Neuropsychol Rev (2008) 18:184–193
DOI 10.1007/s11065-008-9067-9

Measurement in Cross-Cultural Neuropsychology

Otto Pedraza · Dan Mungas

Received: 29 July 2008 / Accepted: 30 July 2008 / Published online: 24 September 2008

Abstract The measurement of cognitive abilities across diverse cultural, racial, and ethnic groups has a contentious history, with broad political, legal, economic, and ethical repercussions. Advances in psychometric methods and converging scientific ideas about genetic variation afford new tools and theoretical contexts to move beyond the reflective analysis of between-group test score discrepancies. Neuropsychology is poised to benefit from these advances to cultivate a richer understanding of the factors that underlie cognitive test score disparities. To this end, the present article considers several topics relevant to the measurement of cognitive abilities across groups from diverse ancestral origins, including fairness and bias, equivalence, diagnostic validity, item response theory, and differential item functioning.

Keywords Cross-cultural neuropsychology · Measurement · Genomics · Ethnicity · Race · Differential item functioning

Although ubiquitous in statistical and scientific texts, the term *measurement* is seldom defined. In its most elemental form, measurement is the process of collecting and recording data for the purpose of estimation, analysis, and interpretation. It ranges from the minute observations and recordings made by individual scientists to the collective effort of large organizations. For a broader conceptualization, Michell (1999) surveys the history and philosophy of measurement in psychological science.

Within the realm of neuroscience and the study of social systems, definitions arguably matter most in the study of human behavior and cognition across cultures, races, and ethnicities. Scientists continue to express widespread disagreement about the terms 'race' and 'ethnicity'. In a 2003 meeting at the National Human Genome Center in Washington, D.C., a multidisciplinary group of sociologists, geneticists, anthropologists, and bioethicists gathered to discuss the state of the science regarding human genome variation and race (Royal and Dunston 2004). The contributors noted that the term 'race' is often undefined and inconsistently used depending on contextual demands (Keita et al. 2004). Colloquially, it may refer to particular cultural, national, social, ethnic, linguistic, genetic, or geographical groups of individuals, and represents a weak surrogate for underlying factors associated with health status. Moreover, skin pigmentation in the US is generally regarded as a proxy for 'race,' and variations in skin pigmentation are frequently and incorrectly understood as a reflection of deeper biological differences among populations (Parra et al. 2004). The term 'race' has a specific taxonomic definition in the natural sciences; namely, natural variations of phylogenetic subspecies with an objective degree of microevolutionary divergence (Keita et al. 2004). Yet, this degree of taxonomic diversification is not applicable to humans. When used instead to denote social or demographic groups, the term 'race' represents a social construction based upon an incorrect usage of the term. As stressed by Keita et al., demographic units in the US are not 'races'.

The degree to which 'racial' classifications are useful in the medical field also continues to be debated (Burchard

O. Pedraza (✉)
Department of Psychiatry and Psychology, Mayo Clinic,
4500 San Pablo Road,
Jacksonville, FL 32224, USA
e-mail: otto.pedraza@mayo.edu

D. Mungas
Department of Neurology, Davis School of Medicine,
University of California,
Sacramento, CA 95817, USA
e-mail: dmmungas@ucdavis.edu

et al. 2003; Cooper et al. 2003). Recent evidence suggests that knowledge about ancestral origins can facilitate our predictions about disease susceptibility and outcome (Bamshad et al. 2004). And as noted by Francis Collins, head of the Human Genome Project, human genetic variation has reasonable predictive validity with regard to geographic ancestral origins (Collins 2004). It is important to bear in mind, however, that within population differences among individuals account for 93–95% of the total genetic variation (Rosenberg et al. 2002). Knowledge that a person has genetic "African ancestry," for instance, is less clinically useful when considering the wide variation in rates of hypertension, diabetes, and other medical variables within this large continental ancestry group (Cooper et al. 2003).

In this context, the current article considers various topics relevant to the measurement of cognitive abilities across groups from diverse ancestral origins. It is understood that these groups may identify themselves as African–American, Hispanic, Caucasian, Asian–American, or other socially-constructed labels that serve as proxies for between-group variation in socioeconomic status, education, language, acculturation, health care access, and geographic ancestry, among other factors. The term 'cross-cultural' is used, therefore, as a general rubric to encompass such diversity.

Fairness and Bias

The *Standards for Educational and Psychological Testing*, published jointly by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education, defines the term 'fairness' in four principal ways (American Educational Research Association [AERA], 1999). The first two definitions are widely supported among educators and psychologists, whereas the third and fourth definitions are contentious and will be presented here only for the sake of completeness.

Bias

Fairness represents the absence of bias. A measurement tool is considered to demonstrate bias if it results in "different meanings for scores earned by members of different identifiable groups" (AERA 1999, p. 74). When a test is biased, it demonstrates systematic errors in construct or predictive validity associated with an examinee's group membership (Beller et al. 2005).

Construct Bias With regard to construct validity, the two most important threats are construct-underrepresentation and construct-irrelevant aspects of the test (Messick 1995). Construct-underrepresentation occurs when a test fails to capture important aspects of the construct being measured. It may be due to a narrow selection of test-item content and is particularly problematic when using brief instruments. Construct-irrelevance refers to nuisance factors or processes that are extraneous to the construct under study. In this case, the item content of the assessment tool is too broad. Van de Vijver and Poortinga (2005) provide an example of bias associated with nuisance or construct-irrelevant factors. They report on an item from the European Values Survey that was intended to measure loyalty. The Spanish scores on this item deviated markedly from the overall pattern of results. As they learned upon closer inspection, the Spanish word for loyalty has an additional connotation of sexual faithfulness, thus skewing the obtained result toward an irrelevant aspect of the construct.

