Journal of Gerontology: PSYCHOLOGICAL SCIENCES
2002, Vol. 57B, No. 6, P548–P558

Education and Sex Differences in the Mini-Mental State Examination: Effects of Differential Item Functioning

Richard N. Jones[1] and Joseph J. Gallo[2]

[1]Hebrew Rehabilitation Center for Aged Research and Training Institute, Boston, Massachusetts.
[2]Department of Family Practice and Community Medicine, University of Pennsylvania, Philadelphia.

Years of completed education is a powerful correlate of performance on mental status assessment. This analysis evaluates differences in cognitive performance attributable to level of education and sex. We analyzed Mini-Mental State Examination responses from a large community sample (Epidemiologic Catchment Area study, $N = 8{,}556$), using a structural equation analytic framework grounded in item response theory. Significant sex and education group differential item functioning (DIF) were detected. Those with low education were more likely to err on the first *serial subtraction*, *spell* world *backwards*, *repeat phrase*, *write*, *name season*, and *copy design* tasks. Women were more likely to err on all *serial subtractions*, men on *spelling* and other language tasks. The magnitude of detected DIF was small. Our analyses show that failing to account for DIF results in an approximately 1.6% overestimation of the magnitude of difference in assessed cognition between high- and low-education groups. In contrast, nearly all (95%) of apparent sex differences underlying cognitive impairment are due to DIF. Therefore, item bias does not appear to be a major source of observed differences in cognitive status by educational attainment. Adjustments of total scores that eliminate education group differences are not supported by these results. Our results have implications for future research concerning education and risk for dementia.

THE correlation of education with tests of cognitive functioning may be the most robust finding in gerontologic public mental health (Bassett & Folstein, 1991; Boone, Ghaffarian, Lesser, Hill-Gutierrez, & Berman, 1993; Brayne & Calloway, 1990; Crum, Anthony, Bassett, & Folstein, 1993; Evans et al., 1993; Fillenbaum, Hughes, Heyman, George, & Blazer, 1988; Ganguli et al., 1991; Jorm, Scott, Henderson, & Kay, 1988; Magaziner, Bassett, & Hebel, 1987; Murden, McRae, Kaner, & Bucknam, 1991; O'Connor, Pollitt, Treasure, Brook, & Reiss, 1989; Wiederholt et al., 1993), and has led some to argue that education may be an important risk factor for dementia (Mortimer & Graves, 1993). The epidemiology of cognitive impairment has been facilitated by the availability of brief cognitive status assessment instruments such as the MMSE (Folstein, Folstein, & McHugh, 1975). The MMSE is a short assessment instrument that assesses orientation to time and place, attention, memory, and ability to follow commands. This article is concerned with the validity of the MMSE as a measure of cognition and with threats to its validity according to education and sex. In the context of Messick's (1989) articulation of the unitary concept of validity, this investigation can be seen as an attempt to demonstrate the extent to which the validity of the MMSE is compromised by test-irrelevant variance. Our goal is to determine the degree to which item-level performance is influenced by level of education after controlling for differences in underlying cognitive ability.

Subgroup Differences in Measured Cognitive Impairment

MMSE item-level differences by education have been reported in clinical samples (Anthony, LeResche, Niaz, von Korff, & Folstein, 1982; Murden et al., 1991; O'Connor et al., 1989) and community samples (Escobar et al., 1986; Ylikoski et al., 1992). Often, the *serial sevens*, *spelling* world *backwards*, *reading*, *writing a sentence,* and *copying polygons* tasks are more difficult for those with less education. Similarly, sex differences in cognition have been reported in educational and psychological research (Guilford, 1967). Women typically show superior performance on tests of verbal ability, and men show superiority on mathematic tests and visuospatial tasks; findings are generally consistent across age groups (Hall, Davis, Bolen, & Chia, 1999; Halpern, 1986; Resnick, 1993; Robert & Tanguay, 1990; Viaud-Delmon, Ivanenko, Berthoz, & Jouvent, 1998). Among the small number of studies of cognitive status assessment in late life explicitly concerned with sex differences, some have found no differences on the MMSE (e.g., Koivisto et al., 1992), whereas others have reported differences similar to those seen in intelligence or achievement testing (Lindal & Stefansson, 1993; O'Connor et al., 1989).

Methodological Approaches to Investigations of Subgroup Differences

With few exceptions (Marshall, Mungas, Weldon, Reed, & Haan, 1997; Teresi et al., 1995; Woodard, Auchus, Godsall, & Green, 1998), previous studies of subgroup differences in the measurement of cognitive impairment have relied on comparing the proportion correct among education groups to identify biased items. Finding one group more likely to err on a specific item might signal a problem with that item and lead to targeted approaches to addressing disturbances in measurement. However, if the two groups differ in the ability presumed to be measured by the test, differences

in error prevalence may simply record group differences in underlying ability. Further, item-to-item variance in proportion correct by group may record differences in how discriminating items are for different levels of underlying ability. Attempts to compare individuals from different groups matched on ability beg the question of identifying comparable members of subgroups. Item response theory (IRT) attempts to address these problems.

IRT

IRT, also referred to as latent trait theory, was developed by educational researchers (Lord, 1952; Rasch, 1960) and has been used in a variety of applied health research settings (Gallo, Anthony, & Muthén, 1994; Gallo, Rabins, & Anthony, 1998; Gibbons, Clarke, VonAmmon, & Davis, 1985; Kessler & Mroczek, 1995; Kirisci, Tarter, & Hsu, 1994; Legler & Ryan, 1997; Linacre, Heinemann, Wright, Granger, & Hamilton, 1994; McHorney, Haley, & Ware, 1997; Suh & Gallo, 1997). For a complete treatment of IRT, see Lord and Novick (1968) and Hambleton, Swaminathan, and Rogers (1991). The key ideas of IRT include the *ability* of an individual responding to a test item, the *difficulty* of the test item, and the accuracy with which the item measures ability (i.e. the item *discrimination*). IRT places person parameters and test item parameters on a unified metric in a self-contained analytic framework. A continuous normal (probit) or log odds (logit) statistical model is used to define the functional form relating person ability to the likelihood of responding correctly to a test item. The shape of this item response function or item characteristic curve (ICC) is sigmoid or normal ogive. Dimensions of the ICC are determined by the item difficulty and discrimination parameters. Highly difficult items have an inflection point near the upper end of the ability distribution. Highly discriminating items provide information on a narrow range of ability and are efficient at separating individuals with underlying ability above or below the level of difficulty for that item. As a consequence of modeling item parameters and person parameters separately, the item can be statistically described in terms independent from the sample to which the test was administered.

Detecting Disturbances in Measurement With IRT

Differential item functioning (DIF) is variously referred to as item bias, item-response bias, measurement noninvariance, measurement bias, measurement disturbance, test-irrelevant variance, or factorial invariance. Educational researchers generally reserve the term *item bias* for items with statistical evidence of DIF and expert content review identifying the source of the DIF as exogenous to the construct being measured. Underlying all terms is a quantitative demonstration of variant measurement properties for members of population subgroups.

IRT-based methods for detecting DIF are superior to methods based on comparing proportion correct or mean score comparisons across groups and are reviewed by Teresi, Kleinman, and Ocepek-Welikson (2000). Briefly, a test item is considered to be free from measurement disturbance or DIF when item response functions or ICCs for two groups are equivalent. That is, individuals matched on ability from different groups should have the same probability of responding correctly to a given item (Angoff, 1993; Camilli & Shepard, 1994; Osterlind, 1983). The degree to which ICCs differ across groups is summarized with DIF statistics: One convenient summary of DIF is the area between ICCs plotted for two groups (Raju, 1988).

