large numbers of items with DIF is a severe threat to the construct validity of tests and the conclusions based on test scores derived from items with and items without DIF.

Previous studies have examined cognitive tests for the presence of potential test bias with respect to a number of different demographic characteristics. These include education [5–14], social class [5, 12, 13], neighbourhood type [15], ethnicity [16–19], and age [11, 14]. Various methodological approaches have been taken in these prior analyses. In this paper, we review the various methodologies for detecting DIF and discuss their strengths and limitations. We then outline a novel approach to the detection of DIF using insights from epidemiological methods applied to an ordinal logistic regression (LR) technique first developed without these insights [20]. We then use this technique to assess items from one specific test, the Cognitive Assessment Screening Instrument (CASI) [21], for DIF with respect to several demographic characteristics.

Methods of assessing DIF in items with dichotomous responses: Mantel–Haenszel and logistic regression techniques

Several techniques have been promulgated for the statistical assessment of DIF. Several excellent reviews are available [1, 2, 22]. Most techniques for DIF assessment were developed in educational settings in which items are generally dichotomously scored as correct or incorrect.

Mantel–Haenszel (MH)-based techniques were initially applied to the problem of assessing DIF. It was recognized by the early 1990s that LR-based techniques were more powerful than MH-based techniques [23–25]. This power may come at the expense of increased type I error rates in LR-based techniques [23].

Two distinct forms of DIF have been recognized. These have been called uniform and non-uniform DIF. Uniform DIF is said to apply when differences between groups in item responses are found at all trait levels, while in non-uniform DIF an interaction is found between trait level, group assignment, and item responses [2, 23]. Uniform and non-uniform DIF are directly analogous to the concepts of confounding and effect modification, respectively, in epidemiological research, though this conceptual relationship has not been previously highlighted in the educational testing literature. LR has been known for some time to be useful for the assessment of effect modification in observational studies, and enables analyses of continuous predictor variables without requiring stratification (unlike MH-based techniques). Not surprisingly, simulation studies from educational testing experts have found that LR-based DIF detection techniques enables the detection of both uniform and non-uniform DIF, while MH techniques are better suited for the analysis of uniform DIF [23–25].

Criteria for detection of uniform and non-uniform DIF in logistic regression based DIF detection techniques

Little attention has been paid to the criteria used for determination of the presence or absence of uniform and non-uniform DIF in LR-based DIF detection approaches. The initial description of an LR-based DIF detection assessed both uniform and non-uniform DIF in a single step by comparing the -2 log likelihood difference between a model containing terms for both the demographic characteristic and an interaction term and a model containing neither of these terms to the χ^2 distribution with two degrees of freedom [25].

Given the parallels between uniform and non-uniform DIF on the one hand, and confounding and effect modification on the other, we have used somewhat different criteria for uniform and non-uniform DIF. For non-uniform DIF, we have used a much more stringent criterion because of the potential for finding DIF when it is in fact not present due to multiple hypothesis testing. In a test such as the CASI with its 41 items, when testing each item for DIF with respect to age, two ethnic group comparisons, educational attainment, and gender, 210 different hypotheses are being tested. At a pre-specified $\alpha = 0.05$ level, it would be likely that several true null hypotheses would be falsely rejected by chance alone. We thus chose to adjust our criterion for non-uniform DIF using the Bonferroni technique (discussed further in the discussion section).

For uniform DIF, we noted that there is a modest literature assessing various techniques for empirically determining whether confounding is present in epidemiological studies. We were especially influenced by the simulation studies of Maldonado and Greenland [26]. In this paper, several strategies were compared, including the statistical significance of the coefficient associated with the candidate confounder compared to the $\alpha = 0.05$ level, as well as a strategy that determined that confounding was present if the coefficient associated with the exposure of interest varied by more than 10 per cent in models with and without the presence of the candidate confounder. The 10 per cent change in coefficient criterion proved to be superior to the $\alpha = 0.05$ level criterion for correctly detecting the presence and absence of confounding relationships. Maldonado and Greenland found that the $\alpha = 0.05$ criterion failed to reject the null hypothesis in a large proportion of cases in which confounding was actually present. They found that $\alpha = 0.20$ was a better criterion than $\alpha = 0.05$ for the candidate confounder, and did not find anything to distinguish the $\alpha = 0.20$ and the 10 per cent change criteria [26].

What is of interest in uniform DIF detection is whether the magnitude of the relationship between overall ability level and item responses is altered significantly when taking into account demographic characteristics. This is the essence of confounding and the essence of uniform DIF. This is precisely the question addressed by the 10 per cent change criterion. We have thus chosen to adopt this criterion for determination that uniform DIF is present in an item.

Estimated LR coefficients are determined more or less precisely depending on a number of factors including sample size. The 10 per cent change criterion could be falsely positive in situations in which the coefficients are estimated from small samples. At present little is known about how large a sample needs to be in order to obtain 'stable enough' estimates of regression coefficients to be satisfied that items flagged with uniform DIF are not falsely identified. More work is needed in this area. We have chosen to analyse our data with respect to categories that were either ubiquitous (i.e. age, educational attainment and sex)

or nearly so (i.e. White vs Black ethnic groups; White vs Asian ethnic groups) in our data set.