Predictive Bias When the purpose of measurement is to predict an outcome, a test may also demonstrate bias if the regressions that relate the predictor to the criterion differ between groups. This predictive relationship can be expressed in simple linear equation of the form

$$bX + a = Y,$$

where X represents the predictor (e.g., LSAT score) and Y represents the predicted outcome (e.g., law school GPA). The slope is expressed by the term b and the y-intercept by a. Predictive bias is usually tested through moderated multiple regression, in which the criterion is regressed on the predictor score, group membership, and the interaction between those two variables. This represents the so-called Cleary model, first described by T. Anne Cleary (1968) in the context of SAT predictions of college performance between black and white students. The model is equivalent to testing the null hypothesis that there is no significant difference in the slopes and intercepts of the linear regressions between groups. Jensen (1980) expands this model further to suggest that a test should be considered a biased predictor when there is a significant between-group difference in slopes, intercepts, or standard error of estimates.

The measurement of cognitive abilities plays a prominent role in the job selection process of many organizations, in which general mental ability is arguably the single best predictor of occupational achievement and performance (Schmidt and Hunter 2004). Meta-analytic studies have shown that African–American adults in the US score approximately one standard deviation lower on measures of quantitative and verbal ability compared to Caucasian adults, and Hispanic adults score about two-thirds of a standard deviation lower than their Caucasian counterparts (e.g., Roth et al. 2001). These findings raise significant concern for academic and employment selection practices whenever cognitive testing contributes to the selection process and prediction of future performance. In a review of the effectiveness of multiple strategies for reducing these

disparities, Sackett et al.(2001) recommend rendering these disparities meaningless by the use of alternate modes of test stimuli presentation, inclusion of noncognitive predictors such as measures of personality and interpersonal skills, use of performance-based assessments, provision of test-taking orientation or coaching, enhancement of test-taking motivation, and extension of the allotted time to complete tests.

Helms (1992) calls into question the unexamined comparison and use of mental ability tests between ethnic groups. In a direct commentary on Sackett et al. (2001), Helms (2002) expresses concern about the authors' implication that cognitive test score disparities reflect irremediable cognitive deficiencies. Building on the work of Darlington (1971) on culture-fair tests, Helms argues that investigators must examine the "cultural contamination" of tests before drawing predictive conclusions. Cultural contamination is present whenever a significant correlation exists between a predictor (e.g., test score) and a cultural indicator (e.g., total score on a racial identity questionnaire) after controlling for the criterion variable. In such a scenario, the predictive validity of the cognitive test may be challenged. Unfortunately, relatively few studies of this sort exist in the psychological or neuropsychological literature.

Item Bias In addition to construct and predictive bias, item bias refers to validity threats that directly affect individual test items. Examples of these threats include ambiguous items, words or phrases that are poorly translated, or words that have different connotations across groups. It should be noted that the term 'item bias' has largely been replaced by the concept of differential item functioning, partly due to a historical sense that the term 'item bias' carried an undertone of deviation from a American-Eurocentric standard (Van de Vijver and Poortinga 2005). Item bias will be discussed in greater detail in the section on differential item functioning.

Equitable Treatment

Fairness may also refer to equitable treatment throughout the testing process. In this respect, measurement tools are not considered intrinsically fair or unfair; it is the manner in which the tool is administered that results in fairness or unfairness. Van de Vijver and Poortinga (2005) refer to this threat to fairness as method bias, which they further subdivide into instrument and administration bias. Instrument bias refers to all the properties associated with an instrument that are not the target of study, but nonetheless result in group differences in test scores. For instance, if a computer is used to measure reaction times in children from families with low versus high socioeconomic status, the differential familiarity with computers by virtue of socioeconomic status is expected to influence the obtained results, regardless of the construct being investigated. Administration bias refers to group differences in test scores due to aspects of the interaction and communication between the examiner and examinee.

Factors such as inappropriate testing conditions, unequal opportunity to familiarize oneself with the test format, unavailability of practice materials and unequal exposure to those materials, unequal performance feedback, and lack of standardized test administration may represent method bias and contribute to unfair measurement. In circumstances in which special accommodations may be necessary, for example due to disability or lack of English language proficiency, those accommodations are expected to result in more comparable test results than if standardized procedures were left unmodified. In these situations, the *Standards* indicate that accommodations would not be considered unfair treatment and in fact may be required by law.

Outcomes

Fairness may refer to comparable testing outcomes across groups. Under this interpretation, comparable rates of passing scores or endorsement responses across two or more groups are necessary for a measurement tool to demonstrate fairness. This notion is widely rejected by education and psychology professionals. Indeed, the *Standards* specify that if a test is otherwise free of bias and the examinees or respondents have been treated fairly throughout the testing process, the conditions of fairness have been met sufficiently. Unequal testing outcomes in themselves are not *de facto* indicators of unfairness, but instead should trigger closer examination of the measurement instrument for possible sources of bias.

Opportunity to Learn

Fairness can also be conceptualized within an educational achievement framework as a comparable opportunity to learn. In this context, measurement of a person's achievement in a particular psychoeducational domain may be unfair if the educational opportunities to learn the subject matter were insufficient or inadequate, thus leading to lower achievement scores than could have otherwise been obtained. This conceptualization of fairness also lacks universal agreement, partly due to the significant difficulty inherent in the operationalization and estimation of such opportunities for learning.