The IRT approach to DIF attempts to disentangle measurement noninvariance from population heterogeneity, and it represents an important development for the evaluation of measurement devices. One statistical approach to this is the multiple indicators–multiple causes, or MIMIC, model (Jöreskog & Goldberger, 1975) for dichotomous variables (Gallo et al., 1994, 1998; Muthén, 1987, 1989).

MIMIC Model Approach to DIF

The MIMIC model is a special parameterization of a general structural equation model (SEM). The MIMIC model for dichotomous dependent variables addresses the same statistical problem as the IRT model (Muthén, Kao, & Burstein, 1991; Oort, 1996; Takane & De Leeuw, 1987; Thissen, Steinberg, & Wainer, 1993), and can easily be extended to tests with a mix of dichotomous, ordinal, and continuous dependent variables. In addition, the MIMIC model has the important advantage of being able to adjust for multiple background variables. In terms of the IRT framework, the MIMIC model approximates a two-parameter IRT model with item discrimination parameters presumed to be equal across groups: Only item thresholds or difficulty parameters are allowed to vary across groups. IRT guessing parameters are not estimated. From the point of view of SEM, the MIMIC model is a confirmatory factor-analytic model for dichotomous dependent variables with covariates. Analytically, the MIMIC model is a multivariate probit regression model with latent variables. DIF is revealed by group differences in item thresholds via regressions termed direct effects. Because discrimination parameters are equivalent across groups, differences in item difficulty are directly proportional to differences in areas between ICCs for the two groups. The only difference between the MIMIC model with ordinal dependent variables and continuous dependent variables is in the interpretation of regression coefficients. Regression paths leading to ordinal dependent variables are interpreted as probit regression coefficients and describe the increase in the normal probability of the outcome per unit increase in the independent variable. Regression paths leading to continuous latent variables are interpreted as in other SEM models, with the change in the latent variable being associated with a unit change in the independent variable.

Hypotheses

The purpose of the current analysis was to examine sex and education group differences in performance on the MMSE at the item level among community-dwelling older adults, using the MIMIC model for dichotomous variables. Our primary hypothesis was that items requiring literacy or computational skill would show evidence of DIF attributable to years of completed education. Furthermore, we hypothesized that men would perform better on computational and spatial relations items and women would perform better on items tapping language skills.

Methods

Study Sample

The study sample was drawn from the National Institute of Mental Health Epidemiologic Catchment Area (ECA) study. The five sites of the ECA study included New Haven, Connecticut; Baltimore, Maryland; Raleigh–Durham, North Carolina; Los Angeles, California; and St. Louis, Missouri. The methods and sample selection procedures used in the ECA study have been reported elsewhere (Eaton & Kessler, 1985; Regier et al., 1984). The ECA study was undertaken to obtain reliable prevalence estimates of mental disorders among community-dwelling adults. At each site, investigators used survey sampling techniques to draw a representative sample of residents within official Mental Health Catchment areas (Leaf, Myers, & McEnvoy, 1991). A lay interviewer gathered standardized health and mental health symptom data (Robins, Helzer, Croughan, & Ratcliff, 1981). Sites taken together had a demographic pattern similar to that of the United States as a whole in 1980 (Leaf et al., 1991). The current project was approved by the Committee on Human Research of the Johns Hopkins University School of Public Health.

The total number of respondents interviewed in the five-site ECA program was 20,861. Excluded participants were those whose age at initial interview was younger than 50 years (n = 10,594), missing data for year of birth (n = 28), not living at home (n = 1,294), did not at least start the MMSE (n = 326), or missing data for educational attainment (n = 63). Characteristics for the final sample of 8,556 are summarized in Table 1.

Table 1. Respondent Characteristics, Sample for Analysis: Five-Site Collaborative Epidemiologic Catchment Area (ECA) Study, 1981–1984

Respondent Characteristic	*n*	%
Total	8,556	100.0
ECA Site		
Yale University, New Haven, CT	3,063	35.8
Johns Hopkins University, Baltimore, MD	1,531	17.9
Washington University, St. Louis, MO	1,082	12.6
Duke University, Durham, NC	2,031	23.7
University of California, Los Angeles	849	9.9
Age Group at Baseline Interview (years)		
50–54	870	10.2
55–59	1,010	11.8
60–64	1,297	15.2
65–69	1,928	22.5
70–74	1,470	17.2
75–80	1,189	13.9
80–84	450	5.3
85+	342	4.0
Self-described Ethnicity		
American Indian or Alaska Native	73	0.9
Asian or Pacific Islander	44	0.5
Black or African American	1,582	18.5
Hispanic	342	4.0
White	6,450	75.4
Other	17	0.2
Not specified	48	0.6
Sex		
Women	5,273	61.6
Men	3,283	38.4
Level of Education (highest grade completed)		
None	112	1.3
1–5 (elementary school)	862	10.1
6–7 (some middle school)	1,000	11.7
8 (completed middle school)	1,388	16.2
9–11 (some high school)	1,653	19.3
12 (completed high school)	1,653	19.3
Grade 13 or higher (postsecondary)	1,888	22.1

Note: N = 8,556, ages 50–98.

Variables Under Study

MMSE.—The MMSE is a short assessment instrument used to grade cognitive mental status; it assesses orientation to time and place, registration, memory, attention and concentration, praxis, constructional and language capacity, and ability to follow commands. The MMSE provides a quick and reliable quantitative assessment of an individual's cognitive state (Folstein et al., 1975). It was designed for use with hospitalized patients (Folstein et al., 1975) and is used widely in primary care (Goldschmidt, Mallin, & Still, 1983; Murden et al., 1991; Tangalos et al., 1996) and in community-based research settings (George, Landerman, Blazer, & Anthony, 1991). Each ECA study site incorporated the MMSE into the lay interview. The interrater reliability of the MMSE in the original presentation was reported as .83 (Folstein et al., 1975). Jones and Gallo (2000) presented data supporting the assumption of unidimensionality of the MMSE in this community sample. Each item of the MMSE was coded 0 if the participant responded correctly to the item and 1 if the participant did not. The effect of this coding was to collapse missing, refusal, or don't-know responses into the error response category. Responses to the *spelling* world *backwards* item were dichotomized. Responses other than *d-l-r-o-w* were counted as errors. This strategy results in an unambiguous method for scaling errors on this task that is consistent across all possible misspellings (cf. Gallo & Anthony, 1994).

Age, sex, and ethnicity.—Three age groups were included in our MIMIC models. Adults in the 50–64 age group served as the reference group for the elderly (age 65–74) and very old (age 75 and older) age groups. Sex was recorded as observed.

Self-reported ethnicity.—Information on self-reported ethnicity was obtained by asking "Would you please look at this card and give me the letter of the group that best describes your racial background?" Respondents then selected from this list: American Indian, Alaska Native, Asian, Pacific Islander, Black–Not Hispanic, Hispanic, White–Not Hispanic. We considered three ethnic groups based on this self-report—White, Black or African American, and all other racial/ethnic groups—in our examination of the distribution of education.

