Published in final edited form as:

Clin Neuropsychol. 2016 October ; 30(7): 999–1005. doi:10.1080/13854046.2016.1211321.

A Commentary on “Generally Representative is Representative of None: Pitfalls of IQ Test Standardization in Multicultural Settings” by A.B. Shuttleworth-Edwards

Preeti Sunderaraman, Laura B. Zahodne, and Jennifer J. Manly

Department of Neurology and Taub Institute for Research on Alzheimer’s Disease and The Aging Brain, College of Physicians and Surgeons, Columbia University, New York, New York, United States of America

Abstract

We offer an appraisal of Professor Shuttleworth-Edwards’s commentary (2016) on the extraordinary challenges of assessment of cognitive function in a culturally, educationally, racially, and linguistically diverse population. First, we discuss the purpose of using intelligence tests in South Africa and beyond in order to clarify the reference group on which norms will be based. Next, we discuss the psychometric consequences of Pearson’s decisions to not adapt their measure of intellectual functioning to the cultural background of the majority of South Africans, and to use a population-matched normative sample in which the disadvantaged group is in the majority. We echo Professor Shuttleworth-Edwards’s call for large-scale empirical studies of cognitive test performance in a multicultural context. We conclude the article by urging the entire community of neuropsychologists to hold test companies accountable to strict, ethical, and comprehensive standards for production of accurate and appropriate measurement of cognitive function.

Keywords

cultural neuropsychology; IQ testing; norms

INTRODUCTION

We are grateful for the opportunity to comment on the detailed, impassioned argument for consideration of cultural issues in cognitive testing provided by Professor Shuttleworth-Edwards. Her commentary honors the legacy of a relatively small but active and inspiring community of neuropsychologists in South Africa who have been outspoken and insightful about the extraordinary challenges of assessment of cognitive function in a culturally, educationally, racially, and linguistically diverse population. We would like to highlight several important points in Professor Shuttleworth-Edwards’s review, attempt to further clarify some complex issues in cultural neuropsychology that were raised in her manuscript, and call on test developers to comprehensively respond to the challenges of cognitive

assessment in culturally diverse and developing countries by following empirically proven best practices for cognitive test development.

Professor Shuttleworth-Edwards's manuscript discusses two previous Wechsler IQ scales for adults that were marketed in South Africa. Her main conclusions are: 1) the tests are not "clinically viable products" because they result in IQ scores that are either inflated or artificially lowered, depending on the background of the testee; 2) norms specific to racial, cultural, educational, and linguistic subgroups are clinically more useful than population-based norms 3) researchers and test developers should commit resources to collection and production of subgroup-specific normative data, taking into account all relevant characteristics that are known to affect cognitive test performance, including first language and quality of education, as well as race, ethnicity, region, years of school, age, and sex.

In response to this article, we will: 1) review the different normative standards for descriptive versus diagnostic use of tests; 2) discuss some of the profound theoretical and psychometric challenges to measurement of intellectual function and cultural neuropsychology highlighted in settings such as South Africa; 3) concur with the author's call for additional research in South Africa, and 4) raise concerns about the practices of test companies whose markets increasingly include non-Western and developing countries.

PURPOSE OF TESTING SHOULD DETERMINE REFERENCE GROUP

The reason that a cognitive assessment is taking place should determine which tests are selected and the normative or reference standard used (Manly & Echemendia, 2007). While Prof. Shuttleworth-Edwards comprehensively describes the difference between population-based and within-group norms, it is unclear whether her critique of population-based norms for the adult Wechsler scales in South Africa relates to the determination of IQ (an estimate of overall level of intellectual or cognitive function as compared to age-matched peers) or for the use of the subtests as neuropsychological instruments for the diagnosis of acquired brain dysfunction or impairment. This is an important point because, 1) demographically adjusted norms should not be used for determination of IQ and, 2) population-based norms should not be used for determination of acquired impairment.

IQ scores are descriptive, population-based comparisons, predicated on the assumption that population performance on a test will be normally distributed (Gaussian) and that all scores reside under the curve (Busch, Chelune, Suchy, Attix, & Welsh-Bohmer, 2006). Population-based norms, such as those collected by Pearson for the Wechsler IQ measures, do not necessarily assume that the population is "relatively homogenous" as is stated by Prof. Shuttleworth-Edwards (p. xxx). IQ scores are intended to be summary measures of how an individual performs in comparison to a census-matched population (as heterogeneous as that population may be) of people of similar age without medical, neurological, psychiatric, or physical conditions that would interfere with cognition or completion of the tests. We do not adjust IQ scores based on background factors such as first language, years or quality of education, occupation, sex, or race. We do not estimate a child's IQ by comparing their score to other children in their school or in their town. Nationally representative samples are used because they reflect the broad context in which an individual is expected to function.