Measurement Equivalence and Diagnostic Validity

Clinical neuropsychologists benefit from a plethora of assessment instruments that are readily available and

thoroughly compiled in standard textbooks (e.g., Lezak et al. 2004; Strauss et al. 2006). A substantial portion of these instruments have sufficient validity and reliability to support their use in everyday clinical practice. Regrettably, the majority of these instruments have not been developed or validated among diverse cultural groups (Manly 2005), thus rendering between-group comparisons questionable. Excellent reliability and validity within a particular cultural group does not guarantee the instrument's ability to provide meaningful data across demographic strata (Liang 2002; Stewart and Napoles-Springer 2003; Teresi and Holmes 2002). To the extent that most neuropsychological measures continue to be developed using predominantly White, middle-class, and highly educated samples, their use with non-White, non-middle-class, and less educated members of our society presents a substantial hurdle for the individual neuropsychologist when choosing clinical tests and interpreting the obtained results.

Equivalence

Cross-cultural measurement at its most basic level means that the same or similar cognitive abilities are being assessed in different cultural groups. This can be accomplished in the least restrictive manner by using tests that are selected, optimized, and have norms for each individual group. Test scores can be derived in this manner using classic psychometric methods, but such scores are not directly comparable across groups and so have meaning only in the context of the group from which they were derived. There are substantial advantages to measures that are comparable across groups. Measures of this type allow for flexible, direct comparison of individuals regardless of differences that may be present in a variety of important background characteristics that vary across and within cultures.

The notion of equivalence in cross-cultural studies may be as psychometrically important as validity and reliability (Johnson 2006; Van de Vijver and Leung 1997). The concept has yet to reach full maturity, however, as considerable ambiguity and disagreement continues to exist. In a review of the literature, Johnson lists 62 terms that have been variously utilized when referring to equivalence. In the absence of a standard nomenclature, Johnson distills these terms into two broad categories: interpretive and procedural equivalence. Interpretive forms of equivalence refer to similarities in the meaning of concepts and constructs, with an emphasis on how measures are interpreted across cultural groups. Interpretive equivalence reflects *shared meaning*. In contrast, procedural forms of equivalence refer to the technical problems of cross-cultural measurement, with an emphasis on the factorial, metric, or structural aspects of the instruments. Procedural equivalence reflects *shared method*.

Procedural equivalence must be established before a measure is to be considered culturally fair (Ramirez et al. 2005). Although a complete discussion of the various methods associated with procedural equivalence is beyond the scope of the present article, the notion of factorial invariance will be presented briefly. Meredith (1993) outlined three forms of factorial invariance that serve as a useful hierarchy, such that each sequential step places additional constraints on the initial factor model (see also Meredith and Teresi 2006). Pattern or weak factorial invariance requires only for the factor loadings or regression weights to be equivalent across groups. If this form of invariance is met, then the differences among item scores can be compared meaningfully between groups (Steemkamp and Baumgartner 1998). Strong factorial invariance places an additional set of constraints, namely, that intercepts of indicator variables for factors are equal across groups. This essentially means that the expected score on the indicator variable is the same for two individuals with the same factor score. This allows for the valid comparison of group means. Finally, strict factorial invariance requires that residual variances of indicator variables be equal across groups.

An example of this approach is found in Culhane et al. (2006), who tested the Need for Cognition Scale-Short Form (NCS-SF) among 608 Hispanic- and Anglo-American college students. The authors sequentially tested confirmatory factor analytic models for weak, strong, and strict factorial invariance. Despite establishing a baseline model of comparable pattern matrices (configural invariance), two of the factor loading coefficients and at least one intercept differed between the groups. The results were interpreted to reflect partial measurement invariance between Hispanic and Anglo-American college students on the NCS-SF.

Although procedural or metric equivalence must be established before considering cultural fairness, the requirement of procedural equivalence, in of itself, does not guarantee cultural equivalence. Additional forms of equivalence (e.g., functional, conceptual, linguistic, contextual)—subsumed in Johnson's analysis under the rubric of interpretive equivalence—must also be established to demonstrate equivalence across cultural groups (Helms 1992, 1997). These forms of equivalence can be achieved in part by close scrutiny of item content, analysis of linguistic differences, representative sampling, and inclusion of "etic" or culture-neutral stimuli, in addition to "emic" or culture-specific stimuli.

Diagnostic Validity

The scarcity of neuropsychological measures with known procedural and interpretive equivalence across cultural groups is also problematic from a clinical diagnostic

standpoint. The vast majority of these measures have not been properly validated for use with ethnic minorities (Manly and Jacobs 2002; Manly 2005). And despite recent large-scale normative (e.g., Artiola i Fortuny et al. 1999; Heaton et al. 2004; Lucas et al. 2005) and battery-development efforts (e.g., Ostrosky-Solis et al. 1999), the overwhelming majority of neuropsychological tests and batteries lack sufficient normative data for African–Americans and Hispanics, the two largest ethnic minority groups in the US. It is not surprising, then, that in the absence of cultural equivalence and appropriate normative data, a sizeable proportion of ethnic minority individuals is misclassified as cognitively impaired.

Diagnostic validity refers to a measure's ability to differentiate persons with and without a specified disorder (Smith et al 2008). It encompasses indices of sensitivity, specificity, and overall classification accuracy or hit rate. For any given cutoff value, sensitivity refers to the proportion of individuals with a given condition that are correctly classified by the test to have the condition. As shown in Fig. 1, this is represented by the number of individuals in cell A, divided by the total number of individuals in cells A and C. In contrast, specificity refers to the proportion of individuals without a given condition who are correctly classified by the test not to have the condition. In Fig. 1, this is represented by the number of individuals in cell D, divided by the total number of individuals in cells B and D. Highly specific tests are useful to 'rule in' a diagnosis, whereas highly sensitive tests are useful to 'rule out' a diagnosis (Fletcher and Fletcher 2005; Smith et al. 2008). Note that this frequency distribution also yields the proportion of individuals with a given condition that are incorrectly classified by the test not to have the condition, or false negatives [C/(A + C)], as well as the proportion of individuals without a given condition who are incorrectly classified by the test to have the condition, or false positives [B/(B + D)]. The hit rate is merely the overall proportion of individuals classified correctly by the test, or [(A + D)/(A + B + C + D)]. The quantitative relationship between sensitivity and specificity,

		Condition of interest	
		Present	Absent
Test result	Positive	A	B
	Negative	C	D

Fig. 1 Diagnostic test characteristics

predictive invariance, procedural (measurement) invariance, and test bias is reviewed in Borsboom et al. (2008).

