Journal of the International Neuropsychological Society (2007), **13**, 228–236.

DOI: 10.1017/S1355617707070257

Do reading tests measure the same construct in multiethnic and multilingual older persons?

STEPHANIE COSENTINO,[1] JENNIFER MANLY,[1,2,3] AND DAN MUNGAS[4,5]

[1]Gertrude H. Sergievsky Center, Columbia University Medical Center, New York, New York
[2]Taub Center for Alzheimer's Disease Research, Columbia University Medical Center, New York, New York
[3]Department of Neurology, Columbia University Medical Center, New York, New York
[4]Department of Neurology, School of Medicine, University of California-Davis, Sacramento, California
[5]Psychology Service, Veterans Affairs, Northern California Health Care System, Martinez, California

(RECEIVED April 10, 2006; FINAL REVISION August 18, 2006; ACCEPTED August 21, 2006)

Abstract

A critical focus of neuropsychological research is to identify unbiased ways to compare heterogeneous groups on background variables relevant to neuropsychological performance. Whereas recent work has pointed to single word reading as a less culturally biased measure of educational experience than years of education, the extent to which reading score captures a broad range of educational experience and does so consistently across ethnic and language groups is unknown. The current study evaluated reading in relation to years of education in English-speaking Whites, Blacks, and Hispanics, and Spanish-speaking Hispanic older persons (n = 342). Consistent with previous work, reading scores at each grade level were lower in English speaking ethnic minorities than in Whites, supporting the idea that variables related to lifetime educational experience are often confounded with ethnicity. Standardized reading scores were highest in the Spanish speakers; however, interpretation of this difference is limited because scores were necessarily derived using separate normative samples. Importantly, the slopes of reading score by years of education were comparable across all groups. That is, reading scores rose consistently with years of education independently of ethnicity or language, suggesting that such scores can be treated comparably for theoretical and statistical purposes in multiethnic and multilingual samples.
(*JINS*, 2007, *13*, 228–236.)

Keywords: Neuropsychological testing, Ethnicity, English, Spanish, Culture, Language

INTRODUCTION

Valid interpretation of neuropsychological performance depends largely on the extent to which raw test scores can be evaluated in the context of an individual's personal characteristics and background. Toward this end, scores are generally compared to those achieved by a cognitively healthy peer group matched on relevant demographic variables such as age, gender, and educational experience. In particular, years of education has proven to be closely associated with performance on a wide range of verbal and nonverbal neuropsychological measures, accounting for a good deal of variability in scores across healthy individuals (Heaton et al., 1996; Howeison et al., 2004). Incorporating years of education into score standardization has offered neuropsychologists greater sensitivity and specificity in detecting clinically significant impairment. For example, normative data may indicate that a given raw score reflects deficient performance in an individual with 20 years of education as compared to average performance for someone with eight years of formal schooling.

Whereas this interpretive process generally facilitates examination of within-group differences, complications arise in making comparisons across heterogeneous groups that differ significantly in socioeconomic, cultural, or educational characteristics. Namely, studies have shown that in cognitively healthy samples matched for years of education, Blacks achieve significantly lower scores than Whites on many traditional neuropsychological measures (Lichtenberg et al., 1994; Manly et al., 1998a). Such findings indicate that total number of educational years does not capture educational experience in a consistent manner across groups.

Correspondence and reprint requests to: Jennifer Manly, Ph.D., Gertrude H. Serviegsky Center, Columbia University Medical Center, 630 West 168th Street, P&S Mailbox 16, New York, NY 10032. E-mail: jjm71@columbia.edu

In fact, accumulating research has now quite clearly outlined the striking differences in the quality of education delivered in ethnic minority *versus* primarily White school systems, such that one year of education likely represents different levels of instruction and learning across multiethnic groups, particularly in older persons (Anderson, 1988; Coleman, 1966). Adjusting raw scores for years of education alone introduces systematic bias into multiethnic research because ethnic minority groups have traditionally been at a disadvantage in terms of student-teacher ratio, per pupil expenditures, and length of school year.

A critical focus of neuropsychological research has thus been to search for relatively unbiased ways in which to equate heterogeneous groups on background variables, such as educational experience, which are relevant to neuropsychological performance. In this vein, a number of studies have pointed to reading ability as a less culturally-biased estimate of educational experience given its correlation with variables such as student-teacher ratio, per pupil expenditures, and length of school year (Hanushek, 1989; Hedges et al., 1994; O'Neill, 1990). Indeed, studies have shown that reading scores in older minority persons fall below those of older White persons matched for years of education (Boekamp et al., 1995; Manly et al., 2002). Most importantly, ethnicity effects on most neuropsychological measures disappear after accounting for reading scores in older African American persons (Manly et al., 2002).