Education.—Education was based on participants' responses to the question "What is the highest grade in school or year of college that you completed?" Responses were assigned values ranging from 0 (indicating no formal education) to 17 (representing graduate school). It should be noted that this method of ascertaining educational attainment is

not wholly consistent with methods used by the National Center for Health Statistics National Health Interview Survey (Aday, 1989) and may result in overestimates of years of completed education. We are unaware of evidence that this method of ascertaining educational attainment performs differentially according to birth cohort, sex, or ethnicity.

A number of alternatives to express education as an analytic variable were considered, motivated by discomfort with the assumption that estimated regression effects are constant across grade levels (implicit when treating education as a continuous variable). Splitting participants into groups on the basis of the sample's mean years of completed education, or on the basis of educational milestones, relies on the dubious assumption that each year of completed education or attainment of milestones implies the same education for all age, sex, and ethnic subgroups in the sample.

We used a new approach to describe educational attainment: a relativistic education indicator. This indicator was created by stratifying the sample by age at initial interview (grouped by ages 50–54, 55–59, 60–64, 65–69, 70–74, 75–79, 80–84, 85+), sex, and self-described ethnicity (White, Black or African American, and other). Within each of the resulting 48 strata, participants with fewer than the median years of completed education were classified as having low education. Five-year intervals were chosen, as finer age strata would result in sparsely populated strata and unstable ranks. This indicator removes the influence of cohort factors (age, sex, ethnicity) in grouping respondents into high- and low-education groups.

A comparison of the performance of the relativistic educational attainment indicator to an indicator split at the overall sample mean (Grade 10) is reported elsewhere (Jones, 1997; Jones & Gallo, 2001). Briefly, the relativistic indicator is not correlated with age (Spearman's $\rho = -0.01, p = .64$), but a mean-split indicator is ($\rho = 0.22, p < .001$). The relativistic indicator is not associated with being Black or African American ($\rho = -0.01, p = .46$) but the mean-split indicator is ($\rho = 0.19, p < .001$). In addition, the relativistic educational attainment indicator and the mean-split educational attainment indicator are comparably associated with cognitive performance after partialing out the effects of age, sex, and ethnicity (standardized regression coefficients, βs $= -0.30$ and -0.32, respectively).

Analytic Approach: The MIMIC Model

Consistent with the modern SEM approach to dichotomous dependent variables (Muthén, 1978; Muthén & Lehman, 1985), our model specified a threshold model for the observed dichotomous MMSE items (y). The likelihood of responding incorrectly to a MMSE item was modeled using multivariate probit regression. Probability of an incorrect response is viewed as a function of a latent ability or a latent trait and background variables. Background variables may influence items directly or indirectly as mediated by the underlying trait. This latent trait (η) fulfills the same role as the latent ability (θ) in the IRT model. Slopes relating the underlying trait to the item responses are probit regression coefficients (λ) and are analogous to factor loadings in confirmatory factor analysis or item-discrimination parameters in IRT. The underlying trait (in this case, cognitive impairment) is assumed to be continuously distributed and normal, conditional on the included covariates.

The latent trait is regressed on background variables (covariates, x). The corresponding linear regression parameters (γ) are referred to as indirect effects. This terminology reflects the fact that these parameters capture the increase in the likelihood of making an error on the item associated with the covariate mediated by the underlying trait. When the covariate is a dichotomous dummy indicator, these regressions can also be conceptualized as expressing mean differences in the underlying trait in line with traditional analysis of covariance models. Our model included background variables for participants older than 75 years (1 if age at interview was 75 years or more, 0 otherwise) and for participants aged between 65 and 74 years at interview (1 if true, 0 otherwise). Participants aged 50–64 formed the reference group. Our model also included indicator variables for Blacks or African Americans, leaving Whites and all other ethnicity groups in the reference group. The model also included indicators for relative educational attainment (1 = low education, 0 otherwise) and male sex. With the exception of education, all background variables were centered at their respective means in the analysis. This modeling strategy results in equivalent model fit and regression parameter estimates, with the exception of mean and threshold parameters, that can be interpreted for participants balanced at the overall sample mean on the other background variables.

DIF is detected by including regressions of the individual items on the background variables. These probit regression parameters (κ) describe differences in item difficulty associated with the background variable. These regressions are termed direct effects and describe differential item difficulty over and above that expected due to the effect of the covariate on the underlying trait. Finally, it is important to note that all of these regressions are estimated simultaneously and thus reflect independent effects conditional on the other parameters in the model.

Assessing Model Fit

Model fit was assessed with the root mean square error of approximation (RMSEA; Browne & Cudeck, 1993; Muthén & Muthén, 1998) and the comparative fit index (CFI; Bentler, 1990; Muthén & Muthén, 1998). The RMSEA provides a measure of discrepancy per model degree of freedom. The RMSEA will approach 0 as model fit improves. Browne and Cudeck recommended rejecting models with RMSEA values greater than 0.1; Hu and Benter (1998) suggested values close to 0.06 or less represent adequately fitting models. The CFI is based on the model chi-square: Values range between 0 and 1 and values greater than 0.95 are generally accepted for adequately fitting models.

MIMIC Model Building and Estimation

The procedure for building a structural probit or MIMIC model to detect DIF has been outlined by Muthén (1988). The first model freely estimated all indirect effects, with all direct effects fixed to zero. Disturbance in measurement by background variables was identified by examining the matrix of first-order derivatives of the fit function corresponding to direct effects. Large values (absolute value) indicate

that if the corresponding parameter were freely estimated, significant improvement in model fit would result. The model was then reestimated with the associated parameter freely estimated. This procedure is similar to that used in general SEM, using the modification index to inform model building (Jöreskog & Sörbom, 1996). If the resulting improvement in model fit was significant ($p < .05$) as judged by the chi-square difference test, the parameter was retained and the process repeated. Iterations were stopped when improvement in model fit was not significant. Models were estimated by using the Mplus program's weighted least squares estimator (Muthén & Muthén, 1998).

Results

Item-Level Performance

The prevalences of MMSE item errors are reported in Table 2. Many items were very easy, having an error prevalence of 1% or less. The most difficult items were the *serial subtraction*, *spell* world *backwards*, and *copy design* items. Men were more likely than women to err on the *what is the month*, *immediate recall*, *spelling backwards*, and *write a sentence* tasks. Women, on the other hand, were more likely to err on the *serial subtraction* tasks, *copy design*, and "*what county . . .* ?" items. Respondents with low education were more likely to err on every item. However, consistent with the IRT conceptual model, excess error prevalence is not sufficient evidence of DIF and possible item bias. To assess DIF, we carried out analyses using the MIMIC model.

MIMIC Model Results

The results of the MIMIC model are shown in Table 3. The overall model RMSEA was 0.036 and the CFI was 1.00, indicating adequate model fit. Inspection of residuals and derivatives from the model-fitting function suggested that model fit would be improved by relaxing assumptions of local independence for items with similar content. In general, patterns of misspecified null residual correlations followed the MMSE factor structure reported previously (Jones & Gallo, 2000). Because the model fit adequately and the assumption of unidimensionality is consistent with the use of MMSE as a summative score, we did not pursue these model modifications and forced unidimensionality and local independence.