Cutoff values in clinical neuropsychology are generally derived from comparison to a normative standard (Busch et al. 2005; Heaton et. al. 2004; Mitrushina et al. 2005; Strauss et al. 2006). As a consequence, the choice of a particular normative dataset and cutoff value directly influences a test's sensitivity and specificity and, hence, diagnostic validity. Appropriate norms reduce the amount of test variance that is associated with nuisance factors, such as demographic variables, thereby increasing specificity (Smith et al. 2008). When cognitively normal ethnic minorities are incorrectly classified as impaired, these errors cast doubt on the diagnostic validity of the test. This situation is most often encountered when ethnic minorities are compared to a dissimilar normative group, resulting in a net reduction in specificity and misattribution of cognitive impairment (Ardila 1995; Ardila etal. 1994; Lucas et al. 2005). Misdiagnosis may unduly cause emotional distress on patients and family members, lead to unnecessary treatments, and potentially result in increased health care burden and costs.

For instance, Fig. 2 presents data for a variety of neuropsychological measures obtained as part of Mayo's Older Adult Normative Studies (MOANS; Ivnik et al. 1990) and Older African American Normative Studies (MOAANS; Lucas et al. 2005). Using cutoff values one standard deviation below the normative mean, 24% to 56% of cognitively normal African–American adults are incorrectly classified across tests as cognitively impaired when the cutoff values are derived from Caucasian norms. In contrast, the discrepancy in impairment rates is substantially attenuated when using cutoff values derived from a normative sample of cognitively normal African American adults. Moreover, Smith et al. (2008) have shown in a sample of demented and nondemented adults that the use of appropriate ethnicity-based norms markedly improved the specificity of selected memory tests.

Occasionally, these errors of diagnostic classification persist even when likely confounding variables such as age, education, gender, and acculturation are partialed out or controlled through matching samples. In such cases, it is possible, and perhaps likely, that additional extraneous factors account for some of the residual variation. An example was presented by Manly et al.(2002) who sought to investigate whether quality of education could explain some of the differences in neuropsychological test performance between non-demented African-American and non-Hispanic, White adults matched on years of education. Their results showed that quality of education, as measured by proxy using the Reading subtest from the Wide Range Achievement Test-3 (WRAT-3; Wilkinson 1993), attenuated the discrepancy in test scores between

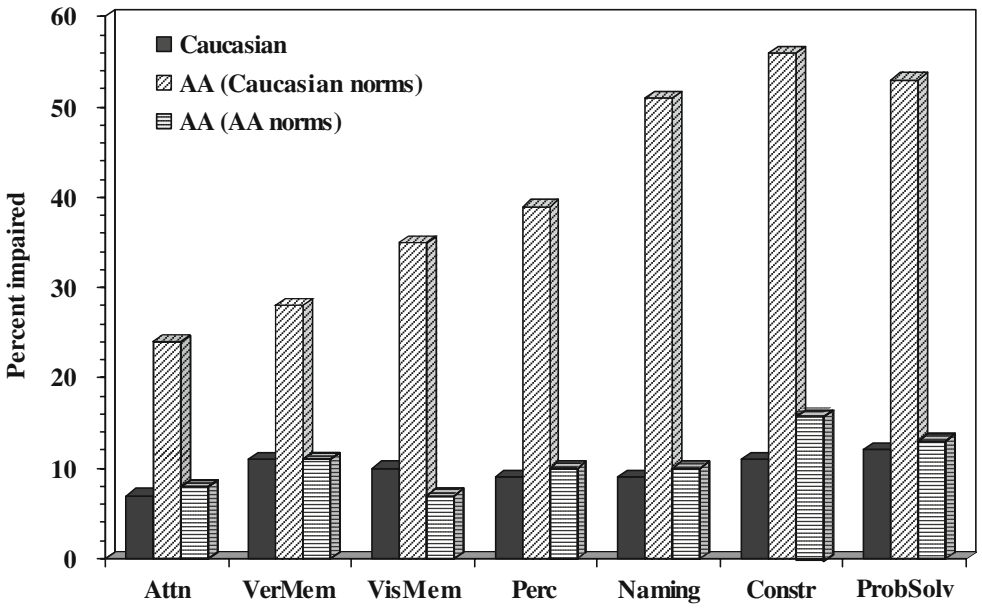

Fig. 2 Percent impairment among African American adults when using Caucasian versus African American norms. *AA* = African American, *Attn* = Attention, *VerMem* = Verbal memory, *VisMem* = Visual memory, *Perc* = Visuoperception, *Constr* = Visuoconstruction, *ProbSolv* = Problem solving. (data courtesy of John A. Lucas)

the two groups. The implication is that years of formal education may be an incomplete indicator of the educational experience of African-American adults, and that additional adjustment for quality of education can improve the specificity of neuropsychological measures in this population. More recently, the use of reading scores as a proxy for quality of education has also been supported among Spanish-speaking, Hispanic adults (Cosentino et al. 2007). Dotson et al. extend this line of investigation by presenting literacy-based (WRAT-3) normative data for African Americans age 30–64 with low socioeconomic status (Dotson et al. 2008).

