Journal of the International Neuropsychological Society (2008), **14**, 746–759.

doi:10.1017/S1355617708081162

Composite scores for executive function items: Demographic heterogeneity and relationships with quantitative magnetic resonance imaging

PAUL K. CRANE,[1] KAAVYA NARASIMHALU,[1] LAURA E. GIBBONS,[1] OTTO PEDRAZA,[2] KALA M. MEHTA,[3] YUXIAO TANG,[4] JENNIFER J. MANLY,[5] BRUCE R. REED,[6,7] AND DAN M. MUNGAS[6]

[1]Department of Medicine, University of Washington, Seattle, Washington
[2]Department of Psychiatry and Psychology, Mayo Clinic, Jacksonville, Florida
[3]Division of Geriatrics, Department of Medicine, University of California-San Francisco, San Francisco, California
[4]Department of Internal Medicine and Rush Institute for Healthy Aging, Rush University Medical Center, Chicago, Illinois
[5]Taub Institute for Research on Alzheimer's Disease and the Aging Brain, the Sergievsky Center, and Department of Neurology, Columbia University, New York, New York
[6]Department of Neurology, University of California-Davis, Sacramento, California
[7]VA Northern California Health Care System, Martinez, California

(RECEIVED August 14, 2006; FINAL REVISION May 30, 2008; ACCEPTED June 13, 2008)

Abstract

Accurate neuropsychological assessment of older individuals from heterogeneous backgrounds is a major challenge. Education, ethnicity, language, and age are associated with scale level differences in test scores, but item level bias might contribute to these differences. We evaluated several strategies for dealing with item and scale level demographic influences on a measure of executive abilities defined by working memory and fluency tasks. We determined the impact of differential item functioning (DIF). We compared composite scoring strategies on the basis of their relationships with volumetric magnetic resonance imaging (MRI) measures of brain structure. Participants were 791 Hispanic, white, and African American older adults. DIF had a salient impact on test scores for 9% of the sample. MRI data were available on a subset of 153 participants. Validity in comparison with structural MRI was higher after scale level adjustment for education, ethnicity/language, and gender, but item level adjustment did not have a major impact on validity. Age adjustment at the scale level had a negative impact on relationships with MRI, most likely because age adjustment removes variance related to age-associated diseases. (*JINS*, 2008, *14*, 746–759.)

Keywords: Composite scores, Item response theory, Dementia, Demographic-adjusted T scores, Ordinal logistic regression, Test bias

INTRODUCTION

Accurate assessment of cognitive ability in individuals from heterogeneous backgrounds is one of the most difficult tasks in neuropsychology. Ethnic diversity is associated with differences in education, language, health, and other factors that may influence test performance. Demographic effects can occur at two distinct levels. Demographic variables can directly effect the cognitive ability measured by the test, and they can be a source of measurement bias. In psychometric theory, observed test scores represent the examinee's ability and measurement error. Bias occurs when ability is systematically under- or over-estimated in one group in comparison with another. When this occurs, measurement error will be systematically different across groups and accuracy of assessment will be compromised.

Tools to account for demographic heterogeneity have been developed using item response theory (IRT). IRT was introduced broadly to psychometrics in 1968 (Lord & Novick, 1968). IRT has revolutionized educational psychology (Hambleton et al., 1991), and has made inroads in other areas (Embretson & Reise, 2000). In IRT, measurement bias is addressed in studies of differential item functioning (DIF). DIF occurs in a test item when individuals from two groups

Correspondence and reprint requests to: Paul Crane, Box 359780, Harborview Medical Center, 325 Ninth Avenue, Seattle, WA 98104. E-mail: pcrane@u.washington.edu

with the same ability have different probabilities of success on that item (Camilli & Shepard, 1994). DIF has received limited attention in neuropsychological assessment, and has been studied primarily in screening tests of global cognition (Crane et al., 2004, 2006a, b; Jones & Gallo, 2001, 2002; Küçükdeveci et al., 2005; Marshall et al., 1997; Teresi et al., 1995, 2000).

Figure 1 illustrates several issues involving demographic effects on test scores and measurement bias. The unobserved cognitive ability is shown in the oval at the top. The observed Composite Score is in the rectangle in the lower half, and observed test item responses 1 through (n+m) are depicted in boxes in the middle. The Composite Score estimates Ability and is created by summing the item responses in some way. Demographic variables can have direct effects on Ability, depicted by the solid arrow, which in turn influences item responses and consequently the Composite Score. When Demographic effects on item responses are entirely due to effects on Ability, an unadjusted Composite Score based on these items provides an unbiased estimate of Ability. However, when Demographic variables have influences on some item response independent of Ability (dotted arrows to items i_{n+1} through i_{n+m}), these items have DIF and introduce bias to the estimate of Ability. In Figure 1, an unadjusted Composite Score derived from items i_1 through i_n would provide an unbiased estimate of Ability, because any effects of Demographics are entirely mediated by their impact on Ability. Adding items i_{n+1} through i_{n+m} without adjustment introduces bias, because the effects of Demographics are mediated in part independent of Ability. The impact of that bias is complexly determined. A primary goal of this study was to identify measurement bias in items assessing executive function, and to empirically evaluate its impact on practical assessment questions.

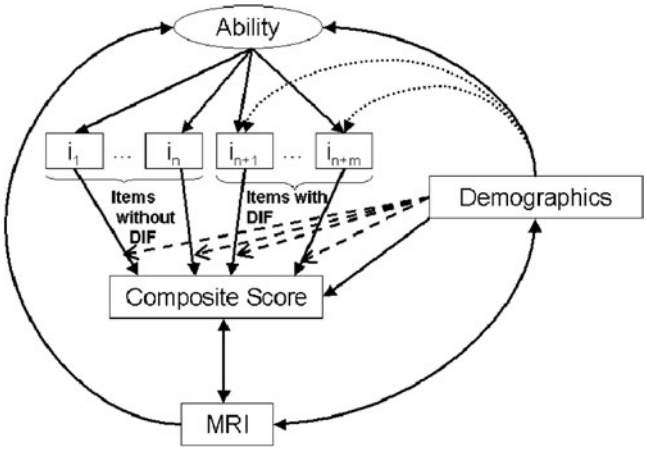

Fig. 1. Schematic representation of relationships analyzed in this study. Ability is represented in an oval at the top. Ability is reflected by item responses on a cognitive test (depicted in boxes as i_1 through i_{n+m}). Demographic characteristics may directly impact Ability (depicted by the solid arrow between Demographics and Ability) and may be associated with item bias (depicted by the dotted arrows to the items with differential item functioning, abbreviated in the Figure as DIF, specifically items i_{n+1} through i_{n+m}). Formulas are used to obtain composite scores from the observed item responses, depicted in the figure by the solid arrows between the item responses and the composite score. Traditional test theoretic composite score formulas include summing up observed responses or summing up average scores. Traditional test theoretic composite score formulas that account for demographic heterogeneity apply the same adjustment to all the items (depicted in the figure as the four dashed arrows extending from Demographics to the solid arrows extending from all of the items to the composite score). Modern psychometric theory formulas (known as item response theory or IRT) empirically calibrate item difficulty across the range of cognitive ability levels, resulting in nonlinear relationships between traditional composite scores and IRT scores. IRT formulas that account for DIF apply adjustments for demographics only to those items found with DIF (depicted in the figure with the rightmost two dashed arrows extending from Demographics to the solid arrows extending from items i_{n+1} through i_{n+m} to the composite score). Finally, we compared these composites based on their strength of relationship with MRI measures of white matter hyperintensity and total brain volume. These evaluations included scale-level accounting for demographic heterogeneity, indicated in the figure by the solid arrow extending from Demographics to the composite score and the double headed arrows between the composite score and MRI, and between MRI and demographics. Note that measurement error is not depicted in the Figure.

IRT approaches that account for DIF remove the demographic influence on items that is independent of ability, or equivalently, that is irrelevant to measuring ability. Only items with DIF (items i_{n+1} through i_{n+m} in Figure 1) are affected by this adjustment, and demographic effects mediated by ability are not removed. In Figure 1, in this strategy, items i_1 through i_n would have no adjustment, while items i_{n+1} through i_{n+m} would be adjusted for DIF related to Demographics, depicted by the rightmost two dashed lines leading from Demographics to the arrows from items i_{n+1} through i_{n+m} to the Composite Score. This strategy will not eliminate measurement error, but will result in measurement error being unrelated to demographics.