Increasingly, researchers view performance on single word reading tests as the most valid available estimate of educational experience in multiethnic samples. Reading scores are often included as independent variables or covariates in statistical analyses evaluating cognitive performance across ethnic groups. This is a particularly important step in the development of psychometrically equivalent tests across cultural groups or in the estimation of premorbid intellectual functioning critical for detecting cognitive decline associated with dementia in older persons. The underlying assumption when implementing reading scores to represent educational quality is that such scores adequately capture the full of range of educational experience, and that they do so consistently across multiethnic and multilingual groups; however, this hypothesis has not been examined directly. This is a challenging issue to explore, both conceptually and methodologically, because an ideal approach would be to evaluate reading scores in each group against a separate and consistent "gold standard" for educational experience. Theoretically, such a standard is unattainable given the developmental, domestic, socioeconomic, and cultural factors that vary across individuals with divergent ethnic backgrounds.

The challenge of matching groups against a consistent educational gold standard grows even more conceptually and methodologically complex when working with multilingual samples. Single word reading in English speakers is assessed with tests such as the American version of the North American Reading Test (AmNART) (Grober & Sliwinski, 1991) and the Wide Range Achievement Test-3 (WRAT-3) (Wilkinson, 1993). These measures evaluate an individual's ability to pronounce a series of phonologically regular and irregular words ranging in difficulty, and are based on the idea that correct pronunciation of the more difficult items requires previous exposure to such words. Because the Spanish language contains only phonologically regular words (i.e., there is a one-to-one correspondence between orthography and phonology), an entirely analogous test is unattainable. Rather, use of accentuation in the Spanish language can provide the opportunity to evaluate subjects' familiarity with words by removing the appropriate accentuation, as is seen on the Word Accentuation Test (WAT) (Del Ser et al., 1997).

Scores on English and Spanish single word reading tests have been shown to correlate with traditional measures of intellectual functioning and years of education in cognitively healthy older persons (Crawford et al., 1989; Del Ser et al., 1997; Grober & Sliwinski, 1991), and are believed to represent similar constructs across English and Spanish speakers. However, there are methodological hurdles to combining multilingual reading scores in a single analysis for use in a cross-cultural sample. In particular, comparison and compilation of standardized scores across tests are problematic because such scores are necessarily derived from two different populations. Whereas an ideal solution may be to administer the WAT and AmNART to bilingual individuals to devise a formula for converting scores from one test to the other, the validity of this approach depends on the extent to which individuals in the validation sample are equally fluent in both languages, or the extent to which fluency in each language can be accurately measured.

Research in multiethnic and multilingual samples therefore poses a number of challenging issues, one of which is the unbiased assessment of complex constructs such as educational experience across groups. Because it may be unrealistic to solve this problem entirely, there are several steps that can be taken to reduce systematic errors in cross-cultural assessment and research. As described earlier, reading scores have been identified as relatively unbiased estimates of educational experience across ethnic and language groups. A next step in evaluating the validity of reading scores is to determine that such scores adequately represent a broad range of educational experience, and that they do so in a consistent fashion across ethnic and language groups. Theoretically speaking, reading scores should be expected to rise as a function of years of education, and the increase in reading score with each additional year of education should be comparable across ethnicity or language groups. However, it is possible that differences in educational quality create an interaction effect between years of education and reading level as a function of ethnicity or language. For example, it is possible that reading scores plateau after a certain grade level in one group, while continuing to rise with years of education in another. The presence of interaction effects would raise the possibility that reading scores do not represent the same construct, either conceptually or methodologically, across ethnic and/or language groups.

Based on previous work indicating that ethnicity is frequently confounded with variables related to lifetime educational experience, we hypothesized that Blacks and Hispanic English speaking older persons would achieve significantly lower AmNART scores than older White persons with similar years of education. However, we predicted that years of education and reading level would be positively correlated to a similar extent in Blacks, Hispanics, and Whites and that no interaction effect would exist as a function of ethnic group. This study also sought to evaluate the extent to which the relationship between reading score and education was consistent across English and Spanish speaking older persons. We could not make predictions regarding mean level of performance given that reading was measured with different instruments, however, we hypothesized that reading score and years of education would be positively correlated to a similar extent in both groups, and that no interaction effect would exist as function of language group.