The extension of logistic regression techniques to polytomous items

The extension of techniques designed for dichotomous items (for example, items scored as 'correct' or 'incorrect') to polytomous items with more than two response categories (for example, 'completely correct,' 'mostly correct,' 'partly correct,' and so on) has been difficult. Good reviews of this topic are available [27–29]. Initial attempts to extend LR techniques to the polytomous case concentrated on recoding data into multiple dichotomies in order to do LR analysis [30, 31]. Recently, it has been noted that the technique of ordinal LR is nicely suited to the task of detection of DIF in polytomous items [20]. Zumbo's ordinal LR technique employs Swaminathan and Rogers' strategy for determining the presence of uniform and non-uniform DIF in a single step by comparing the -2 log likelihood difference between a model containing terms for both the demographic characteristic and an interaction term and a model containing neither of these terms to the χ^2 distribution with two degrees of freedom [20, 25]. Our approach differs from Zumbo's in the criteria employed to determine the presence of uniform and non-uniform DIF.

Item response theory-based methods for assessing differential item functioning

There have been several efforts to use IRT for the detection of DIF in cognitive function tests, including a prior study of DIF with respect to education [32]. We see two major difficulties with IRT-based approaches. The first has to do with large sample sizes. Very large numbers of individuals are needed to fit IRT curves, and when fitting curves for more than one overall group, huge numbers of individuals would be needed to analyse a test for DIF. Embretson and Riese estimate that between 250 and 500 individuals are needed for stable IRT item parameter estimates [33]. For DIF detection this would then require 250–500 individuals in all groups analysed. For predictor variables that are distributed fairly evenly in a sample (e.g. sex) this is a minor inconvenience, but for predictor variables less evenly distributed (e.g. ethnic groups) this requirement may prove to be practically impossible. The second limitation of IRT approaches to DIF detection is that the fitting of curves in separate groups implies that there are clear categories of predictor variables to be examined for DIF. For several demographic characteristics, this is a reasonable assumption. For example, for sex, separating the sample into males and females, fitting IRT curves, and analysing the parameters makes intuitive sense. However, some demographic characteristics are less easy to conceptualize as categorical variables. Specifically with respect to cognitive functioning, education and age are demographic characteristics that should be considered when looking for DIF. IRT techniques require categorizing such continuous variables, and treating those in a particular stratum as identical for the purposes of curve fitting and DIF assessment. This categorization is by its nature somewhat arbitrary. In the paper on the Mattis Dementia Rating Scale, Teresi and colleagues divided educational level into three categories for analysis. They then confirmed their findings by dichotomizing education into two groups and re-running all of their analyses [32]. The techniques we propose here enable education and age (or any continuous demographic variable) to be examined without relying on arbitrary categorization. This improves the power to detect DIF by avoiding categorization of continuous variables into two or three categories, which always results in a loss of power [34].

Importance of DIF with respect to findings of differences between groups

In this study, we developed an ordinal LR technique to analyse a test of cognitive functioning for the presence of DIF. Finding that large numbers of items have DIF leads to the question of whether findings of differences in scores on the test between populations are due to real differences between groups or instead may be due at least in part to the presence of biased test items. The absence of items with DIF would lead to strengthened conclusions regarding differences found between groups with these instruments.

In educational testing settings, items with significant DIF are often discarded. DIF detection is an important first step in the evaluation of test bias. Items flagged with DIF should be examined by content experts to determine whether DIF found is due to a statistical anomaly or instead to item bias [2].

METHODS

Setting and participants

Details of part of the study population used for this paper have been previously published [35, 36]. Adult Changes in Thought (ACT) is a prospective cohort study that focuses on dementia (National Institute on Aging U01 AG06781). The base population for ACT was the Seattle area enrolment of Group Health Cooperative of Puget Sound (GHC) aged 65 years or older. More than 20 000 persons fit that general description when the study began.

Initial cohort enrolment took place between 1994 and 1996. A simple random sample ($N = 6782$) was drawn from the study base (members of GHC aged 65 or older in 1994). Initial medical record review excluded potential subjects who had an existing dementia diagnosis, or those who were in a skilled nursing facility. The remaining consenting subjects were screened to further exclude cases of prevalent dementia.

Of 5422 eligible subjects, 2841 refused to participate in the longitudinal ACT study for a variety of medical, personal and other reasons. The demographic characteristics of the 2841 who refused to participate were similar to those of the overall cohort (data not shown). Levels of educational attainment in the group that refused to participate are not known. There were 2581 non-demented subjects who provided informed consent and were enrolled in the ACT cohort. An additional 359 elderly individuals were added to the study on subsequent cycles with similar recruitment efforts, making a total of 2940 individuals with at least one valid CASI test available for the analyses reported here. Demographic characteristics of these individuals are delineated in Table I.