Table 2. Item-Level Response Data (Proportion Failing to Respond Correctly)

	Educational Attainment		Sex		
MMSE Item	High (n = 4,790)	Low (n = 3,766)	Women (n = 5,273)	Men (n = 3,283)	Total (N = 8,556)
Orientation Items					
What is the year?	0.018	0.049	0.032	0.032	0.032
What season of the year is it?	0.048	0.096	0.067	0.071	0.069
What is the date?	0.144	0.217	0.177	0.176	0.176
What is the day of the week?	0.025	0.040	0.033	0.030	0.032
What is the month?	0.018	0.043	0.025	0.035	0.029
What state are we in?	0.014	0.040	0.030	0.018	0.025
What county?	0.047	0.078	0.074	0.040	0.061
What city or town are we in?	0.010	0.023	0.017	0.012	0.015
What floor of the building . . .?	0.006	0.014	0.011	0.009	0.010
What is this address?	0.018	0.036	0.026	0.025	0.026
Attention and Memory Items					
Can repeat "apple" immediately	0.006	0.013	0.008	0.012	0.009
Can repeat "table" immediately	0.016	0.028	0.019	0.026	0.021
Can repeat "penny" immediately	0.015	0.028	0.017	0.027	0.021
Remembered "apple"	0.109	0.172	0.128	0.152	0.137
Remembered "table"	0.255	0.337	0.277	0.313	0.291
Remembered "penny"	0.273	0.338	0.310	0.289	0.302
Can repeat "no ifs, ands, or buts"	0.183	0.315	0.230	0.259	0.241
Concentration Items					
First serial subtraction	0.194	0.416	0.345	0.206	0.292
Second serial subtraction	0.325	0.571	0.481	0.357	0.433
Third serial subtraction	0.363	0.591	0.512	0384	0.463
Fourth serial subtraction	0.370	0.615	0.534	0.389	0.478
Fifth serial subtraction	0.408	0.653	0.573	0.424	0.516
Spells "world" backwards	0.300	0.540	0.378	0.449	0.405
Language and Praxis Items					
Can name a "watch"	0.011	0.012	0.011	0.012	0.012
Can name a "pencil"	0.011	0.012	0.011	0.012	0.011
Can read and follow instruction	0.044	0.132	0.071	0.102	0.083
Takes paper in right hand	0.156	0.194	0.174	0.172	0.173
Folds paper in half	0.066	0.088	0.074	0.079	0.076
Puts paper down on lap	0.087	0.126	0.100	0.110	0.104
Writes a complete sentence	0.103	0.273	0.157	0.212	0.178
Copies drawing of two polygons	0.339	0.513	0.441	0.375	0.415

Notes: Five-site collaborative Epidemiologic Catchment Area study, 1981–1984 (N = 8,556, ages 50–98 years). MMSE = Mini-Mental State Examination.

Table 3. Results of Differential Item Functioning Analyses Using MIMIC Model

MIMIC Model Parameter	Item Loading	Threshold	Indirect and Direct Effects: Low Vs High Education, Parameter Estimate (*SE*)	Indirect and Direct Effects: Men Vs Women, Parameter Estimate (*SE*)
Indirect Effects				
Latent ability			0.55 (0.02)	0.00 (0.03)[a]
Direct Effects				
Orientation items				
What is the year?	0.74	2.42		
What season . . . ?	0.35	1.77	0.24 (0.04)	
What is the date?	0.49	1.21		
What is the day . . . ?	0.67	2.29		
What is the month?	0.80	2.46		
What state are we in?	0.77	2.42		
What county . . . ?	0.65	1.79		−0.29 (0.05)
What city or town . . . ?	0.73	2.58		
What floor of the . . . ?	0.62	2.78		
What is this address?	0.70	2.39		
Attention and memory items				
Can repeat “apple”	1.0[b]	2.71		
Can repeat “table”	0.95	2.71		0.22 (0.04)
Can repeat “penny”	0.97	2.66		0.28 (0.04)
Remembered “apple”	0.73	1.59		0.13 (0.03)
Rremembered “table”	0.60	1.12		0.17 (0.03)
Remembered “penny”	0.61	1.02	−0.08 (0.03)	
Can repeat phrase	0.53	1.24	0.15 (0.03)	0.12 (0.03)
Concentration items				
First serial subtraction	0.94	1.07	0.11 (0.02)	−0.40 (0.04)
Second serial subtraction	0.92	0.61		−0.34 (0.03)
Third serial subtraction	0.94	0.55		−0.34 (0.03)
Fourth serial subtraction	0.94	0.49		−0.39 (0.03)
Fifth serial subtraction	0.90	0.39		−0.38 (0.03)
Spells “dlrow”	0.68	0.91	0.16 (0.03)	0.23 (0.03)
Language and praxis items				
Can name a “watch”	1.01	2.46	−0.76 (0.05)	
Can name a “pencil”	1.00	2.48	−0.76 (0.05)	0.06 (0.01)
Reads and follows	0.82	2.16		0.34 (0.04)
Right hand	0.55	1.15		
Folds in half	0.84	1.81	−0.23 (0.03)	
Down on lap	0.72	1.70		
Writes a sentence	0.74	1.89	0.24 (0.03)	0.31 (0.03)
Copies drawing	0.53	0.88	0.19 (0.03)	−0.09 (0.03)

Notes: Models include adjustment for ethnicity and age. Direct effects describe excess difficulty for the indicated group; positive values signal an increase in the probability of error. Five-site collaborative Epidemiologic Catchment Area study, 1981–1984 (N = 8,556, ages 50–98 years). All parameter estimates significant at $p <$ .05 unless otherwise noted. MIMIC = multiple indicators–multiple causes.

[a]Nonsignificant, $p > .05$.

[b]Discrimination, or loading parameter fixed to 1.0 to scale of latent construct.

Measurement slopes (λ) reveal that all items were strongly related to the underlying latent trait, a reflection of the essentially unidimensional structure of the MMSE (Jones & Gallo, 2000). The least discriminating MMSE item was the orientation to place item “*what county are we in?*” The metric of the latent trait was set with the attention item *repeat “apple”* (λ = 1). Those with low education had a significantly greater mean level of cognitive dysfunction (indirect effect, γ = 0.55, *SE* = 0.22). Male and female ECA respondents did not differ in level of cognitive dysfunction (indirect effect, γ = 0.00, *SE* = 0.03).

Evidence of DIF

There was evidence of a lower likelihood of error on the *remember* “*penny*” item for those with low education. Those with low education were more likely to err on the first *serial subtraction* but not on subsequent *subtractions*. Although those with low education were more likely to err on *naming items* (Table 2), this excess error rate is lower than would be expected, due to the estimated mean level of cognitive disability for this group (Table 3). In other words, the MIMIC-model–implied mean level of cognitive dysfunction overestimates the likelihood of error on these items for the

low-education group. This was also true for the *fold paper in half* task. Those with low education were more likely to err on the *write a sentence* and *copy design* tasks.

Men were less likely to err on the *county orientation* item ($\kappa = -.29$, $SE = 0.05$). This direct effect provided the shift in the difficulty (or difference in the probability of making an error in the probit scale) for men relative to women, holding constant the effects of education, age, and ethnicity on the construct and item levels. Men were more likely to make an error on two of the three *attention* items, two *delayed recall* items, and on the *repeat phrase* item. Relative to women, men were more likely to err on the *spelling backwards* item but much less likely to err on the serial subtraction items. Men were more likely to err on the *write a sentence* item, but less likely to err on the *copy design* task. Men were also much more likely to err on the *read and follow command* tasks.