Strategies to account for demographic diversity at the item level also have emerged from classical psychometric theory. These strategies include item-level adjustments of scores or, equivalently, adjusting item norms for demographic characteristics. Returning to Figure 1, such strategies adjust all items contributing to a composite score, whether or not those items have DIF. This is depicted by the dashed arrows emanating from the Demographics box and ending on all of the lines from Items to the Composite Score, whether or not those items had DIF. Adjusting in this manner removes demographic variance in individual items that is due to DIF, but in contrast to IRT based DIF adjustment, also removes direct demographic effects mediated by ability. Thus, there may still be a systematic relationship between measurement error and demographics. This is an important theoretical and practical distinction between classical and modern psychometric approaches to accounting for demographic heterogeneity.

Ability in the psychometric sense is the net result of all factors that influence capacity to respond successfully to

https://doi.org/10.1017/S1355617708081162

test items. Differences in ability result from multiple influences. In neuropsychology, we are primarily interested in the effects of brain injuries and diseases on ability. Demographic influences independent of brain variables might lead to erroneous conclusions about the presence and severity of brain injury. For example, a low test score in a highly educated person might be a strong indicator of a dementing illness like Alzheimer's disease, but the same score might be expected in a person with a normal brain who has very limited education. This issue has fueled debate about whether test scores should be adjusted for demographic variables at the scale level. Fundamentally, this is an issue of establishing an estimate of expected performance in the absence of brain injury. The obtained test score can then be compared with this estimate to make an inference about whether disease or brain injury has resulted in cognitive impairment.

Several studies have shown that using unadjusted norms results in false-positive errors among functionally and cognitively normal ethnic minorities and people with few years of education (Fillenbaum et al., 1990; Gasquoine, 1999; Manly et al., 1998; Ramirez et al., 2001; Stern et al., 1992). Demographic adjustment of test scores or group specific norms help to reduce false-positive results in minority and low education groups, and generally make diagnostic sensitivity and specificity more homogenous across diverse groups (Mungas et al., 1996). The advisability of scale level demographic adjustment is not universally accepted (Belle et al., 1996; Brandt, 2007; Kraemer et al., 1998; Reitan & Wolfson, 2004, 2005; Sliwinski et al., 1997). The most compelling argument against it is decreased validity for detecting effects of brain disease and injury (Kraemer et al., 1998). This would be the case if a demographic variable is related to brain variables and exerts effects on ability primarily as a result of this relationship.

Returning to Figure 1, magnetic resonance imaging (MRI) measures of brain structure are used as indicators of brain integrity. In this model, MRI has direct effects on Ability. Demographics can have an indirect impact on Ability as a result of effects mediated by MRI (arrows from Demographics to MRI to Ability), but can also effect Ability through pathways unrelated to MRI (direct arrows from Demographics to Ability). Adjustment to eliminate demographic effects independent of MRI might improve validity for detecting brain injury, but removing Demographics effects that are mediated by MRI could decrease validity in this context.

We examined item and scale level effects of demographic heterogeneity on a composite measure of executive function in this study. Executive function refers to cognitive operations that involve control and coordination of other cognitive activities (Stuss & Levine, 2002) and is generally thought to reflect frontal lobe and frontal system function. The composite measure in this study was based on fluency and working memory tasks. These are not conceptualized as pure frontal measures and likely are influenced by cortical changes in nonfrontal regions, but there is broad agreement that fluency and working memory are important executive function subdomains. We followed a similar approach in previous work (Mungas et al., 2003, 2005a) using different executive tasks in a different sample and found differential effects of brain regions and systems on an executive composite and a psychometrically matched measure of episodic memory (Mungas et al., 2005a).

This study is part of ongoing development of the Spanish and English Neuropsychological Assessment Scales (SENAS). Previous work has developed and validated measures of nonexecutive domains (Mungas et al., 2000, 2005b,c). Measures of fluency and working memory have been added, and validation with respect to independently obtained clinical diagnosis has previously been reported (Mungas et al., 2005c).

Our primary goal was to compare item- and scale-level strategies for handling demographic heterogeneity in a measure of executive function. Demographic variables of interest included age, ethnicity/language, education, and gender. We examined the extent to which DIF distorted test-based estimates of ability. We then examined the extent to which item level and scale level adjustment for demographic variables influenced the relationships of various composite scores with an external criterion, in this case structural MRI measures of total brain matter and white matter hyperintensity (WMH). We chose these MRI measures because they are relevant to executive function (Gunning-Dixon & Raz, 2000; Kramer et al., 2002; Meguro et al., 2003) and directly measure brain structure in a manner that is blind to demographic characteristics of the person being assessed.

METHOD

Participants

Participants were 815 persons recruited by the UC Davis Alzheimer's Disease Center under protocols designed to increase representation of ethnic minorities and maximize heterogeneity of cognitive functioning. There were 271 whites, 544 ethnic minorities (312 Hispanics, 208 black or African Americans, 15 Asians, 1 Native American, and 9 other or missing); 240 Hispanics were tested in Spanish, and all other participants were tested in English. A community screening program designed to identify and recruit individuals with cognitive functioning ranging from normal to demented identified 704 individuals (185 whites, 519 minorities). The remaining 111 (86 whites, 25 minorities) were initially seen at a university memory/dementia clinic and referred for research. We excluded the 25 participants who were not Hispanic, white, or black or African Americans from the present analyses.

All community recruits were 60 years of age or older. Clinical patients under 60 were included if they were being evaluated for cognitive impairment associated with diseases of aging. Inclusion criteria included ability to speak English or Spanish. Participants signed informed consent under protocols approved by institutional review boards at UC Davis, the Veterans Administration Northern California

https://doi.org/10.1017/S1355617708081162

Health Care System, and San Joaquin General Hospital in Stockton, California.

A subsample of participants was referred for clinical evaluation and a research MRI on the basis of SENAS scales measuring episodic memory, semantic memory, attention span, visual spatial abilities, and verbal abstraction. A 25% random sample of those with normal cognition were invited to participate in clinical evaluation and MRI, and all with memory or nonmemory cognitive impairment were invited to participate. Exclusion criteria for selection in this stage included unstable major medical illness, major primary psychiatric disorder (history of schizophrenia, bipolar disorder, or recurrent major depression), and substance abuse or dependence in the past 5 years. These individuals all received a clinical diagnosis based on a comprehensive clinical evaluation, but SENAS results were excluded from consideration in establishing clinical diagnosis. Sampling percentages were used as weights to relate the MRI subsample back to the overall sample and to estimate the prevalence of specific diagnostic categories in the whole sample. Estimated prevalence by diagnosis was: cognitively normal, 57.1%; MCI, 31.0%; and demented, 11.9%. The subsample who received MRI included 171 individuals (83 whites and 88 minorities). Some of these individuals had missing data for executive function items and were excluded from comparative analyses (see footnote to Table 2).

Neuropsychological Measures

The SENAS measures of fluency and working memory are commonly used tasks or are adaptations appropriate for Spanish speaking and/or illiterate individuals. Fluency measures included animals, words beginning with /f/ and /l/ sounds, and total items and categories from the Supermarket Test (Mattis, 1988). Scores were recorded separately for the first and second 30 seconds. Working memory measures included Digit Span Backwards, Visual Span Backwards, and a new List Sorting task. List Sorting has two parts. In part 1, participants are presented with a list of fruits or animals and are asked to repeat all of the elements on the list, but in order from smallest to largest. In part 2, the lists include both fruits and animals and the task is to repeat fruits first, sorted from smallest to largest, and then animals in order from smallest to largest. For both parts 1 and 2, 15 lists of increasing length are presented yielding total scores that range from 0 to 15.

Terms used in neuropsychological assessment may produce some confusion, as often a "scale" comprises a single item (e.g., Trails B), and at other times a "scale" comprises a total score from several items (e.g., the Mattis Dementia Rating Scale). We will refer to the most granular data as an "item" whether or not it is also considered a "scale." Specifically, we incorporated 13 items to create candidate scores measuring executive function. The term "scale" in subsequent discussion refers to candidate scores that summarize performance on the 13 items.

MRI Measures

Brain imaging was obtained at the UC Davis MRI research center on a 1.5T GE Signa Horizon LX Echospeed system or the Veterans Administration at Martinez on a 1.5 T Marconi system. Comparable imaging parameters were used at each site. Detailed methods for obtaining brain and WMH volumes are presented in Appendix 1.

Data Analysis

Generation of candidate scores

We used four techniques to generate scores from fluency and working memory tasks. These techniques are summarized in Table 1, where they are categorized by their underlying psychometric theory (classical *vs*. item response theory) and whether they account for demographic heterogeneity at the item level.