Estimation of underlying cognitive functioning: The CASI

The CASI was designed for cross-cultural comparisons in cross-national studies with populations of Japanese ancestry [21]. It samples a broad range of cognitive abilities, and domains of attention and concentration, verbal and non-verbal memory, language, visual-spatial functions, executive functions, and drawing were established on the basis of face validity by a group of experts. The CASI incorporates elements of the Mini Mental State Exam (MMSE), the Modified Mini-Mental State (3MS), and the Hasegawa Dementia Scale for the Aged. Scores of each of these shorter tests can be derived from CASI results. The MMSE score derived from the CASI was found to have a correlation coefficient of 0.92 compared with the standard MMSE [37].

Table I. Distributions of gender, age and ethnicity in the ACT cohort analysed for this study (see text for details).

Total sample: number	2940
Gender: *number* (%)	
Female	1730 (58.8%)
Male	1210 (41.2%)
Age category: *number* (%)	
65–69	87 (3%)
70–74	657 (22%)
75–79	857 (29%)
80–84	699 (24%)
85+	570 (19%)
Unknown	70 (2%)
Ethnicity: *number* (%)	
White	2646 (90.0%)
Black	132 (4.5%)
Asian	100 (3.4%)
Other	47 (1.6%)
Unknown	15 (0.5%)
Education category: number (%)	
Less than high school	144 (4.9%)
Some high school	244 (8.3%)
Finished high school	765 (26.0%)
Less than 4 years of college	808 (27.5%)
At least 4 years of college	962 (32.7%)
Unknown	17 (0.6%)

Timing of CASI assessments; selection of test for analysis

Research subjects were evaluated with the CASI at baseline and then every 2 years. If there was evidence of cognitive impairment based on CASI scores, individuals were evaluated every year. The data examined in this paper are from each individual's most recent CASI. The most recent test was chosen for each individual in order to maximize the range of cognitive abilities found.

Scoring the CASI

Two methods of estimating cognitive ability based on CASI responses were used in this study. In the first method, total CASI score was used as an operational definition of latent cognitive ability. The CASI has 41 items that assess a number of different cognitive domains. CASI item content is summarized in the footnote to Table II. CASI items have from 2 to 10 response categories. Standard CASI scoring applies pre-determined scoring weights to each of 9 subscales to attain a total test score that can range from 0 to 100 [21].

In the second method, ability scores generated from IRT analysis of the same CASI responses were used as the operational definition of latent cognitive ability. Scores were generated using a two parameter graded response logistic model [38, 39] using PARSCALE 3.0

(Scientific Software International, Chicago, IL, 1997). The CASI was intended to serve as a measure of cognitive functioning, and it has been found to be sufficiently unidimensional for IRT analyses using confirmatory factor analysis (data not shown). CASI items fit IRT models fairly well (data not shown). The psychometric properties of the CASI will be the subject of a future paper.

Demographic characteristics chosen for analysis of potential item bias

Preliminary analysis of data from the ACT cohort indicated that ethnicity, educational level, and age were independent predictors of cognitive ability when controlling for all other demographic variables (data available upon request). Gender was also found to be an independent predictor of cognitive functioning in whites and blacks when CASI scoring was used as the operational definition of cognitive function, and it was an independent predictor of cognitive functioning in whites when IRT scoring was used as the operational definition of cognitive function (data available upon request). Assessments of DIF in the CASI were thus entertained for age, educational level, gender, and two ethnicity comparisons: blacks to whites and whites to Asians.

Ordinal logistic regression technique

A modification of the ordinal LR technique established by Zumbo [20] was used.

The ordinal LR model for these analyses was:

$$f(\text{response}|\text{trait level}, \text{group}) = \beta_0 + \beta_1 * \text{trait level} + \beta_2 * \text{group} + \beta_3 * (\text{group} * \text{trait level})$$

where the function on the left-hand side of the equation is the ordinal logit. The first step of the analysis for each item and each evaluation of DIF was to look for statistically significant interaction terms, that is to say, the Bonferroni-adjusted p value associated with the β_3 term was <0.05 (see multiple testing discussion below). If this was the case, then significant non-uniform DIF (effect modification) was detected. If not, the interaction term was dropped from further analyses.

Subsequent evaluation was designed to detect uniform DIF (confounding). This was accomplished by comparing the β_1 coefficients from models with and without group assignment entered in the model. For the purposes of this study, a 10 per cent difference between the β_1 coefficient with and without the group variable entered into the model was considered to be evidence of significant uniform DIF (confounding) [26].

Adjustment for multiple testing

Because each assessment of non-uniform DIF is performed in each case on 41 items from the CASI, we used several techniques for adjusting the observed p-values to account for multiple comparisons. For purposes of this paper we adjusted using the Bonferrroni method, multiplying the observed p-values by 41, the number of items examined for DIF. The Bonferroni procedure is conservative when many comparisons are involved, so we are reasonably confident of the significance of Bonferroni-adjusted p-values less than 0.05. We also examined results from the Holm, Hochberg and Šidák procedures for adjusting p-values [40, 41].

Table II. Evaluation of non-uniform and uniform DIF for age, gender, education level and Black: White ethnic groups using CASI and IRT scoring.