The use of 'race'-specific norms, however, is not without controversy (e.g., Brandt 2007; Manly 2005; Manly and Echemendia 2007). The arguments against the proliferation of race- or ethnicity-based normative data include: (a) use of separate norms obscures the underlying factors contributing to between-group discrepancies, (b) use of separate norms validates the (incorrect) assumption that race is not socially-constructed, but instead reflects scientific and biological categories among individuals, (c) use of separate norms may promote misinterpretation and misunderstanding about the discrepancy in cognitive test scores, (d) separate norms have limited generalizability and relevance outside of the geographic and cultural milieu in which they are obtained, and (e) use of separate norms unintentionally may result in a conformist attitude toward the discrepancies in cognitive test scores among culturally diverse groups. These arguments have substantial merit, as they move the discourse away from *how much* different one group is from another, to *why* these differences exist at all. To the extent that they advance the science and practice of neuropsychological measurement, these debates appear timely and welcomed.

Modern Psychometric Theory and Differential Item Functioning

Modern psychometric methods provide important tools for the development and evaluation of cross-cultural neuropsychological tests.

Item Response Theory

Initially developed in the 1950s and progressively refined since then, IRT represents state-of-the-art methodology for psychometric test development (Embretson and Reise 2000; Hambleton and Swaminathan 1985; Hambleton et al. 1991; van der Linden and Hambleton 1997). It encompasses a mathematical theory for characterizing item and scale measurement parameters and associated numerical methods for estimating item parameters and ability of examinees.

Ability, item difficulty, and item discrimination Ability is a central concept in modern psychometric theory and refers to the capacity to respond correctly to test items. It represents the net influence of all experiential, environmental, and genetic factors that might influence test performance, without assumptions about the relative contributions of these factors. An individual with higher 'ability' simply has a higher probability of responding correctly to a given test item.

Two fundamental parameters of IRT models are item difficulty and item discrimination. For a given dichotomous item, difficulty corresponds to the ability level associated with a 50% probability of passing the item. The classic psychometric theory analog of item difficulty is the proportion of correct responses. Item discrimination refers to the degree to which small differences in ability are

associated with different probabilities of passing the item, and its classical psychometric analog is the item-total correlation. An IRT model expresses the probabilistic association between a person's observable item responses and their unobservable but estimated ability level.

One-parameter IRT models freely estimate item difficulty and require an assumption that item discrimination remains equal across items. Two-parameter models freely estimate both item difficulty and discrimination. A three-parameter IRT model can be used for multiple-choice items in which guessing might result in a correct response; the third parameter estimates the likelihood of a correct response as a result of guessing. Current IRT models can be applied to dichotomous items as well as items with multiple, ordinal response options.

Item parameters and examinee ability are estimated in an iterative process—item parameters are fixed and ability is estimated, then ability is fixed at those estimates and item parameters freely estimated, and this cycle is repeated until changes in ability estimates and item parameter estimates become small. At the end of this process, there will be estimates of item parameters for each item and ability estimates for each examinee. The ability estimate for an individual examinee is a mathematical function of all the item responses and the item parameters. Item difficulty and examinee ability are expressed on the same scale, and this aspect has interpretive significance.

The estimated item parameters in an IRT model define an item characteristic curve (ICC), a non-linear function relating the probability of passing the item to the person's ability level. On an ICC, difficulty is represented by the location along the x-axis at which point the probability of a correct response for a binary item is 50%, and discrimination is represented by the slope of the trace line at that location parameter. A steeper slope reflects a higher degree of discrimination. The ICCs for all individual items in a scale define the test characteristic curve, a non-linear function relating the expected total score to ability.

Item parameters also define a second test-level summary function, the test information curve, which quantifies scale reliability at each point of the ability continuum. The concept of information corresponds to fidelity of measurement, so that the standard error of measurement at a specific ability value is the inverse square root of the information at that ability (Hambleton and Swaminathan 1985). Thus, an important advantage of IRT over classical psychometric theory is that reliability is not reduced to a single, scale-level reliability coefficient, but instead varies continuously over the entire ability continuum. This permits identification of regions of the ability spectrum where a scale is particularly reliable, or conversely, has poor reliability and limited sensitivity to detect differences in ability.

Invariance If basic assumptions are met and samples include a broad range of variability, IRT has important invariance properties. Item parameters are invariant across samples and ability estimates are invariant across items. Invariance of item parameters means that item difficulty and discrimination are not dependent upon the specific distribution of ability within a sample and should be the same across two samples with similar ranges of performance. In practical terms, this means that differences in base rates of cognitive impairment should not influence IRT results; while in contrast, such differences have well-know effects on item and scale properties when using classic psychometric methods. Invariance of ability estimates means that the ability estimate is not dependent upon the specific items administered. As a result of this property, a scale can include different items for different groups and items can be selected to closely match each individual's ability. This enhances assessment efficiency since items that are not particularly discriminative for a given individual need not be administered.

The invariance properties of IRT are evident in the development and validation of the Spanish–English Neuropsychological Assessment Scales (Mungas et al. 2004, 2005). Participants included adults 60 years of age and older (345 Caucasians tested in English, 353 Hispanics tested in English, and 676 Hispanics tested in Spanish). IRT methods as described above were utilized to develop 12 initial cognitive scales, each with 90–100 items generated in English and translated into Spanish. Back-translation was performed by a team of fully bilingual individuals to minimize the likelihood of linguistic errors. In the initial phase (Mungas et al. 2000), items that showed bias across language groups or lacked desirable measurement properties were eliminated. In a second phase (Mungas et al. 2004), the retained items as well as a new pool of items were calibrated using a new participant sample. As an example of the obtained results, the test information curves for an Object Naming scale were comparable between English and Spanish-speaking participants, and showed high reliability across a broad range of naming ability.