METHODS

Subjects

Participants were 342 older persons who received neuropsychological testing as part of the development process for the Spanish and English Neuropsychological Assessment Scales (SENAS). Participants were recruited *via* a community screening program designed to identify and recruit individuals with a broad range of cognitive functioning ranging from normal to demented. Recruitment was designed to target ethnic minorities, to maximize heterogeneity of demographic characteristics, and to emphasize normal cognition and mild impairment. Inclusion criteria were age 60 or over, and ability to speak either English or Spanish. Individuals were tested in either language according to their preference. Those tested in English included 77 Whites, 114 Blacks, and 33 Hispanics. There were 118 Hispanics tested in Spanish. Thirty-one percent of participants (n = 105) received a clinical evaluation including a detailed medical history, physical and neurological exam, and clinical neuropsychological assessment. A bilingual physician examined Spanish-speaking patients and neuropsychological tests were administered to Spanish speakers by fully bilingual psychometrists. The Clinical Dementia Rating (CDR) was completed based on a structured interview and examination by raters who had completed online training and certification developed by Dr. John Morris at Washington University. A family member or informant in close contact with the participant was interviewed to obtain information about level of independent functioning. Diagnostic neuroimaging and routine dementia work-up laboratory tests were a standard part of the protocol. The final diagnosis was the consensus decision established at a multidisciplinary case conference. A previous study showed that neuropsychological test results were equally related to clinical diagnosis in Whites and Hispanics in this sample (Mungas et al., 2005).

Although one approach to subject selection for the current study would be to include only those older persons with cognitive functioning in the non-impaired range, we could not systematically exclude participants with dementia because only a subsample was clinically evaluated. Further, prior studies of the WAT (Del Ser et al., 1997) and NART (Crawford et al., 1989; Nelson & O'Connell, 1978) have demonstrated comparable reading scores across healthy controls and patients with mild AD, and such scores have been shown to remain stable over time in the context of cognitive decline associated with mild AD (Hart et al., 1986; O'Carroll & Gilleard, 1986; Sharpe & O'Carroll, 1991). However, to be sure that the presence of dementia in one group did not systematically bias our findings, we analyzed the proportion of individuals with this diagnosis across ethnic and language group.

Exclusion criteria 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 last five years. All participants signed informed consent under protocols approved by institutional review boards at UC Davis, the Veterans Administration Northern California Health Care System, and San Joaquin General Hospital in Stockton, California. All data were collected in compliance with the regulations of these institutions' internal review boards.

Reading measures

English reading level was measured with the American version of the Nelson Adult Reading Test (AmNART; Grober & Sliwinski, 1991). This measure includes 45 words with irregular spelling to sound correspondence, whose proper reading would depend on previous knowledge or exposure to the words. Subjects are instructed to read each word aloud. Items are presented in order of increasing difficulty based on the number of subjects in the validation sample (n = 230) who read each word correctly. The authors reported a Cronbach's α coefficient of internal consistency of .93 and found the AmNART to be useful as an estimate of premorbid intellectual functioning among patients with mild to moderate dementia.

Spanish Reading level was measured using the 30-item WAT (Del Ser et al., 1997). This measure was designed to be equivalent to English language measures of reading recognition such as the NART (Nelson & O'Connell, 1978; Nelson, 1982), which consists of words with an irregular pronunciation whose proper reading would depend on previous knowledge or exposure to the words. The authors of the WAT, developed a measure in which the reader is confronted with an ambiguous graphic clue: infrequent, irregularly stressed words written in capital letters without their accent marks. This format allows for some correspondence with the NART, because correct pronunciation depends on previous knowledge of the words. The authors

https://doi.org/10.1017/S1355617707070257 Published online by Cambridge University Press

of the WAT found a Cronbach's α coefficient of internal consistency of .91 and found the measure useful in estimating premorbid intellectual functioning among demented patients.

Data analysis

A multiple group covariance structure analysis with mean structures was performed using Mplus (Muthen & Muthen, 1998, 2004) to evaluate the relationship between years of education and reading ability in the four ethnicity/language groups: Blacks (B), Whites (W), Hispanics tested in English (HE), and Hispanics tested in Spanish (HS). This essentially involved a series of multiple group regression analyses in which reading score was the dependent variable, education was the independent variable, and regression slopes and intercepts were systematically constrained to be either equal across groups, or were allowed to freely vary. In the baseline analyses, regression slopes and intercepts were constrained to be equal. The next model constrained slopes but allowed intercepts to freely vary across groups. The third constrained intercepts, but allowed slopes to vary, and in the final model, both slopes and intercepts were freely estimated within language/ethnicity groups. Variances were freely estimated in all models.

A maximum likelihood estimator (Muthen & Muthen, 1998, 2004) was used to evaluate model fit; that is, to test the extent to which the observed within-group means and covariances were reproduced by the various models. Model fit was assessed with the model χ^2 test, the standardized root mean squared residual (SRMR; Bentler, 1995; Joreskog & Sorbom, 1981), the root mean squared error of approximation (RMSEA; Browne & Cudek, 1993; Cudek & Browne, 1983), the comparative fit index (CFI; Bentler, 1990), and the non-normed fit index (NNFI; Bentler & Bonnett, 1980; Tucker & Lewis, 1973). The χ^2 test is known to be highly sensitive to sample size, and can result in rejection of closely fitting models in large samples. As a result, a number of fit indices have been proposed to better evaluate model fit. SRMR supplemented by CFI and/or NNFI were recommended by Hu and Bentler (1998) based on a simulation study that evaluated a number of different fit indices and tested their relative sensitivity to model misspecification under conditions involving different sample sizes and deviations from distributional assumptions. MacCallum & Austin (2000) recommended the RMSEA as a valuable addition to these fit indices. Generally, CFI and NNFI values of .95 and higher are considered to indicate good fit, RMSEA values of .05 to .06 and lower indicate good fit, whereas values between .06 and .08 indicate adequate fit, and SRMR values under .08 indicate good fit.