Although not summarized, the effects listed in Table 3 include control for DIF attributable to old (age 65–74) and very old (age 75 years and older) age groups (relative to those aged 50–64) and for those self-identifying as Black or African American relative to all other racial/ethnic groups. Briefly, detected direct effects suggest DIF for the *orientation to season*, *delayed recall* tasks, *naming*, *repeat phrase*, *three-step command*, *write a sentence*, and *copy polygon* items for the older age groups. The model detected DIF relative to race/ethnicity for the *registration*, *delayed recall*, *naming*, *repeat phrase*, *read and obey command*, *three-step command*, *write a sentence*, and *copy polygons* tasks. The full matrixes of direct and indirect effects, as well as residual variances, are available from Richard N. Jones on request.

DIF and Estimates of Underlying Ability

To assess the influence of DIF on estimates of the level of underlying ability, we estimated a purposefully misspecified model fixing all direct effects for education to zero (ignoring DIF). The standardized regression coefficients of latent cognitive impairment on low education can be compared between the final model and the purposefully misspecified model to obtain an estimate of the magnitude of bias that is due to educational attainment. The standardized mean latent cognitive dysfunction level for the referent group is presumed to be zero with unit variance, and the standardized regression coefficient for group membership provides the standardized difference in latent cognitive impairment for members of the indicated group. Considering first the educational attainment group differences in latent cognitive dysfunction, the standardized coefficient was .498 for the final model and .490 for the purposefully misspecified model (ignoring DIF). Thus, in both models, the low-education group has about a half a standard deviation greater level of cognitive impairment relative to the high-education group, but only about 1.6% of the difference between estimated latent cognitive dysfunction for the high- and low-education groups is due to DIF. By way of comparison, the standardized coefficient for the regression of latent cognitive impairment on male sex is $-.004$, implying no significant difference in mean cognitive dysfunction by sex (estimate/standard error [z] $= -0.16$, $p = .870$). In a purposefully misspecified model holding all sex direct effects to zero, the standardized regression coefficient was $-.081$ ($z = -4.51$, $p < .001$), implying significantly less cognitive dysfunction for men, but of very small magnitude. Thus, ignoring DIF (the misspecified model) overestimates the level of cognitive dysfunction for women by about 95%. Nearly all of the very small difference in underlying cognitive dysfunction attributable to sex is due to items that perform differently by sex.

Discussion

In general, our hypotheses were confirmed. We found DIF in the MMSE attributable to educational attainment and sex. Those with low education were more likely to err on the first *serial subtraction*, *spelling backwards*, *repeat phrase*, *write a sentence*, *name the season*, and *copy design* MMSE tasks. Men were more likely to err on language tasks, and women were more likely to err on computation and visuospatial tasks. Nevertheless, the detected education DIF has very little influence on estimated education group differences in underlying cognitive impairment. Therefore, mechanisms other than item bias must account for observed education group differences in assessed cognitive functioning.

Limitations

Before putting the results in the context of previous research, it is important to discuss the limitations of the current analysis. The relativistic education indicator addresses some of the difficulties in treating years of completed education as an analytic variable. As it is unknown how early life exposure to education influences cognition in late life, the relativistic indicator may mask a true education effect. On the other hand, raw years of education have been known to capture residual correlation of age-related constructs via the strong association of years of completed education and birth cohort (Cobb, Wolf, Au, White, & D'Agostino, 1995).

Another limitation is the problem of constant bias. Detecting DIF is straightforward when most items on a test do not show DIF. IRT approaches are not able to detect constant bias (Camilli & Shepard, 1994). Constant bias refers to the condition where many or all items on a test are equivalently biased. This situation is conceptually, but not statistically, distinct from a situation where all items are, in fact, not biased, but persons from contrasted groups differ in underlying ability. When a large number of items on a test are similarly biased, DIF analyses will overestimate group differences in underlying ability and reveal a small number of items apparently biased in favor of the minority or disadvantaged group. Interpretation of results is further complicated if group differences in underlying ability exist and a large number of items are biased. If many of the MMSE items are biased in the same way by education level, we would expect exactly the findings observed: Little evidence of item bias or small bias apparently favoring the minority or disadvantaged group and large group differences in underlying ability (Camilli & Shepard, 1994).

Previous Research on Item Bias

One previously published study examined MMSE responses for DIF according to education by using an IRT approach. Teresi and colleagues (1995) found several items in a cognitive battery (that included items from the MMSE) biased by education among participants in a dementia case

register study (n = 550). Teresi and colleagues highlighted the *orientation to state*, *repeat phrase*, *naming*, *read and follow instruction*, *three-step command,* and *write a sentence* items as biased by level of education: All but the *orientation to state* item were more difficult for respondents with low education. Our results agreed for the *write a sentence* and *repeat phrase* items. Unlike Teresi and colleagues, our results suggest DIF for *naming* and the *three-step command* item *fold paper in half* that favors respondents with low education. Our results do not replicate evidence of DIF for the *state orientation* item and detected DIF for the first *serial subtraction*, *spelling*, *orientation to season*, and *copy polygon* items. Although there are many reasons why results of Teresi and colleagues and our own differ, the commonalities deserve mention. The *write a sentence* task was flagged as possibly biased in both studies, and it is plausible to suspect this item of education bias given the manifest content. This finding suggests DIF analyses are capable of identifying items that may need to be dropped, revised, or scored differently for individuals with varying levels of education. Given the results of our investigation of the global impact of measurement bias using purposefully misspecified MIMIC models, however, a recommendation to drop this item is not supported.

There are important differences between the work of Teresi and colleagues (1995) and the analyses we present that could result in variant findings. Firstly, Teresi and colleagues analyzed a set of cognitive assessment items that included tasks from the MMSE, the Blessed (Blessed, Tomlinson, & Roth, 1968), the Short Portable Mental Status Questionnaire (Pfeiffer, 1975), the Kahn-Goldfarb Mental Status Questionnaire (Kahn, Goldfarb, Pollack, & Peck, 1960), the Caregiver Assessment and Referral Evaluation Mental Status Questionnaire (Golden, Teresi, & Gurland, 1984), and the Caregiver Assessment and Referral Evaluation diagnostic scale (Golden, Teresi, & Gurland, 1983). This design feature is an important advantage of the Teresi and colleagues study. If the array of non-MMSE items were not biased, the analysis with the larger item set would have more power to detect bias in MMSE items. However, Teresi and colleagues found that many of the non-MMSE items demonstrated DIF attributable to education. Therefore, the longer item set may also suffer from the problem of constant bias to the same extent that the MMSE suffers from constant bias. Second, Teresi and colleagues fit a three-parameter IRT model (guessing, discrimination, and difficulty parameters) using the LOGIST software package. In contrast, we fit a two-parameter IRT model (discrimination, difficulty) with item-discrimination parameters assumed to be equal across groups. Teresi and colleagues found most guessing parameters to be close to zero, so this does not seem to be an important difference between our studies. However, Teresi and colleagues allowed discrimination and difficulty parameters to differ across groups, whereas we allowed only difficulty parameters to vary across groups. An adequate discussion of the strengths and limitations of these two parameterizations of item-bias detection models is beyond the scope of this report (but see Hambleton, Wright, Crocker, Masters, & van der Linden, 1992). This difference in parameterization could lead to differences in results. Finally, our study concerned a relatively large (N = 8,550), geographically diverse, and representative community sample of older adults. In contrast, Teresi and colleagues conducted their analyses with a much smaller sample (N = 550) of older adults, about a third of whom had been referred to a memory disorder clinic for neurological diagnosis, and with approximately equal numbers of White, Black or African American, and Hispanic participants. Thus, Teresi and colleagues' sample is not representative of community-dwelling older adults, and their low-education group (defined as completion of 8 or fewer years of formal education) was disproportionately represented by minorities. Thus, DIF detected for low education in Teresi and colleagues' work may be due to ethnicity (or language, culture) rather than education. These differences highlight the advantage of the MIMIC model (multiple background variables can be included) and the relativistic educational attainment indicator (reduce confounding by age or ethnicity).