	Age*				Gender		Education			Ethnicity (white: black)			
	CASI		IRT		CASI	IRT	CASI	IRT		CASI	IRT	Total, CASI	Total, IRT
Item	Non-uniform	Uniform	Non-uniform	Uniform	Non-uniform	Non-uniform	Non-uniform	Non-uniform	Uniform	Non-uniform	Non-uniform		
bpl†		Y										1	
byr													
bday													
age													
mnt													
sun													
rgs1		Y		Y								1	1
rgs2		Y		Y								1	1
dba													
dbb	Y		Y				Y					2	1
dbc	Y				Y							2	
rc1a	Y											1	
rc1b													
rc1c	Y											1	
sub3a													
sub3b													
sub3c	Y											1	
yr													
mo													
date													
day													
ssn													
spa													
spb													

anml	Y						Y					2	
sim	Y						Y	Y		Y	Y	3	2
jgmt							Y					1	
rpta		Y		Y								1	1
rptb	Y											1	
read		Y										1	
write													
draw								Y					1
cmd													
rc2a					Y	Y						1	1
rc2b							Y					1	
rc2c	Y		Y									1	1
body	Y											1	
obja													
objb													
rpnm									Y				1
rcobj							Y			Y	Y	2	1
Total	10	5	2	3	2	1	6	2	1	2	2	25, 19 items with DIF	11, 10 items with DIF

*The columns below the 'CASI' headings give the results for traditional CASI scoring. The columns below the 'IRT' headings give the results when IRT scoring was used to estimate cognitive ability. A 'Y' is placed in the 'non-uniform' column if the Bonferroni-corrected p-value for the β_3 coefficient for those items was <0.05. A 'Y' is placed in the 'uniform' column Y if the β_1 coefficient was changed by $>10\%$. See text for details. No uniform DIF was found with respect to gender with either CASI or IRT scoring, and no uniform DIF was found with respect to education with IRT scoring. No DIF was found with respect to White:Asian ethnicity.

†Item names: bpl = birth place; byr = birth year; bday = birth day; age = age; mnt = minutes in an hour; sun = direction of sunset; rgs1 = remembering words, first attempt; rgs2 = second attempt; dba = digits backwards (a); dbb = digits backwards (b); dbc = digits backwards (c); sim = similes; jgmt = judgment; rpta = repeat a phrase (a); rptb = repeat a phrase (b); read = read a command and follow it; write = write a specific sentence; draw = copy interlocked pentagons; cmd = follow three step command; rc2a = recall first word; rc2b = recall second word; rc2c = recall third word; body = naming body parts; obja = identify objects (spoon, coin); objb = identify objects (toothbrush, key, comb); rpnm = repeating names of objects; rcobj = recalling objects when hidden from view. See Reference [21] for full questionnaire.

Statistical analyses

All analyses were performed using STATA 7.0 (Stata Statistical Software, College Station, Texas, 2001).

RESULTS

Demographic information for the 2940 individuals evaluated in this study is shown in Table I. The sample included slightly more women than men. While 90 per cent of the sample was white, sizable numbers of blacks and Asians were also studied. This was a very educated sample; a third of the sample had attended at least 4 years of college, and only one-eighth had not at least graduated from high school.

Results of the ordinal LR technique for the assessment of DIF with respect to age, education, gender and ethnicity are shown in Table II. Non-uniform DIF is indicated with a Y in the appropriate cell in the table when the Bonferroni-adjusted p-values were found to be less than 0.05. For items without non-uniform DIF, further investigation for uniform DIF was undertaken. If the value of the coefficient relating trait level (either CASI score or IRT score) to response category changed by more than 10 per cent with and without inclusion of the group variable in the model, a Y is found in the appropriate cell in the table.

Table II shows that many items displayed DIF with respect to age and education. In each case, more non-uniform DIF was detected than uniform DIF. This is not in line with previous results from the educational testing setting, where more items are generally found to have uniform DIF than non-uniform DIF [23].

The method of assessing cognitive functioning based on CASI results had an impact on the number of items found to have DIF. As shown in Table II, in each case, more items were found to have DIF when using traditional CASI scoring than when using IRT scoring to assess cognitive functioning.

When we used a different criterion for the detection of uniform DIF (the $\alpha = 0.20$ criterion as suggested by Maldonado and Greenland [26]), almost every item was found to have uniform DIF with respect to at least one demographic category (data not shown).

Overall, 19/41 items were found to have DIF in at least one analysis when traditional CASI scoring was used. In contrast, 10/41 items were found to have DIF in at least one analysis when IRT scoring was used. Five items were found to have DIF in more than one analysis when traditional CASI scoring was analysed, while only a single item was found to have DIF in more than one analysis when IRT scoring was analysed.

The Bonferroni, Holm, Hochberg, and Šidák procedures for adjusting p-values for multiple comparisons resulted in the same conclusions regarding the presence of non-uniform DIF in every case.

DISCUSSION

Implications of findings

We found DIF in many CASI items in the ACT cohort, especially with respect to educational level and age (Table II). This study presents the broadest assessment to date of the CASI for DIF in the population for which it was intended. One previous study evaluated the Chinese

version of the CASI for DIF related to educational level in 112 patients with Parkinson's Disease from a behavioural neurology referral clinic in Taiwan and, despite the very small numbers of patients, found DIF in items that assessed language and recent memory [42]. The present study then adds to the evidence base that there is DIF in the CASI with respect to education.