Relative fit of nested models was evaluated using the difference in the χ^2 test values from the two models, which is distributed as χ^2 test with degrees of freedom equal to the difference in degrees of freedom associated with the two models. Two models are nested when both include the same parameters, but parameters that are freely estimated in one model are constrained in the nested model.

A combined reading measure (READ) was created that enabled simultaneous modeling of reading in English and Spanish. The mean and standard deviation of the AmNART was calculated for all the participants tested in English who reported that they were able to read English, and the mean and standard deviation of the WAT were calculated for those tested in Spanish who reported they could read Spanish. These values were used to create standard scores [(reading score − mean)/SD] that could then be combined in subsequent analyses. This variable consisted of the AmNART based standard scores for the B, W, and HE groups and the WAT based standard scores for the HS group. Whereas the mean and SD for the AmNART may not strictly have the same significance as these values for the WAT, this approach nevertheless allows for evaluating if a 1.0 SD difference in the WAT has the same relationship to education as a 1.0 SD difference in the AmNART.

RESULTS

Table 1 shows demographic characteristics of the participant sample by ethnic group and language. Table 2 describes the language background of the Hispanic participants. Of the 105 participants who received a clinical evaluation, 13%

Table 1. Demographic characteristics by language of test administration and ethnicity

Language	Ethnicity	*N*	% Female	Years of Education *M* (*SD*) [Range]	Age *M* (*SD*)	Reading Score *M* (*SD*) [Range]
English	Whites	77	62.3	14.1 (3.2) [7–21]	72.5 (6.8)	31.6 (10.2) [1–45]
English	Blacks	114	71.9	13.1 (2.6) [3–18]	72.5 (7.1)	21.4 (10.3) [0–43]
English	Hispanic	33	63.6	11.5 (3.5) [0–18]	68.5 (6.4)	19.6 (13.4) [0–44]
Spanish	Hispanic	118	72.9	5.1 (4.5) [0–20]	69.4 (7.0)	16.7 (9.7) [0–30]

https://doi.org/10.1017/S1355617707070257 Published online by Cambridge University Press

Table 2. Language characteristics of English speaking and Spanish speaking Hispanics

	Educated in US	Spanish as first language	Bilingual
English	96.9%	72.7%	84.8%
Spanish	2.7%	100%	11.9%

were diagnosed with dementia; 13 cases were mild (CDR ≤ 1), and one was moderate (CDR = 2). The proportion of individuals diagnosed with dementia did not vary as a function of ethnic or language group, $\chi^2(3, N = 105) = 3.82$, $p = .28$, including 22% of Whites (4 of 14), 10% of Blacks (4 of 37), 0% of Hispanic English speakers (0 of 11), and 17% of Hispanic Spanish speakers (6 of 29) who were evaluated.

Fit indices for the separate multiple group Covariance Structure Analyses (CSA) are presented in Table 3. (The fit of the model where all parameters are freely estimated is not presented since it is, by definition, perfect). The baseline model in which both regression slopes and intercepts were constrained to be equal across groups was poorly fitting by all fit indices. The model in which intercepts were freely estimated but slopes were constrained to equality fit very well by all indices, but in contrast, model fit was marginal when intercepts were constrained to equality and slopes were freely estimated (SRMR and CFI indicated adequate fit, but $\chi^2(p < .05)$, NNFI, and RMSEA were sub-optimal). These results indicate that the intercepts of the regressions of reading on education significantly differ across groups, but the slopes are the same. These results are presented in Figure 1, which shows the model derived regression lines for each group based on the final CSA model in which both intercepts and slopes were freely estimated. Figure 1 shows the similarity of slopes across groups. Intercepts for the B and HE groups were very similar, but the intercepts for W and HS were higher.