Perhaps the most intriguing finding of our analyses—a finding that cannot be elicited from previous IRT-based research on the MMSE that derives from our use of the MIMIC model—is the finding that there are at best minor measurement disturbances in MMSE by level of education. Detected DIF is not sufficient to account for education group differences in overall level of estimated cognitive ability. Furthermore, the magnitude of detected DIF by level of education is small relative to the magnitude of effects by sex.

Sex differences detected in how the MMSE measures cognition were strong and to a large extent predicted by neuroendocrine and neuropsychological research. The detection of specific sex differences is important: It highlights that the MMSE is sensitive to group differences expected by theory and that these item-level differences can be detected with the MIMIC model approach. The specific items for which potential item bias may be at work support a theory that the tasks draw on activities learned in early schooling such as spelling, writing, and familiarity with paper-and-pencil tasks. A very important feature separates the importance of DIF attributable to sex and DIF attributable to education. This is that researchers and clinicians often adjust total scores in an attempt to attenuate or remove education-level differences in scores derived from the MMSE (and other instruments) under the assumption that the test is biased. Such adjustments or modifications according to sex are unknown to us. However, researchers often use the *serial subtraction* or *spell* world *backwards* item interchangeably or one or the other exclusively. Using *spell* world *backwards* instead of *serial subtractions* would systematically overestimate the level of cognitive impairment for men. If the research question includes demonstrating sex differences in cognitive function or in the prevalence or incidence of dementia and the MMSE is used as a screening device, researchers should give some consideration to item choice and should perhaps include both the *serial subtraction* and the *spell* world *backwards* items.

Conclusion

Compelling hypotheses or explanations for the association of education and cognitive function in late life include (a) the brain reserve hypothesis (higher education confers

richer neuronal or dendritic density, providing resistance to loss associated with aging; Katzman, 1993), (b) the accelerated (cognitive) aging hypothesis (individuals from socioeconomically disadvantaged groups age at a faster pace), (c) reverse causation (i.e., low intelligence causes lower educational attainment and poor cognitive performance in late life), (d) that overall differences are due to item (Anthony et al., 1982) or test bias (O'Connor, Pollitt, & Treasure, 1991; group differences reflect differences in skills or capacities conferred by education), or (e) that education is either a surrogate for unmeasured lifestyle risk or neuroprotective factors (Orell & Sahakian, 1995; Pitt, 1993) or those with high premorbid intelligence (as evidenced by high academic attainment) experience cognitive decline that is not detected by the available crude mental status assessment instruments (Slater & Roth, 1969).

Our analyses evaluate the test or item bias hypothesis. Our findings reveal that although item bias or DIF attributable to education level was detected, it accounted for a very small fraction of the overall group difference in assessed cognition. Our results set limits on the extent to which item bias influences overall education group differences on assessed cognition. After the presence of DIF was adjusted for, large group differences in underlying cognitive impairment across educational attainment group remained. Furthermore, not only was the magnitude DIF attributable to education not as great as that observed for sex, but the sex DIF was responsible for most of the difference between the genders on assessed cognitive functioning.

As for the other hypotheses or explanations for the association of education and assessed cognitive function in late life, our conclusion that item bias accounts for little of the observed association encourages additional research into the other explanations. Thus, our findings have implications for researchers investigating the role of education in cognitive impairment, cognitive decline, and risk of neurological disease. After DIF was adjusted for DIF, considerable education group differences in underlying cognitive functioning remained. Thus, other mechanism(s) must account for education group differences in assessed cognitive function. Researchers concerned with these hypotheses may appreciate our evidence that measurement bias accounts for only a small portion of observed group differences in assessed cognitive function. Furthermore, our results argue that ad hoc MMSE total score adjustment methods that completely remove education group differences in MMSE scores (e.g., Kittner et al., 1986; Mungas, Marshall, Weldon, Haan, & Reed, 1996) overadjust and replace bias in assessed cognition favoring those with high educational attainment with bias favoring those with lower educational attainment. Also, efforts to remove the influence of education by dropping items (cf. Bazargan, Baker, & Bazargan, 2001) were not supported by the results of this psychometric analysis. The failure of DIF and possible item bias to account for much of the difference in assessed cognitive functioning may be one reason why such total score adjustments fail to substantially improve neuropsychiatric case identification activities (see Belle et al., 1996, for an applied example and Kraemer, Moritz, & Yesavage, 1998, for a review). The failure of detected DIF to account for more than a small fraction of the difference in assessed cognition between high- and low-education groups may simply be evidence that the construct measured bears only a weak link to late-life cognitive functioning. The test may more strongly measure academic achievement and familiarity with the testing situation (cf. Flynn, 1987). If this is true, then no item-level, ad hoc, or item-omission modification will improve the validity of the mental status assessment device. Research using neuropsychological batteries and assessment devices relevant to the daily cognitive functioning of older adults in large, representative, and longitudinal samples are needed to further our understanding of how socioeconomic status and early life experiences influence cognitive functioning in late life.

Acknowledgments

We are grateful for the helpful comments of Jeanne Teresi, EdD, PhD, on an earlier version of this article. Dr. Richard N. Jones's work on this project was supported in part by National Institute of Mental Health (NIMH) Training Grant in Psychiatric Epidemiology No. 5T32MH14592-22 at the Johns Hopkins University Department of Mental Hygiene (Professor William W. Eaton, principal investigator) and by National Institutes of Health Grant AG17680 (Richard N. Jones, principal investigator). Dr. Gallo is a Brookdale National Fellow in Geriatrics. Data gathering was supported by the Epidemiologic Catchment Area program of NIMH Division of Biometry and Epidemiology. The principal investigators and grant award numbers during data gathering were Jerome K. Myers at Yale University (MH34224), Morton Kramer at Johns Hopkins University (MH33870), Lee N. Robins at Washington University (MH33883), Dan Blazer and Linda George at Duke University (MH35386), and Richard Hough and Marvin Karno at the University of California, Los Angeles (MH35865). At NIMH, principal collaborators during data gathering were Darrel A. Regier, Ben Z. Locke, William W. Eaton, Carl A. Taupe, and Jack D. Burke, Jr.

Address correspondence to Richard N. Jones, ScD, HRCA Research and Training Institute, 1200 Centre Street, Boston, MA 02131. E-mail: jones@mail.hrca.harvard.edu

References

Aday, L. A. (1989). *Designing and conducting health surveys: A comprehensive guide*. San Francisco: Jossey-Bass.

Angoff, W. H. (1993). *Perspectives on differential item functioning methodology*. Hillsdale, NJ: Erlbaum.

Anthony, J. C., LeResche, L., Niaz, U., von Korff, M. R., & Folstein, M. F. (1982). Limits of the "Mini-Mental State" as a screening test for dementia and delirium among hospital patients. *Psychological Medicine, 12*, 397–408.