A commonly used technique is to determine means and standard deviations for each item in a battery, which are then used to determine *Z* scores for each individual on each item, which are then averaged across all items. *Unadjusted* T *scores* are re-scaled *Z* scores; instead of N(0,1), *T* scores are N(50,10).

Table 1. Summary of composite scoring techniques for executive function assessment tools

Handling of demographic differences in test scores	Psychometric theory: Classical test theory	Psychometric theory: Item response theory
Demographic differences ignored	Unadjusted *T* score	Unadjusted item response theory score
Item-Level Demographic effects accounted for	Demographic adjusted *T* score	Item response theory score accounting for differential item functioning

https://doi.org/10.1017/S1355617708081162

We used linear regression to determine mean scores appropriate for each education, gender, and ethnicity/language category for each item. Linear regression models for each item included any interactions significant at an alpha level of 0.05. We obtained a pooled standard deviation from the residuals (the difference between the observed and the regression-estimated mean score for each education, gender, and ethnicity/language category). These means and *SD*s were then used to obtain demographic-adjusted *Z* scores for each of the 13 items. We averaged these scores and re-scaled them to generate *demographic-adjusted composite* T *scores*. For a secondary analysis, we repeated these steps including age as a fourth demographic category.

We used the item response theory (IRT) package Parscale (Muraki & Bock, 2003) to obtain *unadjusted IRT scores*. We verified that items were sufficiently unidimensional for IRT purposes using a confirmatory factor analysis approach (McDonald, 1999). Details of the dimensionality and IRT assessments are shown in Appendix 2.

We used a software package we developed called difwithpar (Crane et al., 2006a) to obtain *IRT scores accounting for DIF*. Detailed methods are shown in Appendix 3. These methods determine IRT parameters for each item found to have DIF separately in appropriate demographic subgroups, thus permitting relationships between items and the latent ability to be somewhat different across different demographic groups. We determined the impacts of DIF for each demographic variable (gender, age, education, and ethnicity/language group) for individual participants by subtracting their unadjusted IRT score from their IRT score accounting for DIF related to that covariate. If DIF made no impact this difference would be 0; if DIF had a large impact it would be large. We also determined IRT scores accounting for DIF related to all demographic variables. We used the median standard error of measurement from the entire sample as a benchmark to determine whether DIF had a meaningful or salient impact on individual test scores. IRT estimates ability and the standard error of measurement. We considered differences larger than the median standard error of measurement to indicate meaningful or salient scale-level differential functioning (Crane et al., 2007).

Comparison of candidate scores

For the subsample of 153 participants with complete executive function data and MRI data, we estimated MRI and demographic effects on each candidate score by entering the scores as dependent variables in linear regression models. We evaluated simple effects of MRI variables on the scores with and without demographic covariates in the model (i.e., with and without scale-level adjustment).

We initially examined demographic effects for ethnicity/language, education, and gender, but not age. We ran three models for each composite score. Model A included demographic terms as independent variables. Model B included the MRI variables representing normalized total brain matter and WMH volumes. Model C included all demographic and MRI variables.

We used R^2 values from these models to estimate effect sizes. Simple MRI effects were the R^2 values from the model with only the MRI variables (Model B). We determined incremental MRI effects by subtracting the R^2 from Model A (demographics alone) from the R^2 from Model C (demographics and MRI). We compared these differences using Hotelling's method (Hotelling, 1944). Finally, we repeated the regression analyses including age along with the other demographic variables in Models A and C.

RESULTS

Demographic characteristics of the 791 participants are shown in Table 2. Older participants were more likely to have an MRI. Participants who received an MRI on average were better educated and more likely to be white. While individuals with MRI scans were not representative of the overall population, it is not likely that selection bias drove our results, as executive function scores were not used to determine who was selected for MRI assessment.

DIF findings are summarized in Table 3. One item had DIF related to gender, four had DIF related to age, six had DIF related to education, and seven had DIF related to ethnicity/language group. Only the first 30 seconds of fluency with /l/ was free of DIF for all four covariates.

DIF impact on individual scores is shown in Figure 2. Accounting for DIF related to gender did not change any participant's score by more than the median standard error of measurement, accounting for DIF related to age changed one participant's score, accounting for DIF related to ethnicity/language changed 8 participants' scores (1%), and accounting for DIF related to education changed 70 participants' scores (9%). Accounting for all four sources of DIF simultaneously changed 68 participants' scores (9%).

Results of MRI regression analyses are shown in Table 4. The left three columns of results are the amount of variance explained (R^2) from models with demographics only (column A), MRI only (column B), and the full model with both demographics and MRI variables (column C). The shaded column labeled "Both MRI variables (C-A)" shows the difference in R^2 between the full model and the demographics only model. The remaining two columns provide the unique contributions of WMH alone and total brain volume alone.

Focusing on column A, the demographics only models, it is not surprising that scores that account for demographic heterogeneity have less variability explained by demographics. This is true both for *T* scores, where unadjusted *T* scores have 25% of their variance explained by demographic characteristics, and adjusted *T* scores have only 6% of their variance explained by demographic characteristics, and for IRT scores, where unadjusted IRT scores have 28% of their variance explained by demographic characteristics, and IRT scores accounting for DIF have 20% of their variance explained by demographics.

The model in Figure 1 is helpful in understanding these results. Column A in Table 4 represents the strength of association between the Demographics box and the Composite

https://doi.org/10.1017/S1355617708081162

Table 2. Demographic characteristics of participants with and without MRI[a]

	MRI,[c] complete data (n = 153)		MRI,[c] missing data (n = 18)		No MRI (n = 619)		Total (n = 790)	
	n	%	*n*	%	*N*	%	*n*	%
Age ($p < .001$)								
45–64	19	12%	5	28%	164	27%	188	24%
65–69	23	15%	2	11%	128	20%	153	19%
70–74	36	24%	7	39%	144	23%	187	24%
75–79	37	24%	4	22%	114	18%	155	20%
80+	38	25%	0	0%	69	11%	107	14%
Gender ($p = .74$)								
Male	66	43%	7	39%	246	40%	319	40%
Female	87	57%	11	61%	373	60%	471	60%
Years of education[b] ($p = .058$)								
0–8	40	26%	5	33%	204	34	249	33%
9–13	47	31%	1	7%	178	30	226	30%
14+	66	43%	9	60%	215	36	290	38%
Ethnicity/language ($p < .001$)								
White	79	52%	4	22%	187	29%	270	34%
Black or African-American	31	20%	3	17%	174	27%	208	26%
Hispanic (English)	12	8%	5	28%	55	9%	72	9%
Hispanic (Spanish)	31	20%	6	33%	203	32%	240	30%

[a]*p* values are based on Fisher's exact test.
[b]Education status missing for 3 participants with an MRI and 22 participants without an MRI.
[c]MRIs were obtained on 171 participants. Only 153 of these had complete demographic and complete executive function data, meaning no missing data for any of the 8 elements enumerated in Table 4. These 153 participants were analyzed in regression analyses.

Score box in Figure 1. When demographic variability is removed from all of the items with adjusted *T* scores (dashed lines between Demographics and all of the arrows from the items to the Composite Score), the amount of variability in the Composite Score remaining to be explained by demographics is negligible (6% in this case). However, when only demographic heterogeneity not mediated by Ability is removed from those items in which there is a direct relationship between Demographics and item responses (i.e., in those items with DIF), demographics may still explain a salient amount of the variability in scores, as shown here, where the IRT score that accounts for DIF related to demographics still has 20% of its variability explained by demographics. In essence, because the demographic-adjusted *T* score approach adjusts every item, whether or not the item has DIF, it may over-correct for demographic effects, minimizing the relationship between demographics and the composite score mediated by the effects of demographics on ability.

Column B in Table 4 represents the amount of variance in candidate scores explained by MRI variables alone, without demographic factors in the model. Here the difference between the two IRT scores is negligible (4% for unadjusted IRT scores and 6% for IRT scores accounting for DIF). However, the difference between the two *T* scores is remarkable (6% for unadjusted *T* scores *vs*. 22% for adjusted *T* scores). Returning to Figure 1, for the adjusted *T* scores, because essentially all variability related to Demographics has been removed from the Composite Score, the strength of relationship with MRI is artificially accentuated.

The artifice of this accentuation is discernible when considering the values of the shaded column in Table 4. Here we see that accounting for demographics in regression models for unadjusted *T* scores, unadjusted IRT scores, and IRT scores accounting for DIF improves the strength of relationship with MRI (by 12–13% in each case), while for the adjusted *T* score, the value in column B differs from the value in the shaded column by only 1%. Thus, MRI effects are much stronger after scale level adjustment for demographic effects—unless an adjusted *T* score approach is used. Differences in incremental MRI effects (C-A) across scoring methods were not statistically significant.