Secondary analyses were performed to better characterize the pattern of group differences in intercepts. Slopes were constrained to equality in these analyses. First, intercepts for W and HS were freely estimated but intercepts for B and HE were constrained to be the same. All fit indices indicated good model fit [$\chi^2(4) = 0.7$, $p > .94$, SRMR = .02, CFI = 1.00, NNFI = 1.03, RMSEA = .00], which can

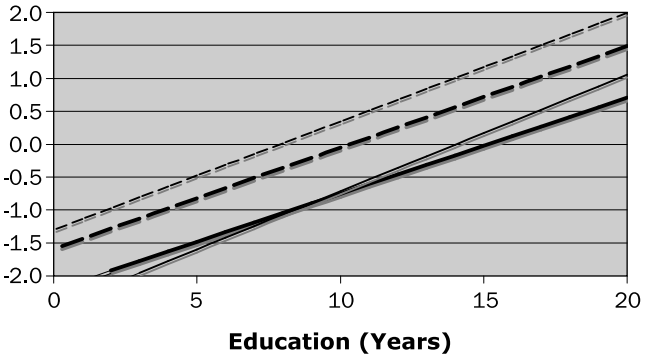

Fig. 1. Model derived regressions of reading score on education in ethnicity/language groups.
In order of highest to lowest *y*-intercept:
Small dotted line: Hispanic speaking Spanish
Large dotted line: English speaking Whites
Thick solid line: English speaking Blacks
Thin solid line: English speaking Hispanic

be interpreted as showing that the intercepts for B and HE did not significantly differ. This served as a baseline model that could be used to evaluate change in model fit associated with constraining intercepts for the W and HS groups. When the intercepts for W, B, and HE were constrained to equality, the overall model fit was poor [$\chi^2(5) = 39.1$, $p < .0001$, SRMR = 0.18, CFI = .67, NNFI = .74, RMSEA = .19] and the χ^2 difference test comparing this model with the baseline model was highly significant [$\chi^2(1) = 38.4$, $p < .0001$]. Results were very similar when the intercepts for HS, B, and HE were constrained to equality; overall model fit was poor [$\chi^2(1) = 41.2$, $p < .0001$, SRMR = .14, CFI = .65, NNFI = .72, RMSEA = .19] and the χ^2 difference test was highly significant (χ^2 [1] = 40.4, $p < .0001$). These results verify that the intercepts for the W and HS groups were significantly higher than those for the B and W groups. In the final analysis, the intercepts for B and HE were constrained to be equal, and the intercepts for W and HS were also constrained to equality. Overall model fit was good [$\chi^2(5) = 5.0$, $p > .40$, SRMR = .05, CFI = 1.00, NNFI = 1.00, RMSEA = .01] but the χ^2 difference test comparing this model with the baseline model was significant [$\chi^2(1) = 4.3$, $p < .05$], suggesting that the intercept was higher for the HS group.

The difference between the HS intercept and the intercepts from the W, B, and HE groups is of unclear signifi-

Table 3. Fit of multiple group covariance structure analysis models

Model	χ^2(DF)	SRMR	CFI	NNFI	RMSEA
Slopes equal intercepts equal	68.2 (6)	.19	.40	.60	.23
Slopes equal intercepts free	.4 (3)	.02	1.00	1.03	.00
Slopes free intercepts equal	7.9 (3)	.06	.95	.94	.09

Note. Values of SRMR under .08, RMSEA under .06, and CFI and NNFI over .95 are considered to indicate good fit.

https://doi.org/10.1017/S1355617707070257 Published online by Cambridge University Press

cance. Literally, this indicates that 0 years of education in HS is associated with a higher reading level relative to the overall distribution of reading in Spanish speakers than is 0 years of education in the other groups relative to the distribution of reading in English speakers. Further studies are required to verify that these metrics are strictly equivalent. The comparisons among the three English-speaking groups are of greater direct interest. As can be seen in Figure 1, the reading level associated with 10 years of education in the W group is about the same as that associated with 14 to 15 years of education in the B and HE groups.

DISCUSSION

Years of education has proven to be an integral consideration in interpreting neuropsychological test performance because it generally accounts for significant variability in performance within cognitively healthy individuals (Heaton et al., 1996; Howeison et al., 2004). Increasingly, however, research involving multiethnic and multilingual samples has challenged neuropsychologists to reconsider traditional means of incorporating educational background into test interpretation as striking differences in the educational quality across Whites and ethnic minorities, particularly in older persons, result in systematic bias when adjusting only for years of education in cross-cultural samples (Kaufman et al., 1997; Loewenstein et al., 1994; Whitfield & Baker-Thomas, 1999). Whereas single word reading has been established as a less culturally biased proxy for educational experience (Hedges et al., 1994; Manly et al., 2002), the degree to which reading score captures a broad range of educational experience and does so in a consistent fashion across groups has not been established.

The current study examined whether reading scores rise predictably with each additional year of education in a multiethnic and multilingual group of older persons, or whether this relationship varies as a function of ethnicity or language. Consistent with previous work, current results highlighted discrepant reading scores across Whites, Blacks and older Hispanic English speaking persons with equivalent years of formal education. That is, at each grade level, Whites achieved significantly higher reading scores than Blacks and Hispanic English speakers. For example, average reading scores achieved by Whites at 10 years of education were roughly equivalent to those achieved by both ethnic minority groups at 15 years of education.