Differential Item Functioning

Under equivalent testing conditions, individuals from different groups but comparable ability level are expected to have a similar probability of responding correctly to a particular test item. Operationally, this means that items with equivalent ICCs for two groups provide unbiased measurement. An item displays differential item functioning (DIF) when the conditional probability of obtaining a correct response differs between individuals matched on the underlying ability construct (Camilli and Shepard 1994; Hambleton et al. 1991).

There are two types of DIF: uniform and nonuniform. Uniform DIF occurs when the probability of a correct response is greater for one group than another across all levels of the ability spectrum. In contrast, nonuniform DIF occurs when the probability of a correct response varies across the ability spectrum (i.e., an interaction exists between group membership and ability). For instance, nonuniform DIF would be present in an object naming test if the probability of a correct response for a particular item is higher for Caucasians than for African Americans at the lower levels of naming ability, but higher for African−Americans than for Caucasians at the higher naming ability levels. In terms of the item characteristic curves described earlier, an item that is DIF-free has ICCs for each group that are comparable and highly overlapping. An item with uniform DIF (characterized by different item difficulty parameters) will display nonoverlapping but relatively parallel curves. In contrast, an item with nonuniform DIF (characterized by different item discrimination parameters) will display nonparallel curves (see Fig. 3 for an illustration).

Under an IRT framework, the basic steps for conducting a DIF analysis can be summarized as follows: (1) test the underlying assumptions (unidimensionality and local independence) for each group, (2) select the appropriate model for analysis, (3) perform an initial DIF analysis to identify items with DIF, (4) continue an iterative purification process to identify a group of DIF-free anchor items, (5) re-estimate DIF for each test item against the set of purified anchor items, and (6) determine which items demonstrate DIF based upon significance testing or effect-size estimation methods. A recent special issue of the journal *Medical Care* was devoted to the theory and methods of DIF-detection and provides detailed examples of a number of different approaches (Teresi et al. 2006).

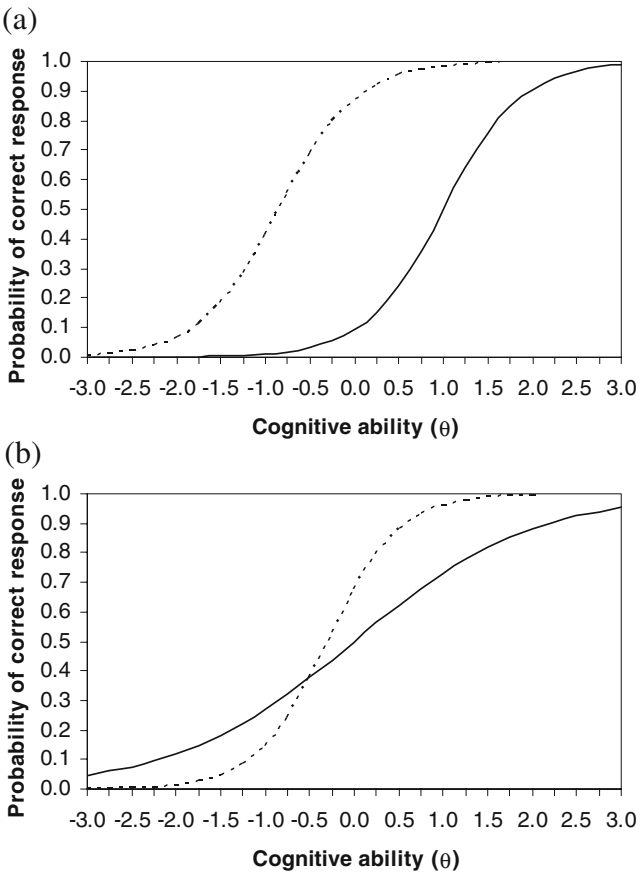

Fig. 3 Item characteristic curves of a sample item with uniform (**a**) or nonuniform (**b**) differential item functioning

IRT methods can be powerful tools for cross-cultural neuropsychological test development. DIF analyses can be used either to identify DIF-free items that provide unbiased measurement, or alternately to guide group-specific calibration of item parameters that can be used to calculate unbiased ability estimates. Test information curves can be used to guide item selection so that the resulting scale has desired characteristics. In neuropsychology, scales that have high and unchanging reliability across a broad range of ability are particularly effective and desirable. These scales have linear measurement properties such that a specific difference in underlying true ability is associated with the same difference in the expected test score, regardless of whether the starting point is high or low on the ability continuum. Moreover, these types of scales do not have floor or ceiling effects (see the Glymour et al. paper in this issue), tend to be normally distributed, and are particularly effective for measuring longitudinal change in populations with considerable variability of ability. An important consequence of the invariance properties associated with IRT is that different items can be used to measure a specific ability across different groups. In short, IRT methods can be used as the basis for including items in scales that have the desired measurement properties and that are psychometrically matched for different cognitive domains and across different cultural groups.

Concluding Remarks

The measurement of cognitive abilities across diverse cultural, racial, and ethnic groups has a contentious history, with broad political, legal, economic, and ethical repercussions. Driven in part by the impetus to enhance fairness across standardized psychoeducational measures, modern psychometric tools have propelled a new era of test development, standardization, and validation. Test developers now can capitalize from these tools in the construction of neuropsychological instruments with sufficient interpretive and procedural equivalence as well as normative reference standards to facilitate cross-cultural measurement. Simultaneously, a convergence of geneticists,

bioethicists, anthropologists, and additional experts from the social and natural sciences is contributing to a fuller understanding of human diversity.