Several studies have reported that observed CASI scores differ on the basis of education and/or age. A prior analysis of the individuals from the ACT cohort published curves demonstrating the relationships between age, education, and predicted CASI score in order to help interpret CASI scores [36]. A study from China also noted the need to adjust the CASI score for educational attainment [43]. Two studies have reported that observed CASI scores differ between different ethnic groups [35, 44]. As one of these studies commented [35], it is impossible to know whether the finding that groups differ in cognitive test scores is due to true differences in cognitive functioning, or to using biased tests. Had no DIF been found in the CASI in this study, our confidence that the CASI was not biased would be increased. The presence of DIF with respect to these demographic factors in so many CASI items should prompt content experts to see whether this DIF is due to bias or instead may represent a statistical anomaly [2].

A surprising finding from this study was that the method of scoring the CASI mattered with respect to how much DIF was found. CASI scoring was designed to give each of nine sub-scales roughly equal weight and to provide a total score of 100 [21]. The weighting of individual item responses in IRT is somewhat more complicated. Briefly, items are weighted by how well they discriminate between individuals of different ability levels [4, 33, 45]. It may be that DIF led to poorer fit for some items, in turn leading to lower discrimination and thus less weighting when used in scoring. This finding of less DIF with IRT scoring of the CASI, if replicated in other settings, would be another argument for using IRT scoring rather than traditional sum scores in cognitive tests [46, 47].

Even using IRT scoring techniques, 10 of the 41 CASI items were found to have DIF in this study (and 19 of the 41 items were found to have DIF when traditional CASI scoring was used). This finding is not to suggest that the CASI is *a priori* a bad test; it may well be the best test of its kind [48]. Future work will need to be done to assess DIF in the CASI and other tests of cognitive functioning in other settings.

Should the large number of items with DIF in the CASI be confirmed in subsequent analyses and should bias be confirmed by content experts, several implications are clear: 1. We should strive to understand the particular items that produced DIF with an eye towards writing less biased items. However, previous research in educational settings has demonstrated that experts find it difficult to predict which items will be found to have DIF, and clear understandings of why DIF exists in a particular item are often elusive [49, 50]. 2. We should look for DIF in the measures we use when asking epidemiological questions about the relationship(s) between group membership and cognitive functioning, and epidemiological methods of adjusting for effect modification and confounding should be applied. 3. If the finding of differential DIF between traditional and IRT scoring is confirmed in subsequent studies, a general shift toward IRT scoring should be made in order to minimize the effect of biased items.

To date, there has been little thinking about how the finding that many test items had DIF would impact the assessment of changes over time. It may be the case that DIF prevents valid conclusions at one point of time but does not preclude the validity of following changes in an individual person over time. The finding of large amounts of non-uniform DIF, however,

would lead us to be concerned that if cognitive ability indeed does change in an individual over time, the interactions between that individual's demographic features and new cognitive ability may produce un-interpretable results. As mentioned earlier, the finding of non-uniform DIF is analogous to finding that a demographic variable modifies the relationship between test responses and ability levels. In this case, standard statistical methods of adjusting for confounding relationships will fail. Further thinking and studies are needed to delineate how much of an effect the presence of so many items with non-uniform DIF will have on longitudinal assessments of individuals. These comments are especially true in the context of observational studies. In randomized trials, the randomization process should help distribute both measured and unmeasured attributes evenly between groups. This would probably lead to increased validity compared to observational studies despite the use of a biased assessment tool. However, the presence of non-uniform DIF, analogous to effect modification relationships, makes the serial assessment of individuals over time a problematic exercise. More thinking needs to be done in this area before choosing the CASI or other measures found to have large numbers of items with non-uniform DIF as outcome measures for clinical trials [51].

Implications of methods

Ordinal LR modelling is useful for the assessment of DIF because it retains the full information from the original data. As exhibited in this study, the model can handle dichotomous demographic characteristics such as gender, categorical demographic characteristics such as ethnicity, and continuous demographic characteristics such as age or educational level. This is an advantage over IRT-based techniques, which require categorization of exposures of interest into two or a few groups for DIF assessments.

The techniques delineated here can also detect suspect items with smaller numbers of respondents than is the case with IRT-based approaches. For example, we found significant levels of DIF with respect to Blacks vs Whites in our analyses. However, we did not have enough Blacks in our sample for stable IRT parameter estimates, despite having more than 2900 elderly King County citizens in our sample. Like IRT analyses, though, ordinal LR techniques do require significant numbers of patients for stable parameter estimates. Many of the item analyses failed to converge to stable estimates, especially in the comparison of Blacks and Asians (data not shown). Despite small numbers of Asians and Blacks, valid assessments for DIF with respect to both of those groups in comparison to Whites were possible using LR techniques, while with IRT stable item characteristic curves could not be obtained for either minority group, thus precluding the assessment of DIF with respect to ethnicity using IRT techniques in this data set. IRT techniques for detecting DIF require the estimation not only of parameters for the difficulty and discrimination of each item, but also the ability of each of the test takers, while LR requires only the estimation of the β parameters as described here. Thus, fewer individuals are required for stable parameter estimates with LR than with IRT.