The intercept difference across the groups has several potential interpretations that are not mutually exclusive. There are variables confounded with ethnicity and language that are important to consider, particularly those that relate to socioeconomic status (SES). Although we did not formally measure SES or occupation, it is very likely that the White and ethnic minority older persons in our study differed in these regards (Smith, 1984; Smith & Welch, 1977). Economic disadvantages often result in larger teacher-student ratios and lower per pupil expenditures, variables that have been shown to correlate with reading achievement (Hedges et al., 1994), and may have impacted current findings. It is also possible that early environmental factors related to SES such as access to health care and community resources (e.g., libraries), and exposure to informal educational experiences in the home differed systematically between minorities and non-minorities, placing the former at a disadvantage during formal schooling. Similarly, different degrees of exposure to reading in adulthood, likely as a function of occupation, may have contributed to intercept differences in reading scores across ethnicity.

Additionally, the individuals in this study come from diverse cultural backgrounds, another difference which may have contributed to discrepant reading performance. The effects of culture on cognitive testing can be multifold (Ardila, 2005; Geertz, 2000). At a general level, culture specific values and beliefs might influence perceptions of the testing environment, the purpose and value of the research, and the significance of the reading evaluation. Existing work has shown that the degree of acculturation within minority populations impacts performance on cognitive tests such that individuals who are more assimilated into the majority culture demonstrate higher cognitive scores (Manly et al., 1998b). It is also possible that test specific qualities, such as the words included in the AmNART, are tailored to reading experiences of the dominant culture and thus might systematically bias this test as a measure of reading ability in minorities. This hypothesis would be directly testable using modern psychometric methods to evaluate item-level differential item function (DIF) and its impact on the overall reading score. This is an important area for further study and one that we plan to address specifically with data from this project.

Despite discordant reading scores across English speaking groups matched for total years of education, however, an important finding is that the magnitude of differences (slope) did not vary by grade level. That is, we found a similar linear relationship between reading scores and years of education in Whites, African Americans, and Hispanics such that approximately four years of additional education were associated with a 0.5 SD improvement in reading scores in all three groups. Equivalent reading score slopes across minorities and non-minorities suggest that despite the discrepancy in scores, which exists at every grade, each additional year of schooling provides an equivalent improvement in reading scores across ethnic groups. That is, the factors that lead to lower reading scores in the ethnic minority groups do not appear to have a cumulative effect over time that would result in a larger discrepancy at the higher grades.

The second goal of this study was to determine whether or not reading scores rise consistently with years of education across English and Spanish speakers. Essentially, the same pattern emerged as when the analyses were constricted to English speakers, such that WAT and AmNART scores increased to a similar extent with each additional year of education. Results support the use of reading scores as estimates of educational quality in multiethnic and multi-

lingual groups, regardless of the number of years that an individual attended formal schooling. Importantly, the absence of interaction effects as a function of ethnicity or language group also allows for the combination of reading score and years of education within a single statistical analysis.

Whereas WAT scores were significantly higher than AmNart scores at each grade level, the true import of this finding is unclear. Because the two tests were normalized based on performance within the Spanish and English speaking groups, respectively, the distributions of scores are not directly comparable. There were very few English speakers in this study with less than eight years of formal education. The fact that the Spanish speaking sample had significantly fewer years of education than the English speaking group raises one possible interpretation of the data. A high proportion of subjects attaining very low raw scores on the WAT would in effect shift the distribution of standardized scores to the right, resulting in a higher reading *z*-score at zero years of education in the Spanish speakers than in the English speaking groups. It is also possible that the two reading tests measure different aspects of reading ability, or differ in their overall level of difficulty, and additional research is certainly needed to explore these issues. For practical purposes, however, differences in the *y*-intercept across the WAT and AmNART can be adjusted to accommodate statistical analysis.

In sum, the broad goal of this study was to evaluate the extent to which reading scores measure the same construct in multilingual and multiethnic older persons. Consistent with previous work, intercept differences across the groups suggest that variables which are frequently confounded with ethnicity, seem to impact lifetime educational experience, and reading tests reflect an aspect of this educational experience not captured by total years of education. Despite differences in reading scores at each grade level, the magnitude of differences from one grade level to the next was linear across all years of education and comparable across ethnic and language groups. Documenting this pattern of scores across each group is an important step in establishing that reading tests measure similar constructs in multiethnic and multilingual groups, and can be treated comparably in statistical analyses required for cross cultural work.