The field of neuropsychology is poised to benefit from these advances in psychometric measurement and human genetics. Over the past 100 years, the bulk of cross-cultural studies on human cognition have centered upon measuring *how much* social groups differ from each other. Estimating these differences has been driven by economic, political, and legal motivations in the academic and employment fields. In aggregate, these studies have demonstrated a sizable gap between majority and minority ethnic groups. In clinical settings, the lower test scores obtained by African-American and Hispanic individuals presents a considerable challenge to diagnostic validity, a challenge that has been partly mitigated by the ongoing development and uniform use of appropriate normative comparison standards.

Yet, a complete understanding of the fundamental factors underlying the observed disparities continues to elude the scientific community. The recent advances in our conceptualization of measurement equivalence and invariance, use of modern psychometric methods derived from item response theory, refinement in the development and application of demographic norms whenever applicable, and progress in the biological and social sciences in weighing the contribution of genomic variation and nonbiological factors in our definition of 'race' hold substantial promise. Ideally, these advances will enhance our future capability to accurately measure cognitive ability and dysfunction, regardless of cultural background.

References

American Education Research Association, American Psychological Association, and the National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, D.C.: American Education Research Association.

Ardila, A. (1995). Directions of research in cross-cultural neuropsychology. *Journal of Clinical and Experimental Neuropsychology*, *17*, 143–150.

Ardila, A., Rosselli, M., & Puente, A. E. (1994). *Neuropsychological evaluation of the Spanish speaker*. New York: Springer.

Artiola i Fortuny, L., Romo, D. H., Heaton, R. K., & Pardee, R. E. (1999). *Manual de normas y procedimientos para la batería neuropsicológica en Español*. Tucson: AZ: m.

Bamshad, M., Wooding, S., Salisbury, B. A., & Stephens, J. C. (2004). Deconstructing the relationship between genetics and race. *Nature Reviews Genetics*, *5*, 598–609.

Beller, M., Gafni, N., & Hanani, P. (2005). Constructing, adapting, and validating admissions tests in multiple languages: The Israeli case. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3–38). Mahwah: Erlbaum.

Borsboom, D., Romeijn, J. W., & Wicherts, J. M. (2008). Measurement invariance versus selection invariance: Is fair selection possible? *Psychological Methods, 13*, 75–98.

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

Burchard, E. G., Ziv, E., Coyle, N., Gomez, S. L., Tang, H., Karter, A. J., Mountain, J. L., Perez-Stable, E. J., Sheppard, D., & Risch, N. (2003). The importance of race and ethnic background in biomedical research and clinical practice. *New England Journal of Medicine*, *348*(12), 1170–1175.

Busch, R. M., Chelune, G. J., & Suchy, Y. (2005). Using norms in neuropsychological assessment of the elderly. In D. K. Attix, & K. A. Welsh-Bohmer (Eds.), *Geriatric Neuropsychology* (pp. 133–157). New York: Guilford.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items. Volume 4*. Thousand Oaks: Sage.

Cleary, T. A. (1968). Test bias: prediction of grades of Negro and White students in integrated colleges. *Journal of Educational Measurement*, *5*, 115–124.

Collins, F. S. (2004). What we do and don't know about 'race', 'ethnicity', genetics and health at the dawn of the genome era. *Nature Genetics*, *36*(11, Suppl.), 13–15.

Cooper, R. S., Kaufman, J. S., & Ward, R. (2003). Race and genomics. *New England Journal of Medicine*, *348*(12), 1166–1170.

Cosentino, S., Manly, J., & Mungas, D. (2007). Do reading tests measure the same construct in multiethnic and multilingual older persons? *Journal of the International Neuropsychological Society*, *13*, 228–236.

Culhane, S. E., Morera, O. F., & Watson, P. J. (2006). The assessment of factorial invariance in Need for Cognition using Hispanic and Anglo samples. *The Journal of Psychology*, *140*(1), 53–67.

Darlington, R. B. (1971). Another look at "cultural fairness. *Journal of Educational Measurement*, *8*, 71–82.

Dotson, V. M., Kitner-Triolo, M., Evans, M. K., & Zonderman, A. B. (2008). Literacy-based normative data for low socioeconomic status African Americans. *The Clinical Neuropsychologist*, *21*, 1–29.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah: Erlbaum.

Fletcher, R. W., & Fletcher, S. W. (2005). *Clinical epidemiology: The essentials* (4th ed.). Baltimore: Lippincott.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory. Principles and applications*. Boston: Kluwer-Nijhoff.

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

Heaton, R. K., Miller, S. W., Taylor, M. J., & Grant, I. (2004). *Revised comprehensive norms for an expanded Halstead-Reitan battery: Demographically-adjusted neuropsychological norms for African American and Caucasian adults*. Lutz: Psychological Assessment Resources, Inc.

Helms, J. E. (1992). Why is there no study of cultural equivalence in standardized cognitive ability testing? *American Psychologist*, *47*, 1083–1101.

Helms, J. E. (1997). The triple quandary of race, culture, and social class in standardized cognitive ability testing. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 517–532). New York: Guilford.

Helms, J. E. (2002). A remedy for the Black-White test-score disparity [Comment]. *American Psychologist*, *57*(4), 303–305.

Ivnik, R. J., Malec, J. F., Tangalos, E. R., Peterson, R. C., Kokmen, E., & Kurland, L. T. (1990). The Auditory Verbal Learning Test (AVLT): norms for ages 55 years and older. *Psychological Assessment*, *2*, 304–312.

Jensen, A. R. (1980). *Bias in mental testing*. New York: Free.

Johnson, T. P. (2006). Methods and frameworks for crosscultural measurement. *Medical Care*, *44*(11 Suppl. 3), S17–S20.

Keita, S. O. Y., Kittles, R. A., Royal, C. D. M., Bonney, G. E., Furbert-Harris, P., Dunston, G. M., & Rotimi, C. N. (2004).