Age was not used in the regression models presented in Table 4, and was not used in the demographic-adjusted *T* score (i.e., the demographic-adjusted *T* score included adjustments for ethnicity/language, gender, and education, but not age). We performed additional analyses with demographic-adjusted *T* scores that accounted for age differences as well as ethnicity/language group, gender, and education. Again, demographic effects (column A in Table 4) were negligible, with 7% of the variance. Including age in

https://doi.org/10.1017/S1355617708081162

Table 3. Differential item functioning for executive function items related to age, education, gender, and ethnicity/language group

	Age		Education		Gender		Ethnicity/ Language	
Item	U	NU	U	NU	U	NU	U	NU
Animals 1	2%	0.19	6%	0.67	0%	0.31	**7%**	0.55
Animals 2	2%	0.19	**15%**	0.2	0%	0.66	**15%**	**0.03**
F 1	2%	0.96	2%	0.21	0%	0.43	1%	**0.05**
F 2	1%	0.44	3%	0.32	0%	**0.01**	2%	0.65
L 1	0%	0.25	1%	0.64	0%	0.84	0%	0.4
L 2	0%	0.86	3%	0.33	0%	0.34	1%	**0.03**
Supermarket Items 1	1%	0.14	**11%**	0.24	0%	0.41	6%	**0.03**
Supermarket Items 2	2%	**<0.01**	2%	0.22	0%	0.13	0%	0.79
Supermarket Categories	2%	**<0.01**	2%	0.39	1%	0.86	5%	0.19
Digit Span Backward	3%	0.09	**16%**	**<0.01**	0%	0.08	**10%**	0.91
Visual Span Backward	1%	**<0.01**	**18%**	0.64	1%	0.35	**12%**	**0.04**
List Sorting 1	0%	**<0.01**	**10%**	0.11	0%	0.12	5%	0.84
List Sorting 2	1%	0.43	**9%**	0.13	0%	0.24	4%	0.49

Note. Numbers in the "U" columns represent uniform DIF findings. Uniform DIF occurs in an item if members of one group are at a consistent advantage or disadvantage for that item relative to another group for every executive function level. For example, people with lower levels of educational attainment had lower expected scores on the animals item compared to people with higher educational attainment at all levels of executive function. The numbers shown here represent the proportional change in the β_1 coefficient from including or excluding the group term or terms from models 2 and 3. Changes with an absolute value of at least 7% are shown in bold font. Larger values indicate larger differences in expected score across groups for a given executive function level. Numbers in the "NU" columns represent non-uniform DIF findings. Non-uniform DIF occurs in an item if there is an interaction between executive function level, group membership, and expected scores. For example, this occurred for the supermarket categories item related to age. There are two possible relationships when there is non-uniform DIF. First, the probabilities may cross, so that older people have (for example) a higher expected score at high executive function levels, but lower expected scores at lower executive function levels. Second, the probabilities may not cross, but be more extreme at one end of the executive function spectrum than the other, so that older people have (for example) a lower expected score at high executive function levels, but much lower expected scores at low executive function levels. The numbers shown here are p values associated with the likelihood difference between models 1 and 2 (including or excluding the interaction term or terms). All p values less than .05 are shown in bold font. Here, smaller values indicate more statistically significant interaction between the group term and overall executive function. See Appendix 3 for details.

the T scores diminished the strength of association with MRI, with R^2 of 16% as opposed to 22% when the demographic-adjusted T score did not account for age. Differences in incremental effects of MRI were also diminished for the demographic-adjusted T score that included age, with the difference in R^2 of 15% as opposed to 23% when the demographic-adjusted T score did not include age.

We repeated the regression analyses used in the generation of Table 4, this time including age as an independent variable in each analysis. In each case the incremental amount of variance explained by the MRI variables was diminished compared with regression models that excluded age. This result is shown graphically in Figure 3. In each case, including age (either adjusting norms for age, as in T scores, or regression models for age, as in Figure 3) reduced the strength of association with MRI scores.

DISCUSSION

Test bias is present when individuals from different groups who have the same ability have different expected test or item scores (Camilli & Shepard, 1994). Ability in psychometric theory is a latent construct. It is measured by items, and item responses are combined in some way to arrive at a score that estimates ability. If bias exists at the item level, this could lead to a biased estimate of ability. Conversely, group differences in means and distributions of test scores do not necessarily indicate that bias is present. DIF adjustment as used in this study essentially removes the effects of measurement bias. This item level adjustment allows for evaluating valid effects of demographic variables on ability apart from measurement bias.

Different pathways may produce the same latent ability. We modeled the independent effects of demographic and MRI variables on candidate scores, comparing item- and scale-level approaches. IRT scores accounting for DIF are unbiased estimates of individual ability and can be used in regression analyses to determine which variables to include in scale level adjustments. The adjusted T score approach, in contrast, produces a biased estimate of individual ability, and different candidate scores are required to determine whether adjustments should be made. Here, the best adjusted T score necessitated results from the MRI analyses to show that age should not be used. The IRT/DIF approach separates internal and external considerations into two steps, facilitating better understanding of relationships between

https://doi.org/10.1017/S1355617708081162

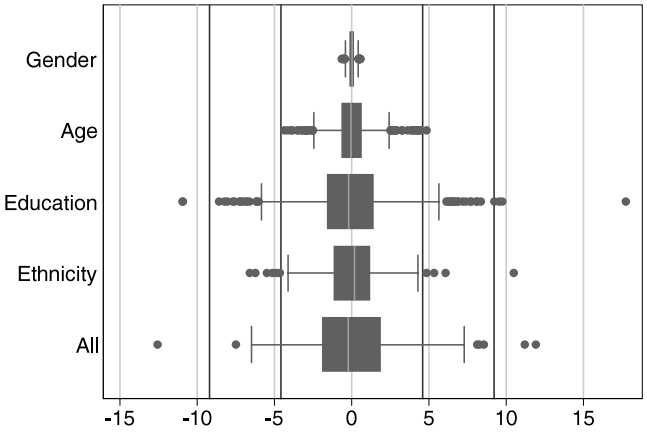

Fig. 2. Impact on estimated executive function scores of differential item functioning related to gender, age, education, and race separately, and related to all four covariates simultaneously. The x-axis maps the distribution of the difference scores obtained between individuals' executive function scores accounting for DIF and executive function scores that ignore DIF (i.e., If DIF made no impact on scores, then the difference in scores would be 0). All scores were transformed such that 1 standard deviation is 15 points. For each adjustment strategy, the distribution is illustrated with a box-and-whiskers plot (the box defines the 25th, 50th, and 75th percentiles, while the whiskers define 1 $\frac{1}{2}$ times the interquartile range; individual observations more extreme than this are indicated with dots). The vertical lines indicate the median value of the standard error of measurement for the population and twice the median value of the standard error of measurement for the population; the range of the standard error of measurement was 3.9 to 7.3 points. Differences when accounting for DIF greater than the median standard error of measurement are referred to as "salient scale-level differential functioning."

variables of interest and immediate access to unbiased scores.

In this study, 9% of subjects had salient or meaningful impact from DIF. Controlling for ethnicity/language, education, and gender substantially strengthened relationships with MRI variables; adjusting for age weakened these relationships. These findings suggest that it is important to understand scale level demographic effects, and to control for some but not all of these effects to optimally measure brain-related cognitive effects, especially in demographically heterogeneous samples.

The results shown here are to our knowledge the first report of DIF in executive function items. Figure 2 is very helpful in documenting the impact of DIF on individual scores. Much of the impact of DIF is related to education and to race/ethnicity. Some participants had scores that were affected by DIF by as much as 1/3 of a standard deviation. DIF impact of this magnitude is most likely to be problematic when using cutoff scores to determine whether an individual is impaired. DIF could impact the validity of clinical diagnosis in this context.