A major advantage of LR (and ordinal LR) techniques over Mantel–Haenszel-based techniques is that the trait level being assessed by the test (in this case, cognitive functioning) is examined as a continuous variable in a single analysis, enabling the use of all of the information contained in the data set [27, 52]. These techniques also have the capacity to handle large numbers of analyses very quickly and with a minimum of cost; these are two of the criteria established for potential polytomous DIF detection techniques by Potenza and Dorans [29]. In order to facilitate the rapid evaluation of data sets for the assessment of DIF

using these techniques, we have developed a STATA package called DIFdetect that uses the methods described in this paper. The package is available for downloading from our website (http://www.alz.washington.edu/DIFDETECT/welcome.html).

We debated among ourselves about the wisdom of adjustment for multiple comparisons when assessing items for DIF. There is a case to be made for not adjusting. That case states that the search for DIF is hypothesis-generating rather than hypothesis-confirming, and that the danger would come from false-negative rather than false-positive assessment of an item. When we did not adjust for multiple comparisons, almost every item was found to have non-uniform DIF in at least one assessment (data not shown). We thus chose to err on the side of caution, and adjusted using the most stringent criterion of which we were aware, the Bonferroni technique. As mentioned above, adjustments with several other less-stringent techniques (the Holm, Hochberg, and Šidák techniques) did not alter those items found to have non-uniform DIF. A graphical analysis of observed p-values along the lines suggested by Schweder and Spjotvøl [53] led to the estimation that there were roughly 80 truly negative null hypotheses in the assessment of non-uniform DIF in these analyses, well over the 27 we present in Table II (data not shown). Further discussion of the multiple comparisons issue (with the plot of p-values derived from these analyses) can be found as the Statistical Rule of the Month for June, 2002 (http://www.vanbelle.org) [54].

Limitations

The principal strength and the main weakness of this study is the cohort of individuals in whom the CASI was used. Members of Group Health Cooperative tend to be better educated and economically more secure than average for King County. Certainly all of these individuals have health insurance and a regular source of medical care. The ACT cohort on the whole is better educated than other large cohorts that have taken the CASI. The finding that there is significant DIF based on educational level is thus quite significant. Our findings may have been different if the educational levels of non-participants differed from those of respondents, and if the relationships between item responses, ability levels, and educational levels differed in these respondents compared to those available for analysis in this study.

A significant limitation is the small number of individuals with dementia included in this cohort. The ACT study was designed to detect the incidence of dementia rather than following individuals with known dementia over time, so individuals with low scores or other findings of dementia were excluded at baseline. To maximize the spread of CASI scores available for this study, the most recent CASI test available for each subject was used in this analysis; some individuals have had significant declines in their CASI score over time. The findings of this study are probably best extended to similar populations not thought to have significant levels of cognitive dysfunction at baseline. Further analyses in cohorts of individuals with cognitive dysfunction will be necessary to determine whether DIF is present at lower cognitive ability levels as well.

The small numbers of Blacks and Asians enrolled in this study decreased the power of this analysis to detect DIF in these ethnic groups. Indeed, for many items, convergence of an ordered LR analysis was not achieved when ethnicity status was examined for DIF. Other data sources with larger numbers of Blacks and Asians should be examined for the presence of DIF with respect to ethnicity.

CONCLUSIONS

This study evaluated a test designed for use in cross-cultural settings and found evidence of DIF with respect to ethnicity, age, and educational status. These findings call into question results of studies that use this and related instruments as measures of cognitive functioning in groups that differ with respect to ethnicity, age, or educational status. These analyses will be repeated with data from a second large cohort study (the Seattle Kame population) [55].

Careful attention to measurement issues and bias are clearly necessary when designing and interpreting studies of cognitive function and cognitive decline [56]. This study presents for the first time in a medical setting the technique of ordinal LR analysis for the assessment of DIF. Our ordinal LR technique provides a flexible and powerful technique for the detection of DIF in polytomous items. In our view, this technique should be considered first when faced with the task of analysing DIF in tests with polytomous items. Software implementing this technique is available for downloading from the website http://www.alz.washington.edu/DIFDETECT/welcome.html.

ACKNOWLEDGEMENTS

Dr van Belle was supported by the National Alzheimer's Coordinating Center (NACC) (NIA AG 16976), the University of Washington's Alzheimer's Disease Research Center (ADRC) (NIA AG 05136), and University of Washington's Alzheimer's Disease Patient Registry (NIA AG 06781).

The authors thank Lance Jolley for help with data management and STATA programming, and wish to also thank two anonymous reviewers for comments that strengthened the manuscript.