The cross-sectional design of the current study limits interpretation of the relationship between reading scores and years of education in all groups. Because the same individuals were not followed over time, we cannot comment on the rate of reading score increase with additional education in a given individual. That is, those whose highest grade achieved was 6 years did not go onto achieve 7 or 12 years of school. Thus, average reading scores per grade level are defined by individuals who subsequently stopped attending school at that level, an issue that may effectively lower the average reading score per grade when compared to a longitudinal design. That is, reading scores at the 6th grade level in a sample of individuals who ultimately obtained a high school education might be higher than those seen in a cross-sectional sample, because the former group might experience more success in an academic environment. However, we should not assume that mean scores in a longitudinal design would rise consistently in each ethnic or language group. Reading score slopes might vary across groups because of a cumulative effect of differences in educational quality. It is also possible that the factors that lead an individual to leave school after a particular grade level differ across ethnic and linguistic groups; this might also result in discrepant reading score slopes across ethnic groups that could only be observed in a longitudinal design.

Nonetheless, the current study establishes an important aspect of between-group psychometric equivalence for reading scores, a finding which facilitates the study of more sophisticated questions examining educational experience as it relates to cognitive functioning in multiethnic and multilingual samples. In particular, the degree to which years of education and reading score differentially impact aspects of cognition such as episodic memory, executive functioning, visuospatial skills, and semantic knowledge in healthy and clinical populations is unclear. Moreover, it remains to be determined whether or not such relationships vary as a function of ethnicity or language group, an issue that we are examining in a follow up study.

Although 4% of our sample carried a dementia diagnosis, we could not systematically examine their reading scores in comparison to those of older cognitively healthy persons because only 30% of all participants underwent clinical evaluation. It is possible that reading scores in participants with dementia lowered the mean scores in our sample (Patterson et al., 1994; Storandt et al., 1995; Taylor et al., 1996); however, the majority of dementia cases in our sample were mild (13 of 14), and many studies have suggested that single word reading is preserved in early dementia (Crawford et al., 1989; Del Ser et al., 1997; Hart et al., 1986; Nelson & O'Connell, 1978; O'Carroll & Gilleard, 1986; Sharpe & O'Carroll, 1991). Importantly, the proportion of individuals with dementia in our sample did not differ by ethnicity or language, suggesting that any potential effect of cognitive impairment on the current data would not vary across group. Further, it does not seem that either the WAT or AmNART was differentially sensitive to the presence of cognitive impairment, because an interaction between reading and education was not present across language group. In the future, it will be important to clarify the way in which reading score relates to years of education and cognitive abilities over the course of dementia in multiethnic and multilingual groups.

ACKNOWLEDGMENTS

This work was supported by the National Institute on Aging (AG10220, AG10129, AG00261-07) and by a grant from the Alzheimer's Association for the 2005 Friday Harbor Psychometrics Workshop.

REFERENCES

Anderson, J.D. (1988). *The education of blacks in the South, 1960–1935*. Chapel Hill, NC: University of North Carolina Press.

Ardila, A. (2005). Cultural values underlying psychometric cognitive testing. *Neuropsychology Review*, *15*, 185–195.

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

Bentler, P.M. (1995). *Eqs structural equations program manual*. Encino, CA: Multivariate Software.

Bentler, P.M. & Bonnett, D.G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*, 588–606.

Boekamp, J.R., Strauss, M.E., & Adams, N. (1995). Estimating premorbid intelligence in African-American and White elderly veterans using the American version of the North American Reading Test. *Journal of Clinical and Experimental Neuropsychology*, *17*, 645–653.

Browne, M. & Cudek, R. (1993). Alternate ways of assessing model fit. In K. Bollen & J. Long (Eds.), *Testing structural equation models*. Thousand Oaks, CA: Sage.

Coleman, J. (1966). *Equality of educational opportunity*. Washington, D.C.: Government Printing Office.

Crawford, J.R., Parker, D.M., Stewart, L.E., Besson, J.A.O., & De Lacey, G. (1989). Prediction of WAIS IQ with the National Adult Reading Test: Cross-validation and extension. *British Journal of Psychology*, *27*, 181–182.

Cudek, R. & Browne, M.W. (1983). Cross-validation of covariance structures. *Multivariate Behavioral Research*, *18*, 147–167.

Del Ser, T., Gonzalez-Montalvo, J., Martinez-Espinosa, S., Delgado-Villapalos, C., & Bermejo, F. (1997). Estimation of premorbid intelligence in Spanish people with the Word Accentuation Test and its application to the diagnosis of dementia. *Brain and Cognition*, *33*, 343–356.

Geertz, C. (2000). *Interpretation of Culture*. New York: Basic Books.

Grober, E. & Sliwinski, M. (1991). Development and validation of a model for estimating premorbid verbal intelligence in the elderly. *Journal of Clinical and Experimental Neuropsychology*, *13*, 933–949.

Hanushek, E. (1989). The impact of differential expenditures on school performance. *Educational Researcher*, *18*, 45–51.

Hart, S., Smith, C.M., & Swash, M. (1986). Assessing intellectual deterioration. *British Journal of Clinical Psychology*, *25*, 119–124.