This study is unique in that it examined the impact of DIF on test validity, using a culture-blind validation criterion, structural MRI. Previous studies of cognition and MRI have shown modest associations of volume of WMH and executive function (Gunning-Dixon & Raz, 2000; Kramer et al., 2002; Meguro et al., 2003). Those studies included much more homogeneous samples of participants, and the amount of variability explained by MRI variables was greater than that explained by MRI alone in our study (column B in Table 4). After accounting for scale-level gender, ethnicity, and education effects—but *not* for age—we found a similar

Table 4. Variance of composite executive function scores explained by MRI variables and ethnicity/language group, education, and gender (but not age)

	R^2 values			Incremental R^2 values for MRI variables		
Score	Demographics only (A)	MRI only (B)	Full model (both demographics and MRI) (C)	Both MRI variables (C-A)	WMH (C-A-total brain volume)	Total brain volume (C-A-WMH)
Unadjusted *T* score	0.25	0.06	0.44	0.19	0.06	0.11
Demographic-adjusted *T* score	0.06	0.22	0.29	0.23	0.08	0.12
Unadjusted IRT score	0.28	0.04	0.45	0.17	0.06	0.10
IRT score accounting for DIF	0.20	0.06	0.38	0.18	0.06	0.10

Note. Values in the columns labeled "R^2 values" are the R^2 from Model A("Demographics") including three demographic variables (ethnicity/language group, education, gender), Model B ("MRI only") including both MRI variables (total brain volume and WMH [white matter hyperintensity volume]), and Model C ("Full model") including the three demographic and two MRI variables. Values in the columns labeled "Incremental R^2 values for MRI variables" represent the difference in R^2 between the full model (column C) and models that include all the variables with the exception of the elements named in the heading of the column. Thus, for the column labeled "Both MRI variables," the values shown are the difference in R^2 between the full model (C) and the demographics alone model (A). Values in the next two columns (WMH and Total brain volume) are the difference in R^2 between the full model and a model with demographics plus the other MRI value. We performed Hotelling's tests on the values in the shaded column. None of these values was statistically different from any other value (all p values $> .10$).

https://doi.org/10.1017/S1355617708081162

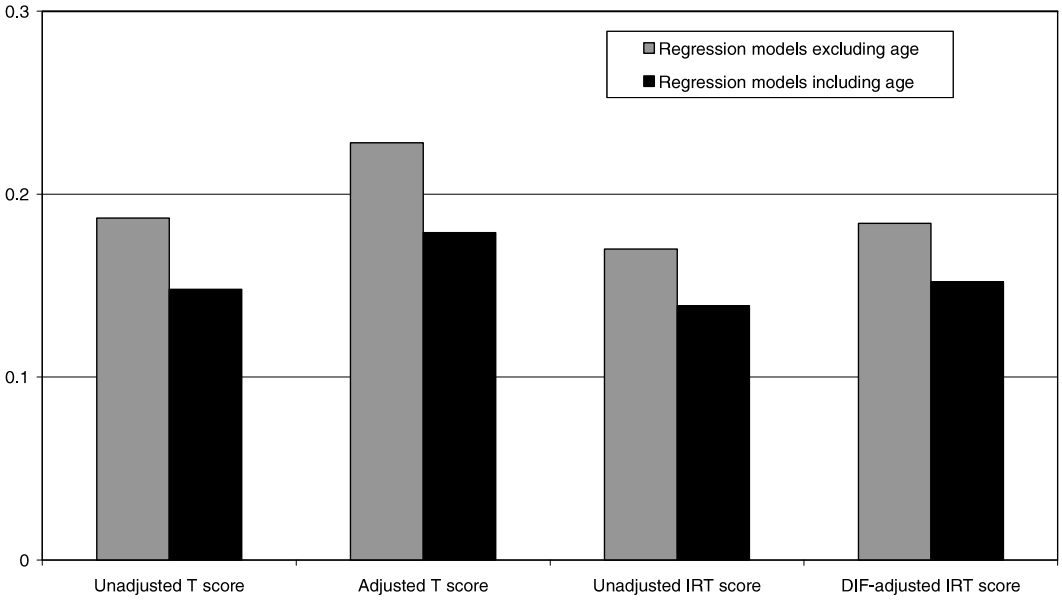

Fig. 3. Incremental variance explained by structural MRI variables in Executive Function composite scores not adjusting for age (gray bars) and adjusting for age (black bars). Values represent the R^2 for a full model with both MRI variables and demographics minus the R^2 for a model with only demographics [see the shaded column in Table 4 labeled "Both MRI variables (C-A)"]. Age was included as a demographic variable in the age adjusted model and was not included in the model without age adjustment. Adjusted *T* scores were adjusted for gender, ethnicity/language, and education, but not age.

amount of variability explained by MRI variables as had been found in these prior studies. When accounting for age as well, the relationship with MRI variables was actually smaller.

Scale level adjustment for demographic variables like ethnicity and education was beneficial because it removed a variance component from the total ability estimate unrelated to brain structure, thereby making the brain structure effect more salient. Age adjustment had a negative impact on test validity. The effect was consistent across all composite score strategies. The adverse effect of age adjustment likely is because age is strongly related to disease processes that result in cognitive impairment. Removing age-related variance effectively decreases disease-related variance and decreases relationships of test scores with structural changes in the brain associated with disease. Age was strongly related to MRI variables in this sample ($R^2 = 0.40$). While simple bivariate correlations of age with executive function measures were significant (*r*'s in the 0.20s), age was unrelated to executive function independent of MRI variables (incremental R^2s ranged from < 0.01 to 0.02). In contrast, the relationship of the other demographic variables to test scores were equally strong after controlling for MRI variables (not shown).

An intermediate finding was that cognitive tasks involving category fluency, phonemic fluency, and working memory were sufficiently unidimensional to be combined into a composite executive function measure (see Appendix 2). This finding is consistent with findings that executive function tasks are highly correlated (Salthouse, 2005). The fluency and working memory tasks used in this study may be influenced by multiple brain regions. From a substantive perspective, including different tasks expands the brain regions being monitored, and adding items increases reliability. These characteristics are likely to increase sensitivity to broad disease-based effects on frontal systems, but at the expense of specificity to more specific frontal lobe structures of systems. Ultimately, the utility of any test is an empirical question, and depends on the intended purpose for the test. Consequently, if the goal of neuropsychological assessment is to identify relatively small focal lesions, a broader measure may be problematic. If the goal is to monitor broader disease effects on frontal subcortical systems then a broader measure has much to offer.

This study has several limitations. It examined a limited set of specific cognitive measures in a specific and unusually diverse sample, and different results might be found with different cognitive domains and different populations. Additionally, different results might be found with different techniques for identifying items with DIF (Millsap, 2006). Furthermore, while the MRI measures are presumably culture-free, the relative validity findings are limited to the extent that MRI measures of WMH and total brain volume capture important features related to executive function.

Accounting for DIF had demonstrable benefits in this study in terms of improving accuracy of estimation of individual ability. DIF adjustment can be accomplished with no additional testing time or burden. Addressing DIF in neuropsychological test development, however, requires a substantial investment, particularly in obtaining a sufficiently large development sample to permit DIF analyses. There is also an investment needed in analytic and computational infrastructure to use IRT algorithms that account for DIF,

https://doi.org/10.1017/S1355617708081162

although this barrier is continually diminishing. As neuropsychological tests are used with increasingly diverse patient populations, this level of investment may become a minimum requirement for demonstrating psychometric properties appropriate to the population of interest. This study shows the neuropsychological relevance of demographic influences on test performance, and highlights the need for further studies with heterogeneous populations and broader measures of cognition.

ACKNOWLEDGMENTS

Data collection was supported by grants AG10220 (Mungas), AG10129 (DeCarli), and AG021028 (DeCarli). Data analysis was begun at the 2005 Friday Harbor Psychometrics Workshop, with conference grant support from the Alzheimer's Association. Additional data analysis was supported by grant K08 AG22232 (Crane) and R01 AG029672 (Crane). Dr. Gibbons was supported by NIH grant P50 AG05136. Parts of these analyses were presented at the 2006 International Conference on Alzheimer's Disease and Related Disorders in Madrid, Spain, July 2006. None of the authors has any conflict of interest.

REFERENCES

Baker, F.B. & Kim, S.-H. (2004). *Item Response Theory: Parameter Estimation Techniques* (2nd revised and expanded ed.). New York: Marcel Dekker.

Belle, S.H., Seaberg, E.C., Ganguli, M., Ratcliff, G., DeKosky, S., & Kuller, L.H. (1996). Effect of education and gender adjustment on the sensitivity and specificity of a cognitive screening battery for dementia: Results from the MoVIES Project. Monongahela Valley Independent Elders Survey. *Neuroepidemiology*, *15*, 321–329.

Brandt, J. (2007). 2005 INS Presidential Address: Neuropsychological crimes and misdemeanors. *The Clinical Neuropsychologist*, *21*, 553–568.

Camilli, G. & Shepard, L.A. (1994). *Methods for Identifying Biased Test Items*. Thousand Oaks, CA: Sage.

Crane, P.K., Gibbons, L.E., Jolley, L., & van Belle, G. (2006a). Differential item functioning analysis with ordinal logistic regression techniques: DIFdetect and difwithpar. *Medical Care*, *44*(Suppl. 3), S115–S123.