REFERENCES

1. Clauser BE, Mazor KM. Using statistical procedures to identify differentially functioning test items. *Educational Measurement*: *Issues and Practice* 1998; **17**:31–44.
2. Camilli G, Shepard LA. *Methods for Identifying Biased Test Items*. Sage: Thousand Oaks, 1994.
3. Hambleton RK, Swaminathan H, Rogers HJ: *Fundamentals of Item Response Theory*. Sage: Newbury Park, 1991.
4. McDonald RP. *Test Theory: A Unified Treatment*. Lawrence Erlbaum: Mahwah, NJ, 1999.
5. O'Connor DW, Pollitt PA, Treasure FP. The influence of education and social class on the diagnosis of dementia in a community population. *Psychological Medicine* 1991; **21**(1):219–224.
6. Jones RN, Gallo JJ. Education bias in the mini-mental state examination. *International Psychogeriatrics* 2001; **13**(3):299–310.
7. Jorm AF, Scott R, Henderson AS, Kay DW. Educational level differences on the mini-mental state: the role of test bias. *Psychological Medicine* 1988; **18**(3):727–731.
8. Black SA, Espino DV, Mahurin R *et al*. The influence of noncognitive factors on the mini-mental state examination in older Mexican–Americans: findings from the Hispanic EPESE. Established population for the epidemiologic study of the elderly. *Journal of Clinical Epidemiology* 1999; **52**(11):1095–1102.
9. Schmand B, Lindeboom J, Hooijer C, Jonker C. Relation between education and dementia: the role of test bias revisited. *Journal of Neurology and Neurosurgery and Psychiatry* 1995; **59**(2):170–174.
10. Ainslie NK, Murden RA. Effect of education on the clock-drawing dementia screen in non-demented elderly persons. *Journal of the American Geriatrics Society* 1993; **41**(3):249–252.
11. Liu HC, Teng EL, Lin KN *et al*. Performance on a dementia screening test in relation to demographic variables. Study of 5297 community residents in Taiwan. *Archives of Neurology* 1994; **51**(9):910–915.
12. Ortiz IE, LaRue A, Romero LJ, Sassaman MF, Lindeman RD. Comparison of cultural bias in two cognitive screening instruments in elderly Hispanic patients in New Mexico. *American Journal of Geriatric Psychiatry* 1997; **5**(4):333–338.
13. Jagger C, Clarke M, Anderson J, Battcock T. Misclassification of dementia by the mini-mental state examination—are education and social class the only factors? *Age and Ageing* 1992; **21**(6):404–411.

14. Kraemer HC, Moritz DJ, Yesavage J. Adjusting mini-mental state examination scores for age and educational level to screen for dementia: correcting bias or reducing validity? *International Psychogeriatrics* 1998; **10**(1): 43–51.
15. Espino DV, Lichtenstein MJ, Palmer RF, Hazuda HP. Ethnic differences in mini-mental state examination (MMSE) scores: where you live makes a difference. *Journal of the American Geriatrics Society* 2001; **49**(5):538–548.
16. Hohl U, Grundman M, Salmon DP, Thomas RG, Thal LJ. Mini-mental state examination and Mattis Dementia rating scale performance differs in Hispanic and non-Hispanic Alzheimer's disease patients. *Journal of the International Neuropsychological Society* 1999; **5**(4):301–307.
17. Blesa R, Pujol M, Aguilar M *et al*. Clinical validity of the 'mini-mental state' for Spanish speaking communities. *Neuropsychologia* 2001; **39**(11):1150–1157.
18. Ripich DN, Carpenter B, Ziol E. Comparison of African–American and white persons with Alzheimer's disease on language measures. *Neurology* 1997; **48**(3):781–783.
19. Marshall SC, Mungas D, Weldon M, Reed B, Haan M. Differential item functioning in the mini-mental state examination in English- and Spanish-speaking older adults. *Psychology and Aging* 1997; **12**(4):718–725.
20. Zumbo BD. *A Handbook on the Theory and Methods of Differential Item Functioning* (*DIF*): *Logistic Regression Modelling as a Unitary Framework for Binary and Likert-type* (*Ordinal*) *Item Scores*. Directorate of Human Resources Research and Evaluation, Department of National Defense: Ottawa, Ont., 1999.
21. Teng EL, Hasegawa K, Homma A *et al*. The cognitive abilities screening instrument (CASI): a practical test for cross-cultural epidemiological studies of dementia. *International Psychogeriatrics* 1994; **6**(1):45–58 (discussion 62).
22. Millsap RE, Everson HT. Methodology review: statistical approaches for assessing measurement bias. *Applied Psychological Measurement* 1993; **17**(4):297–334.
23. Jodoin MG, Gierl MJ. Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education* 2001; **14**(4):329–349.
24. Rogers HJ, Swaminathan H. A comparison of the logistic regression and Mantel–Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement* 1993; **17**(2):105–116.
25. Swaminathan H, Rogers HJ. Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement* 1990; **27**(4):361–370.
26. Maldonado G, Greenland S. Simulation study of confounder-selection strategies. *American Journal of Epidemiology* 1993; **138**(11):923–936.
27. Zwick R, Thayer DT. Evaluating the magnitude of differential item functioning in polytomous items. *Journal of Educational and Behavioral Statistics* 1996; **21**(3):187–201.
28. Welch C, Hoover HD. Procedures for extending item bias detection techniques to polytomously scored items. *Applied Measurement in Education* 1993; **6**(1):1–19.
29. Potenza MT, Dorans NJ. DIF Assessment for polytomously scored items: a framework for classification and evaluation. *Applied Psychological Measurement* 1995; **19**(1):23–37.
30. French AW, Miller TR. Logistic regression and its use in detecting differential item functioning in polytomous items. *Journal of Educational Measurement* 1996; **33**(3):315–332.
31. Miller TR, Spray JA. Logistic discriminant function analysis for DIF identification of polytomously scored items. *Journal of Educational Measurement* 1993; **30**(2):107–122.
32. Teresi JA, Kleinman M, Ocepek-Welikson K. Modern psychometric methods for detection of differential item functioning: application to cognitive assessment measures. *Statistics in Medicine* 2000; **19**(11–12):1651–1683.
33. Embretson SE, Reise SP. *Item Response Theory for Psychologists*. Lawrence Erlbaum: Mahwah, NJ, 2000.
34. van Belle G. *Statistical Rules of Thumb*. Wiley: New York, 2002.
35. Shadlen M-F, Larson EB, Gibbons LE *et al*. Ethnicity and cognitive performance among older African Americans, Japanese Americans, and Caucasians: the role of education. *Journal of the American Geriatrics Society* 2001; **49**(10):1371–1378.
36. McCurry SM, Edland SD, Teri L *et al*. The cognitive abilities screening instrument (CASI): data from a cohort of 2524 cognitively intact elderly. *International Journal of Geriatric Psychiatry* 1999; **14**(10):882–888.
37. Graves AB, Larson EB, Kukull WA, White LR, Teng EL. Screening for dementia in the community in cross-national studies: comparison between the cognitive abilities screening instrument and the mini-mental state examination. In *Alzheimer's Disease*: *Advances in Clinical and Basic Research*, Corain B, Iqbal K, Nicolini M *et al*. (eds). Wiley: New York, 1993; 113–119.
38. Samejima F. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph* 1969; No. **17**.
39. Samejima F. Graded response model. In *Handbook of Modern Item Response Theory*, van der Linden WJ, Hambleton RK (eds). Springer: New York, 1997; 85–100.
40. Hsu J. *Multiple Comparisons: Theory and Methods*. Chapman and Hall: London, 1996.
41. Schouten HJ: Combined evidence from multiple outcomes in a clinical trial. *Journal of Clinical Epidemiology* 2000; **53**(11):1137–1144.