Bassett, S. S., & Folstein, M. F. (1991). Cognitive impairment and functional disability in the absence of psychiatric diagnosis. *Psychological Medicine, 21*(1), 77–84.

Bazargan, M., Baker, R. S., & Bazargan, S. (2001). Sensory impairments and subjective well-being among aged African-American persons. *Journal of Gerontology: Psychological Sciences, 56B*, P268–P278.

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. *Neuroepidemiology, 15*, 321–329.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*, 238–246.

Blessed, G., Tomlinson, B., & Roth, M. (1968). The association between quantitative measures of dementia and of senile change in the cerebral grey matter of elderly subjects. *British Journal of Psychiatry, 114*, 797–811.

Boone, K. B., Ghaffarian, S., Lesser, I. M., Hill-Gutierrez, E., & Berman, N. G. (1993). Wisconsin Card Sorting Test performance in healthy, older adults: Relationship to age, sex, education, and IQ. *Journal of Clinical Psychology, 49*, 54–60.

Brayne, C., & Calloway, P. (1990). The association of education and socioeconomic status with the Mini Mental State Examination and the clinical diagnosis of dementia in elderly people. *Age & Ageing, 19*, 91–96.

Browne, M., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. Bollen & J. Long (Eds.), *Testing structural equation models* (pp. 136–162). Thousand Oaks, CA: Sage.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items*. Newbury Park, CA: Sage.

Cobb, J., Wolf, P., Au, R., White, R., & D'Agostino, R. (1995). The effect of education on the incidence of dementia and Alzheimer's disease in the Framingham study. *Neurology, 45*, 1707–1712.

Crum, R. M., Anthony, J. C., Bassett, S. S., & Folstein, M. F. (1993). Population-based norms for the Mini-Mental State Examination by age and educational level. *Journal of the American Medical Association, 269*, 2386–2391.

Eaton, W. W., & Kessler, L. G. (1985). *Epidemiologic field methods in psychiatry: The NIMH Epidemiologic Catchment Area program*. New York: Academic Press.

Escobar, J. I., Burnam, A., Karno, M., Forsythe, A., Landsverk, J., & Golding, J. M. (1986). Use of the Mini-Mental State Examination (MMSE) in a community population of mixed ethnicity: Cultural and linguistic artifacts. *Journal of Nervous and Mental Disease, 174*, 607–614.

Evans, D. A., Beckett, L. A., Albert, M. S., Hebert, L. E., Scherr, P. A., Funkenstein, H. H., et al. (1993). Level of education and change in cognitive function in a community population of older persons. *Annals of Epidemiology, 3*, 71–77.

Fillenbaum, G. G., Hughes, D. C., Heyman, A., George, L. K., & Blazer, D. G. (1988). Relationship of health and demographic characteristics to Mini-Mental State Examination score among community residents. *Psychological Medicine, 18*, 719–726.

Flynn, J. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101*, 171–191.

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). "Mini-mental state:" A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research, 12*, 189–198.

Gallo, J. J., & Anthony, J. C. (1994). Re: A scoring error in the Mini-Mental State test. *Canadian Journal of Psychiatry, 39*, 384–385.

Gallo, J. J., Anthony, J. C., & Muthén, B. O. (1994). Age differences in the symptoms of depression: A latent trait analysis. *Journal of Gerontology: Psychological Sciences, 49*, P251–P264.

Gallo, J. J., Rabins, P. V., & Anthony, J. C. (1998). Sadness in older persons: 13-year follow-up of a community sample in Baltimore, Maryland. *Psychological Medicine, 29*, 341–350.

Ganguli, M., Ratcliff, G., Huff, F. J., Belle, S., Kancel, M. J., Fischer, L., et al. (1991). Effects of age, gender, and education on cognitive tests in a rural elderly community sample: Norms from the Monongahela Valley Independent Elders Survey. *Neuroepidemiology, 10*, 42–52.

George, L., Landerman, R., Blazer, D., & Anthony, J. (1991). Cognitive Impairment. In L. Robins & D. Regier (Eds.), *Psychiatric disorders in America* (pp. 291–327). New York: Free Press.

Gibbons, R. D., Clarke, D. C., VonAmmon, C. S., & Davis, J. M. (1985). Application of modern psychometric theory in psychiatric research. *Journal of Psychiatric Research, 19*, 43–55.

Golden, R., Teresi, J., & Gurland, B. (1983). Detection of dementia and depression cases with the Comprehensive Assessment and Referral Evaluation interview schedule. *International Journal of Aging and Human Development, 16*, 242–254.

Golden, R. R., Teresi, J. A., & Gurland, B. J. (1984). Development of indicator scales for the Comprehensive Assessment and Referral Evaluation (CARE) interview schedule. *Journal of Gerontology, 39*, 138–146.

Goldschmidt, T. J., Mallin, R., & Still, C. N. (1983). Recognition of cognitive impairment in primary care outpatients. *Southern Medical Journal, 76*, 1264–1265, 1270.

Guilford, J. P. (1967). *The nature of human intelligence*. New York: McGraw-Hill.

Hall, C. W., Davis, N. B., Bolen, L. M., & Chia, R. (1999). Gender and racial differences in mathematical performance. *Journal of Social Psychology, 139*, 677–689.

Halpern, D. (1986). *Sex differences in cognitive abilities*. Hillsdale, NJ: Erlbaum.

Hambleton, R., Wright, B., Crocker, L., Masters, G., & van der Linden, W. (1992). IRT in the 1990s: Which models work best. *Rasch Measurement Transactions, 6*, 215–217.

Hambleton, R. K., Swaminathan, H., & Rogers, H. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hu, L., & Benter, P. (1998). Fit indices in covariance structure analysis: Sensitivity to underparameterized model misspecifications. *Psychological Methods, 3*, 424–453.

Jones, R., & Gallo, J. (2001). Education bias in the Mini-Mental State Examination. *International Psychogeriatrics, 13*, 299–310.

Jones, R. N. (1997). *Construct validity and item analysis of the Mini-Mental State Examination by level of educational attainment and sex*. Unpublished doctoral dissertation, Johns Hopkins University. Baltimore, MD.

Jones, R. N., & Gallo, J. J. (2000). Dimensions of the Mini-Mental State Examination among community dwelling older adults. *Psychological Medicine, 30*, 605–618.

Jöreskog, K., & Goldberger, A. (1975). Estimation of a model of multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association, 10*, 631–639.

Jöreskog, K., & Sörbom, D. (1996). LISREL 8: User's reference guide. Chicago: Scientific Software International.

Jorm, A. F., Scott, R., Henderson, A. S., & Kay, D. W. (1988). Educational level differences on the Mini-Mental State: The role of test bias. *Psychological Medicine, 18*, 727–731.

Kahn, R., Goldfarb, A., Pollack, M., & Peck, A. (1960). Brief objective measures for the determination of mental status in the aged. *American Journal of Psychiatry, 117*, 326–328.

Katzman, R. (1993). Education and the prevalence of dementia and Alzheimer's disease. *Neurology, 43*, 13–20.

Kessler, R. C., & Mroczek, D. K. (1995). Measuring the effects of medical interventions. *Medical Care, 33*(Suppl.), AS109–AS119.

Kirisci, L., Tarter, R. E., & Hsu, T. C. (1994). Fitting a two-parameter logistic item response model to clarify the psychometric properties of the Drug Use Screening Inventory for adolescent alcohol and drug abusers. *Alcoholism Clinical & Experimental Research, 18*, 1335–1341.