Crane, P.K., Gibbons, L.E., Jolley, L., van Belle, G., Selleri, R., Dalmonte, E., & De Ronchi, D. (2006b). Differential item functioning related to education and age in the Italian version of the Mini-mental State Examination. *International Psychogeriatrics*, *18*, 505–515.

Crane, P.K., Gibbons, L.E., Narasimhalu, K., Lai, J.S., & Cella, D. (2007). Rapid detection of differential item functioning in assessments of health-related quality of life: The Functional Assessment of Cancer Therapy. *Quality of Life Research*, *16*, 101–114.

Crane, P.K., Hart, D.L., Gibbons, L.E., & Cook, K.F. (2006c). A 37-item shoulder functional status item pool had negligible differential item functioning. *Journal of Clinical Epidemiology*, *59*, 478–484.

Crane, P.K., van Belle, G., & Larson, E.B. (2004). Test bias in a cognitive test: Differential item functioning in the CASI. *Statistics in Medicine*, *23*, 241–256.

DeCarli, C., Maisog, J., Murphy, D.G., Teichberg, D., Rapoport, S.I., & Horwitz, B. (1992). Method for quantification of brain, ventricular, and subarachnoid CSF volumes from MR images. *Journal of Computer Assisted Tomography*, *16*, 274–284.

DeCarli, C., Massaro, J., Harvey, D., Hald, J., Tullberg, M., Au, R., Beiser, A., D'Agostino, R., & Wolf, P.A. (2005). Measures of brain morphology and infarction in the Framingham heart study: Establishing what is normal. *Neurobiology of Aging*, *26*, 491–510.

DeCarli, C., Miller, B.L., Swan, G.E., Reed, T., Wolf, P.A., Garner, J., Jack, L., & Carmelli, D. (1999). Predictors of brain morphology for the men of the NHLBI twin study. *Stroke*, *30*, 529–536.

DeCarli, C., Murphy, D.G., Teichberg, D., Campbell, G., & Sobering, G.S. (1996). Local histogram correction of MRI spatially dependent image pixel intensity nonuniformity. *Journal of Magnetic Resonance Imaging*, *6*, 519–528.

Embretson, S.E. & Reise, S.P. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: Erlbaum.

Fillenbaum, G., Heyman, A., Williams, K., Prosnitz, B., & Burchett, B. (1990). Sensitivity and specificity of standardized screens of cognitive impairment and dementia among elderly black and white community residents. *Journal of Clinical Epidemiology*, *43*, 651–660.

Gasquoine, P.G. (1999). Variables moderating cultural and ethnic differences in neuropsychological assessment: The case of Hispanic Americans. *The Clinical Neuropsychologist*, *13*, 376–383.

Gibbons, R.D., Bock, R.D., Hedeker, D., Weiss, D., Bhaumik, D.K., Kupfer, D., Frank, E., Grochocinski, V., & Stover, A. (2007). Full-information item bi-factor analysis of graded response data. *Applied Psychological Measurement*, *31*, 4–19.

Gunning-Dixon, F.M. & Raz, N. (2000). The cognitive correlates of white matter abnormalities in normal aging: A quantitative review. *Neuropsychology*, *14*, 224–232.

Hambleton, R.K., Swaminathan, H., & Rogers, H.J. (1991). *Fundamentals of Item Response Theory*. Newbury Park, CA: Sage Publications.

Holland, P.W. & Wainer, H. (Eds.). (1993). *Differential Item Functioning*. Hillsdale, NJ: Erlbaum.

Hotelling, H. (1944). The selection of variates for use in prediction, with some comments on the general problem of nuisance parameters. *Annals of Mathematical Statistics*, *11*, 271–283.

Hu, L.-t. & Bentler, P.M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. *Psychological Methods*, *3*, 424–453.

Hu, L.-t. & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55.

Jack, C.R., Jr., O'Brien, P.C., Rettman, D.W., Shiung, M.M., Xu, Y., Muthupillai, R., Manduca, A., Avula, R., & Erickson, B.J. (2001). FLAIR histogram segmentation for measurement of leukoaraiosis volume. *Journal of Magnetic Resonance Imaging*, *14*, 668–676.

Jones, R.N. & Gallo, J.J. (2001). Education bias in the mini-mental state examination. *International Psychogeriatrics*, *13*, 299–310.

Jones, R.N. & Gallo, J.J. (2002). Education and sex differences in the Mini-Mental State Examination: Effects of differential item functioning. *The Journals of Gerontology, Series B, Psychological Sciences and Social Sciences*, *57*, P548–P558.

Kraemer, H.C., Moritz, D.J., & Yesavage, J. (1998). Adjusting Mini-Mental State Examination scores for age and educational level to screen for dementia: Correcting bias or reducing validity? *International Psychogeriatrics*, *10*, 43–51.

Kramer, J.H., Reed, B.R., Mungas, D., Weiner, M.W., & Chui, H.C. (2002). Executive dysfunction in subcortical ischaemic vascular disease. *Journal of Neurology, Neurosurgery, and Psychiatry*, *72*, 217–220.

https://doi.org/10.1017/S1355617708081162

Küçükdeveci, A.A., Kutlay, S., Elhan, A.H., & Tennant, A. (2005). Preliminary study to evaluate the validity of the mini-mental state examination in a normal population in Turkey. *International Journal of Rehabilitation Research*, *28*, 77–79.

Lord, F.M. & Novick, M.R. (1968). *Statistical Theories of Mental Test Scores, with Contributions by Allan Birnbaum*. Reading, MA: Addison-Wesley.

MacCallum, R.C. & Austin, J.T. (2000). Applications of structural equation modeling in psychological research. In S.K. Fiske, D.L. Schacter & C. Zahn-Waxler (Eds.), *Annual Review of Psychology*: Vol. 51. (pp. 201–226). Palo Alto, CA: Annual Reviews.

Maldonado, G. & Greenland, S. (1993). Simulation study of confounder-selection strategies. *American Journal of Epidemiology*, *138*, 923–936.

Manly, J.J., Jacobs, D.M., Sano, M., Bell, K., Merchant, C.A., Small, S.A., & Stern, Y. (1998). Cognitive test performance among nondemented elderly African Americans and whites. *Neurology*, *50*, 1238–1245.

Marshall, S.C., Mungas, D., Weldon, M., Reed, B., & Haan, M. (1997). Differential item functioning in the Mini-Mental State Examination in English- and Spanish-speaking older adults. *Psychology and Aging*, *12*, 718–725.

Mattis, S. (1988). *Dementia Rating Scale*. Odessa, FL: Psychological Assessment Resources.

McCullagh, P. & Nelder, J.A. (1989). *Generalized Linear Models* (2nd ed.). London: Chapman and Hall.

McDonald, R.P. (1999). *Test Theory: A Unified Treatment*. Mahwah, NJ: Erlbaum.

Meguro, K., Constans, J.M., Shimada, M., Yamaguchi, S., Ishizaki, J., Ishii, H., Yamadori, A., & Sekita, Y. (2003). Corpus callosum atrophy, white matter lesions, and frontal executive dysfunction in normal aging and Alzheimer's disease. A community-based study: The Tajiri Project. *International Psychogeriatrics*, *15*, 9–25.

Millsap, R.E. (2006). Comments on methods for the investigation of measurement bias in the Mini-Mental State Examination. *Medical Care*, *44*(Suppl. 3), S171–S175.

Mungas, D., Reed, B.R., Marshall, S.C., & Gonzalez, H.M. (2000). Development of psychometrically matched English and Spanish language neuropsychological tests for older persons. *Neuropsychology*, *14*, 209–223.

Mungas, D., Harvey, D., Reed, B.R., Jagust, W.J., DeCarli, C., Beckett, L., Mack, W.J., Kramer, J.H., Weiner, M.W., Schuff, N., & Chui, H.C. (2005a). Longitudinal volumetric MRI change and rate of cognitive decline. *Neurology*, *65*, 565–571.

Mungas, D., Reed, B.R., Haan, M.N., & Gonzalez, H.M. (2005b). Spanish and English neuropsychological assessment scales: Relationship to demographics, language, cognition, and independent function. *Neuropsychology*, *19*, 466–475.

Mungas, D., Reed, B.R., Tomaszewski Farias, S., & DeCarli, C. (2005c). Criterion-referenced validity of a neuropsychological test battery: Equivalent performance in elderly Hispanics and non-Hispanic Whites. *Journal of the International Neuropsychological Society*, *11*, 620–630.

Mungas, D., Marshall, S.C., Weldon, M., Haan, M., & Reed, B.R. (1996). Age and education correction of Mini-Mental State Examination for English and Spanish-speaking elderly. *Neurology*, *46*, 700–706.