Heaton, R.K., Ryan, L., Grant, I., & Matthews, C.G. (1996). Demographic influences on neuropsychological test performance. In I. Grant & K.M. Adams (Eds.), *Neuropsychological assessment of neuropsychiatric disorders* (2nd ed.). New York: Oxford University Press.

Hedges, L.V., Laine, R.D., & Greenwald, R. (1994). Does money matter? A meta-analysis of studies of the effects of differential school inputs on student outcomes. *Educational Researcher*, *23*, 5–14.

Howeison, D.B., Loring, D.W., & Hannay, H.J. (2004). Neurobehavioral Variables and diagnostic issues. In M.D. Lezak, D.B. Howeison, & D.W. Loring (Eds.), *Neuropsychological Assessment*. New York: Oxford University Press.

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

Joreskog, K.G. & Sorbom, D. (1981). *Lisrel v: Analysis of linear structural relationships by the method of maximum likelihood*. Chicago, IL: National Educational Resources.

Kaufman, J.S., Cooper, R.S., & McGee, D.L. (1997). Socioeconomic status and health in blacks and whites: The problem of residual confounding and resilience of race. *Epidemiology*, *8*, 621–628.

Lichtenberg, P.A., Ross, T., & Christensen, B. (1994). Preliminary normative data on the Boston Naming Test for an older urban population. *Clinical Neuropsychologist*, *8*, 109–111.

Loewenstein, D.A., Arguelles, T., Arguelles, S., & Linn-Fuentes, P. (1994). Potential culture bias in the neuropsychological assessment of the older adult. *Journal of Clinical and Experimental Neuropsychology*, *16*, 623–629.

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*. Palo Alto, CA: Annual Reviews.

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

Manly, J.J., Jacobs, D.M., Touradji, P., Small, S.A., & Stern, Y. (2002). Reading level attenuates differences in neuropsychological test performance between African American and White elders. *Journal of the International Neuropsychological Society*, *8*, 341–348.

Manly, J.J., Miller, S.W., Heaton, R.K., Byrd, D., Reilly, J., Velasquez, R.J., Saccuzzo, D.P., & Grant, I. (1998b). The effect of African-American acculturation on neuropsychological test performance in normal and HIV-positive individuals. The HIV Neurobehavioral Research Center (HNRC) Group. *Journal of the International Neuropsychological Society*, *4*, 291–302.

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

Muthen, L.K. & Muthen, B.O. (1998). *Mplus User's Guide* (2nd ed.). Los Angeles, CA: Muthen & Muthen.

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

Nelson, H.E. (1982). *The National Adult Reading Test (NART): Test Manual*. Windsor, Berks, U.K.: NFER-Nelson.

Nelson, H.E. & O'Connell, A. (1978). Dementia: The estimation of premorbid intelligence levels using the New Adult Reading Test. *Cortex*, *14*, 234–244.

O'Carroll, R.E. & Gilleard, C.J. (1986). Estimation of premorbid intelligence in dementia. *British Journal of Clinical Psychology*, *25*, 157–158.

O'Neill, J. (1990). The role of human capitol in earning differences between Black and White men. *Journal of Economic Perspectives*, *4*, 25–45.

Patterson, K., Graham, N., & Hodges, J.R. (1994). Reading in dementia of the Alzheimer's Type: A preserved ability? *Neuropsychology*, *3*, 395–407.

Sharpe, K. & O'Carroll, R.E. (1991). Estimating premorbid intellectual level in dementia using the National Adult Reading Test: A Canadian study. *British Journal of Clinical Psychology*, *30*, 381–384.

Smith, J.P. (1984). Race and human capital. *American Economic Review*, *4*, 685–698.

Smith, J.P. & Welch, F. (1977). Black-White male wage ratios: 1960–1970. *American Economic Review*, *67*, 323–328.

Storandt, M., Stone, K., & LaBarge, E. (1995). Deficits in reading performance in very mild dementia of the Alzheimer type. *Neuropsychology*, *9*, 174–176.

Taylor, K.I., Salmon, D.P., Rice, V.A., Bondi, M.W., Hill, L.R., Ernesto, C.R., & Butters, N. (1996). Longitudinal examination of American National Adult Reading Test (AMNART) performance in dementia of the Alzheimer type (DAT): Validation and correction based on degree of cognitive decline. *Journal of Clinical and Experimental Neuropsychology*, *18*, 883–891.

Tucker, L.R. & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*, 1–10.

Whitfield, K.E. & Baker-Thomas, T. (1999). Individual differences in aging minorities. *International Journal of Aging and Human Development*, *48*, 73–79.

Wilkinson, G.S. (1993). *Wide Range Achievement Test 3—Administration manual*. Wilmington, DE: Jastak Associates, Inc.