Conceptualizing human variation. *Nature Genetics*, *36*(11 Suppl.), S17–S20.

Lezak, M. D., Howieson, D. B., & Loring, D. W. (2004). *Neuropsychological assessment* (4th ed.). New York: Oxford University Press.

Liang, J. (2002). Assessing cross-cultural comparibility in mental health among older adults. In J. H. Skinner, J. A. Teresi, D. Holmes, S. M. Stahl, & A. L. Stewart (Eds.), *Multicultural measurement in older populations* (pp. 11–21). New York: Springer.

Lucas, J. A., Ivnik, R. J., Willis, F. B., Ferman, T. J., Smith, G. E., Parfitt, F. C., Petersen, R. C., & Graff-Radford, N. R. (2005). Mayo's older African Americans normative studies: normative data for commonly used clinical neuropsychological measures. *The Clinical Neuropsychologist*, *19*, 162–183.

Manly, J. J. (2005). Advantages and disadvantages of separate norms for African Americans. *The Clinical Neuropsychologist*, *19*, 270–275.

Manly, J. J., & Echemendia, R. J. (2007). Race-specific norms: using the model of hypertension to understand issues of race, culture, and education in neuropsychology. *Archives of Clinical Neuropsychology*, *22*(3), 319–325.

Manly, J. J., & Jacobs, D. M. (2002). Future directions in neuropsychological assessment with African Americans. In F. R. Ferraro (Ed.), *Minority and cross-cultural aspects of neuropsychological assessment* (pp. 79–96). Lisse: Swets & Zeitlinger.

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.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543.

Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care*, *44*(11 Suppl. 3), S69–77.

Messick, S. (1995). Validity of psychological assessment. *American Psychologist*, *50*(9), 741–749.

Michell, J. (1999). *Measurement in psychology: A critical history of a methodological concept.* New York: Cambridge University Press.

Mitrushina, M., Boone, K. B., Razani, J., & D'Elia, L. F. (2005). *Handbook of normative data for neuropsychological assessment* (2nd ed.). New York: Oxford University Press.

Mungas, D., Reed, B. R., Crane, P. K., Haan, M. N., & González, H. (2004). Spanish and English Neuropsychological Assessment Scales (SENAS): further development and psychometric characteristics. *Psychological Assessment*, *16*(4), 347–359.

Mungas, D., Reed, B. R., Haan, M. N., & González, H. (2005). Spanish and English Neuropsychological Assessment Scales: relationship to demographics, language, cognition, and independent function. *Neuropsychology*, *19*(4), 466–475.

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

Ostrosky-Solis, F., Ardila, A., & Rosselli, M. (1999). NEUROPSI: a brief neuropsychological test battery in Spanish with norms by age and education level. *Journal of the International Neuropsychological Society*, *5*, 413–433.

Parra, E. J., Kittles, R. A., & Shriver, M. D. (2004). Implications of correlations between skin color and genetic ancestry for biomedical research. *Nature Genetics*, *36*(11 Suppl.), S54–S60.

Ramirez, M., Ford, M. E., Stewart, A. L., & Teresi, J. A. (2005). Measurement issues in health disparities research. *HSR: Health Services Research*, *40*(5), 1640–1657.

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A., & Feldman, M. W. (2002). Genetic structure of human populations. *Science*, *298*, 2381–2385.

Roth, P. L., Bevier, C. A., Bobko, P., Switzer, F. S., & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: a meta-analysis. *Personnel Psychology*, *54*, 297–330.

Royal, C. D. M., & Dunston, G. M. (2004). Changing the paradigm from 'race' to human genome variation. *Nature Genetics*, *36*(11 Suppl.), S5–S7.

Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2001). High-stakes testing in employment, credentialing, and higher education. Prospects in a post-affirmative action world. *American Psychologist*, *56*, 302–318.

Schmidt, F. L., & Hunter, J. (2004). General mental ability in the world of work: occupational attainment and job performance. *Journal of Personality and Social Psychology*, *86*(1), 162–173.

Smith, G. E., Ivnik, R. J., & Lucas, J. A. (2008). Assessment techniques: Tests, test batteries, norms, and methodological approaches. In J. Morgan, & J. Ricker (Eds.), *Textbook of Clinical Neuropsychology* (pp. 38–57). New York: Taylor & Francis.

Steenkamp, J. B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*, 78–90.

Stewart, A. L., & Napoles-Springer, A. M. (2003). Advancing health disparities research. Can we affort to ignore measurement issues? *Medical Care*, *41*(11), 1207–1220.

Strauss, E., Sherman, E. M. S., & Spreen, O. (2006). *A compendium of neuropsychological tests: Administration, norms, and commentary* (3rd ed.). New York: Oxford University Press.

Teresi, J. A., & Holmes, D. (2002). Some methodological guidelines for cross-cultural comparisons. In J. H. Skinner, J. A. Teresi, D. Holmes, S. M. Stahl, & A. L. Stewart (Eds.), *Multicultural measurement in older populations* (pp. 3–10). New York: Springer.

Teresi, J. A., Stewart, A. L., Morales, L. S., & Stahl, S. M. (2006). Measurement in a multi-ethnic society. Overview to the special issue. *Medical Care*, *44*(11 Suppl. 3), S3–4.

Van der Linden, W. J., & Hambleton, R. K. (1997). Item response theory: Brief history, common models, and extensions. In W. J. van der Linden, & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 1–28). New York: Springer.

van de Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Newbury Park: Sage.

Van de Vijver, F. J. R., & Poortinga, Y. H. (2005). Conceptual and methodological issues in adapting tests. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 39–63). Mahwah: Erlbaum.

Wilkinson, G. S. (1993). *Wide Range Achievement Test 3*. Wilmington: Wide Range.