Mungas, D., Reed, B.R., & Kramer, J.H. (2003). Psychometrically matched measures of global cognition, memory, and executive function for assessment of cognitive decline in older persons. *Neuropsychology*, *17*, 380–392.

Muraki, E. & Bock, D. (2003). *PARSCALE for Windows* (Version 4.1). Chicago: Scientific Software International.

Muthen, L.K. & Muthen, B.O. (1998–2004). *Mplus User's Guide* (3rd ed.). Los Angeles: Muthen & Muthen.

Ramirez, M., Teresi, J.E., Silver, S., Holmes, D., Gurland, B., & Lantigua, R. (2001). Cognitive assessment among minority elderly: Possible test bias. *Journal of Mental Health and Aging*, *7*, 91–118.

Reitan, R.M. & Wolfson, D. (2004). Clinical and forensic issues regarding age, education, and the validity of neuropsychological test results: A review and presentation of a new study. *Journal of Forensic Neuropsychology*, *4*, 1–32.

Reitan, R.M. & Wolfson, D. (2005). The effect of age and education transformations on neuropsychological test scores of persons with diffuse or bilateral brain damage. *Applied Neuropsychology*, *12*, 181–189.

Salthouse, T.A. (2005). Relations between cognitive abilities and measures of executive functioning. *Neuropsychology*, *19*, 532–545.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, No. 17.

Samejima, F. (1997). Graded response model. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of Modern Item Response Theory* (pp. 85–100). New York: Springer.

Sliwinski, M., Buschke, H., Stewart, W.F., Masur, D., & Lipton, R.B. (1997). The effect of dementia risk factors on comparative and diagnostic selective reminding norms. *Journal of the International Neuropsychological Society*, *3*, 317–326.

Stern, Y., Andrews, H., Pittman, J., Sano, M., Tatemichi, T., Lantigua, R., & Mayeux, R. (1992). Diagnosis of dementia in a heterogeneous population. Development of a neuropsychological paradigm-based diagnosis of dementia and quantified correction for the effects of education. *Archives of Neurology*, *49*, 453–460.

Stuss, D.T. & Levine, B. (2002). Adult clinical neuropsychology: Lessons from studies of the frontal lobes. *Annual Review of Psychology*, *53*, 401–433.

Teresi, J.A., Golden, R.R., Cross, P., Gurland, B., Kleinman, M., & Wilder, D. (1995). Item bias in cognitive screening measures: Comparisons of elderly white, Afro-American, Hispanic and high and low education subgroups. *Journal of Clinical Epidemiology*, *48*, 473–483.

Teresi, J.A., Kleinman, M., & Ocepek-Welikson, K. (2000). Modern psychometric methods for detection of differential item functioning: Application to cognitive assessment measures. *Statistics in Medicine*, *19*, 1651–1683.

APPENDIX 1. DETAILED METHODS OF NEUROIMAGE DATA ANALYSIS

Analysis of brain and WMH volumes was based on a Fluid Attenuated Inversion Recovery (FLAIR) sequence designed to enhance WMH segmentation (Jack et al., 2001). WMH segmentation was performed in a two-step process (DeCarli et al., 1992, 1999). In brief, nonbrain elements were manually removed from the image by operator guided tracing of the dura matter within the cranial vault including the middle cranial fossa, but excluding the posterior fossa and

cerebellum. The resulting measure of the cranial vault was defined as the total cranial volume (TCV) to correct for differences in head size.

The first step in image segmentation required the identification of brain matter. Image intensity nonuniformities (DeCarli et al., 1996) were then removed from the image and the resulting corrected image was modeled as a mixture of two Gaussian probability functions with the segmentation threshold determined at the minimum probability between these two distributions (DeCarli et al., 1992). Once brain matter segmentation was achieved, a single Gaussian distribution was fitted to the image data and a segmentation threshold for WMH was a priori determined at 3.5 *SD*s in pixel intensity above the mean of the fitted distribution of brain parenchyma. Morphometric erosion of two exterior image pixels was also applied to the brain matter image before modeling to remove the effects of partial volume CSF pixels on WMH determination. Reliability estimates for these methods are high (DeCarli et al., 2005).

Two MRI measures were used independently in subsequent analyses. These were normalized brain volume (brain matter/TCV) and normalized white matter hyperintensity (white matter hyperintensity/TCV).

APPENDIX 2. DIMENSIONALITY AND IRT ANALYSES

We used confirmatory factor analyses implemented by MPlus version 3.0 (Muthen & Muthen, 1998–2004). We used McDonald's bi-factor method (McDonald, 1999). Each item is specified to have loadings on a single general factor as well as on a more specific subdomain factor defined *a-priori* based on theoretical considerations. This approach has recently been discussed by Gibbons et al. (2007). We assigned each fluency and working memory item to one of three subdomains (category fluency, phonemic fluency, or working memory). A graphical summary of this model is shown in Appendix Figure 1. The general executive function factor is defined by all of the items, while the category fluency, phonemic fluency, and working memory subdomain factors are defined by a few of the items, as shown in Figure 1. McDonald suggests that if standardized loadings on the general factor all exceed 0.30, then the scale is sufficiently unidimensional for applications requiring unidimensionality. If loadings on subdomains also exceed 0.30, then one could use subdomains for some applications and summary scores of the general factor for other applications, as appropriate (McDonald, 1999). Because all of the items had many response categories, we treated them all as continuous indicators. We assessed model fit using the comparative fit index (CFI), the Tucker-Lewis index (TLI), and the root mean squared error of approximation (RMSEA). These three summary fit indices have been recommended for evaluating model fit due to their ability to robustly detect model misfit in data sets with a variety of violations of basic assumptions (Hu & Bentler, 1998, 1999; MacCallum & Austin, 2000). CFI and TLI values > 0.95 indicate good model fit; RMSEA < 0.08 indicates adequate fit, and RMSEA < 0.06 indicates good fit (Hu & Bentler, 1999).

The bi-factor model fit the data well, with CFI 0.98, TLI 0.96, and RMSEA 0.052. Loadings for the bi-factor model are summarized in Appendix Table 1. Loadings for each item on the general executive function factor ranged from 0.42 to 0.77, well in excess of the 0.30 threshold for salience (McDonald, 1999). We thus considered executive function as assessed by these items to be a sufficiently unidimensional construct to proceed with IRT.

We used Parscale 4.1 (Muraki & Bock, 2003) for IRT modeling. We used Samejima's graded response model (Samejima, 1969, 1997), which is an extension of a 2-parameter logistic (2PL) model for dichotomous items to items with many response categories ("polytomous" items). We used a normal prior for *expectation a posteriori* scoring; results were similar when we used maximal likelihood scoring.

Parscale uses an iterative approach to determining item and person parameters. The 2-parameter logistic (2PL) model

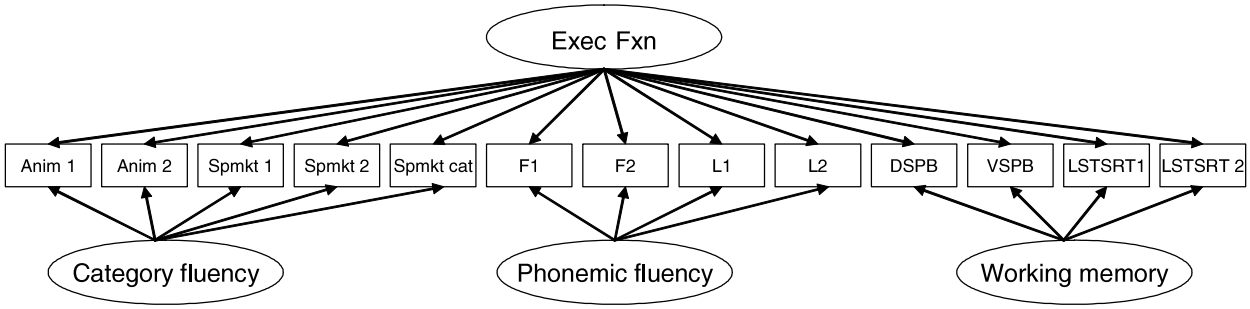

Fig. A1. Schematic representation of the executive function bi-factor confirmatory factor analysis. Abbreviations: Exec Fxn = executive function; Anim 1 = animal fluency, 1st 30 seconds; Anim 2 = animal fluency, 2nd 30 seconds; Spmkt 1 = supermarket items, 1st 30 seconds; Spmkt 2 = supermarket items, 2nd 30 seconds; Spmkt cat = number of categories of supermarket items over 60 seconds; F1 = words beginning with f produced in the 1st 30 seconds; F2 = words beginning with f produced in the 2nd 30 seconds; L1 = words beginning with l produced in the 1st 30 seconds; L2 = words beginning with l produced in the 2nd 30 seconds; DSPB = digit span backwards; VSPB = visual span backwards; LSTSRT 1 = list sorting 1; LSTSRT 2 = list sorting 2.