42. Pai MC, Chan SH. Education and cognitive decline in Parkinson's disease: a study of 102 patients. *Acta Neurologica Scandinavica* 2001; **103**(4):243–247.
43. Liu HC, Chou P, Lin KN *et al*. Assessing cognitive abilities and dementia in a predominantly illiterate population of older individuals in Kinmen. *Psychological Medicine* 1994; **24**(3):763–770.
44. Yano K, Grove JS, Masaki KH *et al*. The effects of childhood residence in Japan and testing language on cognitive performance in late life among Japanese American men in Hawaii. *Journal of the American Geriatrics Society* 2000; **48**(2):199–204.
45. Thissen D, Nelson L, Swygert KA. Item response theory applied to combinations of multiple-choice and constructed-response items—approximation methods for scale scores. In *Test Scoring*, Thissen D, Wainer H (eds). Lawrence Erlbaum: Mahwah, NJ, 2001; 293–341.
46. Mungas D, Reed BR. Application of item response theory for development of a global functioning measure of dementia with linear measurement properties. *Statistics in Medicine* 2000; **19**(11–12):1631–1644.
47. du Toit M. *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFACT*. Scientific Software International: Lincolnwood, IL, 2002.
48. Sherrell K, Buckwalter KC, Bode R, Strozdas L. Use of the cognitive abilities screening instrument to assess elderly persons with schizophrenia in long-term care settings. *Issues in Mental Health and Nursing* 1999; **20**(6):541–558.
49. Engelhard G Jr, Hansche L, Rutledge KE. Accuracy of bias review judges in identifying differential item functioning on teacher certification tests. *Applied Measurement in Education* 1990; **3**(4):347–360.
50. Hambleton RK, Jones RW. Comparison of empirical and judgmental procedures for detecting differential item functioning. *Educational Research Quarterly* 1994; **18**(1):21–36.
51. Wong WJ, Liu HC, Fuh JL *et al*. A double-blind, placebo-controlled study of Tacrine in Chinese patients with Alzheimer's disease. *Dementia and Geriatric Cognitive Disorders* 1999; **10**(4):289–294.
52. Zwick R, Thayer DT, Mazzeo J. *Describing and Categorizing DIF in Polytomous Items*. Educational Testing Service: Princeton, NJ, 1997 (GRE Board Report No. 93-10P).
53. Schweder T, Spjøtvoll E. Plots of *P*-values to evaluate many tests simultaneously. *Biometrika* 1982; **69**:493–502.
54. van Belle G. June 2002: multiple comparisons (rule 6.12)., www.vanbelle.org, accessed 5/12/2003.
55. Graves AB, Larson EB, Edland SD *et al*. Prevalence of dementia and its subtypes in the Japanese American population of King County, Washington state. The Kame Project. *American Journal of Epidemiology* 1996; **144**(8):760–771.
56. Teresi JA, Holmes D. Methodological issues in cognitive assessment and their impact on outcome measurement. *Alzheimer Disease & Associated Disorders* 1997; 11(Suppl 6): 146–155.