Kittner, S., White, L., Farmer, M., Wolz, M., Kaplan, E., Moes, E., et al. (1986). Methodological issues in screening for dementia: The problem of education adjustment. *Journal of Chronic Disease, 39*, 163–170.

Koivisto, K., Helkala, E. L., Reinikainen, K. J., Hanninen, T., Mykkanen, L., Laakso, M., et al. (1992). Population-based dementia screening program in Kuopio: The effect of education, age, and sex on brief neuropsychological tests. *Journal of Geriatric Psychiatry and Neurology, 5*, 162–171.

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.

Leaf, P., Myers, J., & McEnvoy, L. (1991). Procedures used in the Epidemiologic Catchment Area study. In L. N. Robins & D. A. Regier (Eds.), *Psychiatric disorders in America* (pp. 11–32). New York: Free Press.

Legler, J., & Ryan, L. (1997). Latent variable models for teratogenesis using multiple binary outcomes. *Journal of the American Statisical Association, 92*, 13–20.

Linacre, J. M., Heinemann, A. W., Wright, B. D., Granger, C. V., & Hamilton, B. B. (1994). The structure and stability of the Functional Independence Measure. *Archives of Physical Medicine & Rehabilitation, 75*, 127–132.

Lindal, E., & Stefansson, J. G. (1993). Mini-Mental State Examination scores: Gender and lifetime psychiatric disorders. *Psychological Reports, 72*, 631–641.

Lord, F. (1952). A theory of test scores. *Psychometric Monographs, 7*, 1–84.

Lord, F., & Novick, M. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Magaziner, J., Bassett, S. S., & Hebel, J. R. (1987). Predicting performance on the Mini-Mental State Examination: Use of age- and education-specific equations. *Journal of the American Geriatrics Society, 35*, 996–1000.

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.

McHorney, C. A., Haley, S. M., & Ware, J. E., Jr. (1997). Evaluation of the MOS SF-36 Physical Functioning Scale (PF-10): II. Comparison of relative precision using Likert and Rasch scoring methods. *Journal of Clinical Epidemiology, 50*, 451–461.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). London: Collier Macmillan.

Mortimer, J., & Graves, A. (1993). Education and other socioeconomic determinants of dementia and Alzheimer's disease. *Neurology, 43*(Suppl. 4), S39–S44.

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.

Murden, R. A., McRae, T. D., Kaner, S., & Bucknam, M. E. (1991). Mini-Mental State exam scores vary with education in Blacks and Whites. *Journal of the American Geriatrics Society, 39*, 149–155.

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika, 43*, 551–560.

Muthén, B. (1987). *LISCOMP: Analysis of linear structural equations using a comprehensive measurement model. A program for advanced research*. Moorseville, IN: Scientific Software.

Muthén, B. (1988). Some uses of structural equation modeling in validity studies: Extending IRT to external variables. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 213–238). Hillsdale, NJ: Erlbaum.

Muthén, B., & Lehman, J. (1985). Multiple group IRT modeling: Applications to item bias analysis. *Journal of Educational Statistics, 10*, 133–142.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations: Meetings of Psychometric Society (1989, Los Angeles, California, and Leuven, Belgium). *Psychometrika, 54*, 557–585.

Muthén, B. O., Kao, C. F., & Burstein, L. (1991). Instructionally sensitive psychometrics: Application of a new IRT-based detection technique to mathematics achievement test items. *Journal of Educational Measurement, 28*, 1–22.

Muthén, L. K., & Muthén, B. O. (1998). *Mplus user's guide*. (Version 2). Los Angeles, CA: Author.

O'Connor, D. W., Pollitt, P. A., & Treasure, F. P. (1991). The influence of education and social class on the diagnosis of dementia in a community population. *Psychological Medicine, 21*, 219–224.

O'Connor, D. W., Pollitt, P. A., Treasure, F. P., Brook, C. P., & Reiss, B. B. (1989). The influence of education, social class and sex on Mini-Mental State scores. *Psychological Medicine, 19*, 771–776.

Oort, F. (1996). *Using restricted factor analysis in test construction*. Amsterdam, The Netherlands: Faculteit der Psychologie, Universiteit van Amsterdam.

Orell, M., & Sahakian, B. (1995). Education and dementia. *British Medical Journal, 310*, 951–952.

Osterlind, S. J. (1983). *Test item bias*. Beverly Hills, CA: Sage.

Pfeiffer, E. (1975). A short portable mental status questionnaire for the assessment of organic brain deficit in elderly patients. *Journal of the American Geriatrics Society, 23*, 433–441.

Pitt, B. (1993). Social factors and old age. In D. Bhughra & J. Leff (Eds.), *Social psychiatry* (pp. 315–330). Oxford, England: Blackwell Scientific Publications.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53*, 495–502.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute of Educational Research.

Regier, D. A., Myers, J. K., Kramer, M., Robins, L. N., Blazer, D. G., Hough, R. L., et al. (1984). The NIMH Epidemiologic Catchment Area program: Historical context, major objectives, and study population characteristics. *Archives of General Psychiatry, 41*, 934–941.

Resnick, S. M. (1993). Sex differences in mental rotations: An effect of time limits? *Brain & Cognition, 21*, 71–79.

Robert, M., & Tanguay, M. (1990). Perception and representation of the Euclidean coordinates in mature and elderly men and women. *Experimental Aging Research, 16*, 123–131.

Robins, L. N., Helzer, J. E., Croughan, J., & Ratcliff, K. S. (1981). National Institute of Mental Health Diagnostic Interview Schedule. Its history, characteristics, and validity. *Archives of General Psychiatry, 38*, 381–389.

Slater, E., & Roth, M. (1969). *Clinical psychiatry*. London: Balliere Tindall.

Suh, T., & Gallo, J. J. (1997). Symptom profiles of depression among general medical service users compared with specialty mental health service users. *Psychological Medicine, 27*, 1051–1063.

Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*, 393–408.

Tangalos, E. G., Smith, G. E., Ivnik, R. J., Petersen, R. C., Kokmen, E., Kurland, L. T., et al. (1996). The Mini-Mental State Examination in general medical practice: Clinical utility and acceptance. *Mayo Clinic Proceedings, 71*, 829–837.

Teresi, J., Golden, 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.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–113). Hillsdale, NJ: Erlbaum.

Viaud-Delmon, I., Ivanenko, Y., Berthoz, A., & Jouvent, R. (1998). Sex, lies and virtual reality. *Nature Neuroscience, 1*(1), 15–16.

Wiederholt, W. C., Cahn, D., Butters, N. M., Salmon, D. P., Kritz-Silverstein, D., & Barrett-Connor, E. (1993). Effects of age, gender and education on selected neuropsychological tests in an elderly community cohort. *Journal of the American Geriatrics Society, 41*, 639–647.

Woodard, J. L., Auchus, A. P., Godsall, R. E., & Green, R. C. (1998). An analysis of test bias and differential item functioning due to race on the Mattis Dementia Rating Scale. *Journal of Gerontology: Psychological Sciences, 53B*, P370–P374.

Ylikoski, R., Erkinjuntti, T., Sulkava, R., Juva, K., Tilvis, R., & Valvanne, J. (1992). Correction for age, education and other demographic variables in the use of the Mini Mental State Examination in Finland. *Acta Neurologica Scandinavica, 85*, 391–396.

Received May 12, 2000
Accepted February 7, 2002