https://doi.org/10.1017/S1355617708081162

Table A1. Bi-factor model results of executive function items

Item	Standardized loading on general executive function factor	Standardized loading on subdomain factor	Name of subdomain factor
Animals 1	0.63	0.27	
Animals 2	0.50	0.19	
Supermarket 1	0.58	0.46	Category Fluency
Supermarket 2	0.58	0.35	
Supermarket categories	0.42	0.44	
F 1	0.72	0.42	
F 2	0.57	0.49	Phonemic Fluency
L 1	0.77	0.39	
L 2	0.60	0.32	
Digit span backwards	0.67	0.45	
Visual span backwards	0.60	0.39	Working Memory
List sorting 1	0.69	0.48	
List sorting 2	0.55	0.42	

has 2 parameters for each item—a difficulty and a discrimination parameter. The equation for the 2PL is as follows: $P(Y = 1|\theta, a, b) = \exp(Da(\theta - b))/[1 + \exp(Da(\theta - b))]$.

Here $P(Y = 1)$ means the probability of success on item Y, θ is the subject's ability level, b is item difficulty, a is item discrimination, and D is a constant that makes the logistic curve approximate the normal ogive. The probability of success is 50% where $\theta = b$. The logistic curve varies around that point proportional to the discrimination parameter a. Samejima's graded response model is an extension of the 2PL model to multiple categories using the proportional odds assumption. A single slope parameter is estimated for each item, but multiple difficulty parameters are estimated as the thresholds between adjacent response categories. Details of equations and estimation procedures can be found in Baker and Kim (Baker & Kim, 2004). We rescaled raw Parscale output so that the mean score was 100 and the standard deviation was 15 using a linear transformation.

APPENDIX 3: METHODS FOR IDENTIFYING ITEMS WITH DIFFERENTIAL ITEM FUNCTIONING (DIF)

The specific issue of item-level bias is addressed in studies on differential item functioning (DIF). The definition of DIF is a conditional one: when controlling for the underlying ability measured by the test, DIF occurs when the probabilities of success on an item differ related to group membership. Thus, for a given level of ability, for a specific item, members of Group A have a higher probability of success than members of Group B. Two types of DIF are identified in the literature: uniform and nonuniform DIF. In an item with uniform DIF, the advantage Group A members over Group B members is constant across the spectrum of abilities measured by the test. In nonuniform DIF, however, the advantage varies across the spectrum of abilities measured by the test, and even the direction may change. Thus, in an item with nonuniform DIF, members of Group A with high ability levels may have a higher probability of success on the item than members of Group B with high ability levels, while members of Group A with low ability may have a lower probability of success on the item than members of Group B with low ability.

We have developed an approach to DIF assessment that combines ordinal logistic regression and IRT. Details of this approach are outlined in earlier publications (Crane et al., 2004, 2006c).

We use IRT executive function scores to evaluate items for DIF. We examine three ordinal logistic regression models for each item for each demographic category (labeled here as "group") selected for analysis:

$$f(\text{item response}) = cut + \beta_1 * \theta + \beta_2 * \text{group} + \beta_3 * \theta * \text{group} \quad \text{(model 1)}$$

$$f(\text{item response}) = cut + \beta_1 * \theta + \beta_2 * \text{group} \quad \text{(model 2)}$$

$$f(\text{item response}) = cut + \beta_1 * \theta \quad \text{(model 3)}$$

In these models, *cut* is the cutpoint for each level in the proportional odds ordinal logistic regression model (McCullagh & Nelder, 1989), and θ is the IRT estimate of executive function.

To detect nonuniform DIF, we compare the log likelihoods of models 1 and 2 using a χ^2 test, $\alpha = .05$. To detect uniform DIF, we determine the relative difference between the parameters associated with θ (β_1 from models 2 and 3) using the formula $|(\beta_{1(\text{model 2})} - \beta_{1(\text{model 3})})/\beta_{1(\text{model 3})}|$. If the relative difference is large, group membership interferes with the expected relationship between ability and item responses. There is little guidance from the literature regarding how large the relative difference should be. A

https://doi.org/10.1017/S1355617708081162

	Demographic Category A	Demographic Category B	Comments
DIF free items 1 through m	Original Responses	Original Responses	Item parameters estimated from all people in sample
Items with DIF (m+1) through (m+n)	Original Responses	Missing	Item parameters estimated from those in Demographic Category A
Items with DIF (m+1) through (m+n)	Missing	Original Responses	Item parameters estimated from those in Demographic Category B

Fig. A2. Handling of items by their differential item functioning (DIF) status. In this schematic there are a total of $(n + m)$ items included in the test; n of these items are found with DIF, while m items do not have DIF.

simulation by Maldonado and Greenland on confounder selection strategies used a 10% change criterion in a very different context (Maldonado & Greenland, 1993). We have previously used 10% (Crane et al., 2004) and 5% (Crane et al., 2006c) change criteria. In this data set, when we used a 5% change criterion, almost every item had either uniform or nonuniform DIF related to ethnicity/language, resulting in unstable parameter estimates. We thus used a 7% criterion. For the other three covariates, the difference in impact on individual scores when accounting for DIF using a 5% *versus* a 7% change criterion was negligible. It may be appropriate to determine the change criterion used empirically from the data, selecting a level that still leaves a few items free of DIF to serve as anchor items (Crane et al., 2007). Anchor items are items that have the same parameters in all demographic groups; they serve to anchor comparisons between groups. Anchor items for each comparison are those in Table 5 not flagged with DIF.

We have developed an approach to generate scores that account for DIF (Crane et al., 2006c). When DIF is found, we create new data sets as summarized in Appendix Figure 2. Items without DIF have item parameters estimated from the whole sample, while items with DIF have demographic-specific item parameters estimated.

Spurious false-positive and false-negative results may occur if the ability score (θ) used for DIF detection includes many items with DIF (Holland & Wainer, 1993). We therefore use an iterative approach for each covariate. We generate IRT scores that account for DIF, and use these as the ability score to detect DIF. If different items are identified with DIF, we repeat the process outlined in Appendix Figure 2, modifying the assignments of items based on the most recent round of DIF detection. If the same items are identified with DIF on successive rounds, we are satisfied that we identified items with DIF (as opposed to spurious findings). In the present analyses only 1–3 iterations were required for each demographic variable.

We have modified this approach for demographic categories with more than two groups (such as age, education, and ethnicity/language). Indicator terms (dummy variables) for each group are generated, and interaction terms are generated by multiplying θ by the indicator terms. All indicator terms and interaction terms are included in model 1; all indicator terms are included in model 2; and only the ability term θ is included in model 3. For the determination of nonuniform DIF, we compared the likelihoods of models 1 and 2 to a χ^2 distribution with degrees of freedom equal to the number of groups minus 1. The determination of uniform DIF is unchanged, except all of the indicator terms are included in model 2.

We performed DIF analyses in two ways. First, we analyzed DIF related to each covariate in turn. Second, we analyzed DIF related to all four covariates simultaneously. We began with unadjusted scores and analyzed items for DIF related to gender. We proceeded to analyze items for DIF related to age using the IRT score that accounted for DIF related to gender. If an item had DIF related to gender, it was analyzed separately in males and females for DIF related to age. We then analyzed items for DIF related to education and ethnicity/language using analogous steps.

In addition to item-level DIF findings, we also show the scale-level impact of accounting for DIF. We determined the median standard error of measurement. Differences in individual scores larger than the median standard error of measurement are termed "salient scale-level differential functioning." In other work we have indexed these findings to the minimally important difference established for a scale to detect "relevant scale-level differential functioning" (Crane et al., 2007); no minimally important difference has been established for the executive function scale.

We performed no adjustment for multiple comparisons in our DIF analyses. There is little cost to declaring an item has DIF using our technique—the item is still used to help determine scores, using demographic-specific item parameters as appropriate. A more thorough discussion of adjusting DIF analyses for multiple comparisons can be found in (Crane et al., 2006a). Several hundred individuals were available for DIF analyses. Sufficient overall and subgroup sample sizes for DIF detection are not known. For further discussion of this issue also refer to (Crane et al., 2006a).

Stata .do files for all of the DIF analyses are available for free download. To access the programs type "scc install difwithpar" at the Stata prompt.

https://doi.org/10.1017/S1355617708081162