IQ scores must be demonstrated to be both reliable and valid. If they are, and the testee is a member of the general population included in the normative cohort, IQ scores can't be "too lenient", "strict" or "inflated"; the obtained score (and the associated 95% CI) represents the best estimate of where an individual's true IQ lies with respect to their age-matched peers across the nation.

If the goal is to determine whether an adult has an acquired impairment, demographically adjusted norms (described as within-group norms by Shuttleworth-Edwards) are appropriate. Diagnosis of acquired impairment, brain dysfunction, cognitive decline or change is typically measured against pre-injury or pre-decline scores (if one is lucky enough to have them), or our best estimate of what those scores are expected to have been, given the demographic features of the examinee. This is the traditional purpose of neuropsychological assessment. Because the person cannot be returned to the pre-injury or pre-illness state for a comparative assessment with him or herself, demographic characteristics are a substitute technique to estimate how that person would have performed, but-for the injury or illness. The more similar the comparative sample is to the examinee, the more precise the estimate will be, and the better the diagnostic accuracy of the measures. The normative standard for cognitive tests when used diagnostically is not population based, it is criterion or deficit based and uses the best estimate of the examinee's premorbid function as the normative standard. Demographic adjustments are not available for IQ index scores because the comparison standard of people with the exact same background as the examinee would be inappropriate for a score that is an intended comparison to the general population.

Commonly considered demographic variables that are used in developing demographically adjusted norms are age, race, ethnicity, sex, and education. Dr. Shuttleworth-Edwards reviews evidence that additional background factors have a profound impact on test performance both between and within racial/ethnic groups and should also be considered, including geographic region, first language and degree of multilingualism, school quality, acculturation, and migration or immigration status. In nations with wide socioeconomic, health, and social inequalities, deeply rooted residential and educational segregation across racial or ethnic lines, and complex linguistic diversity, these additional background factors may not only apply to a greater number of potential testees than in Western developed countries, but because of wider variation, may also have a stronger association with premorbid test performance.

DIVERSITY PRESENTS PROFOUND CHALLENGES TO MEASUREMENT OF INTELLECTUAL FUNCTION

Whether the purpose of testing is descriptive (such as determination of IQ) or diagnostic (such as diagnosis of acquired brain dysfunction), if the test is not a reliable and valid measure of the cognitive ability it is intended to measure, the representativeness and size of the normative sample will not improve the measure. If IQ accurately predicts outcomes thought to relate to intellectual function in well-educated Whites, but fails to do so in poorly educated Blacks, the test is biased. If reliability of a memory measure declines among people living in poverty, collection of normative data across the economic spectrum will not

improve that test's sensitivity and specificity for memory impairment among people with low socioeconomic status. In other words, demographic diversity presents challenges that go beyond the adequacy of the normative samples that are required to either measure intellectual functioning or to determine cognitive impairment.

In confronting these challenges, Professor Shuttleworth-Edwards and her colleagues in South Africa have provided vivid and thought-provoking questions about the assumptions underlying IQ assessment and how intellectual functioning is defined. People belonging to the dominant US culture (White, well-educated, US-born, and English speaking) developed, refined, and revised the Wechsler IQ scales using subtests, items, and standard test administration consistent with the values and assumptions of their culture. As a result, people whose cultural, educational, or linguistic experience are different than the dominant culture obtain lower scores on these tests than people from the dominant culture, even though their capacity to "act purposefully, to think rationally, and to deal effectively" with their environment (Wechsler, 1944, p. 3) and their intellectual potential is the same as people from the dominant culture. In the US, because the dominant culture is also in the majority, the typical performance of census-matched normative samples largely reflects the scores of White, well-educated, US-born, and English speaking people, so average IQ scores reflect dominant cultural background.

Pearson has apparently changed very little of the test content, subtest structure, and standard test administration of the US version of the Wechsler IQ scales for use in South Africa. In other words, the assumptions underlying the South African Wechsler IQ scales remain embedded in dominant US culture (White, well-educated, US-born, and English speaking). Just as in the US, people with disparate cultural, educational, or linguistic experiences obtain lower test scores than well-educated South African Whites, even though their intellectual capacity/potential is the same as their peers from the dominant culture. However, in South Africa, because the dominant culture is NOT in the majority, the typical performance of census-matched normative samples largely reflects the scores of Black, African first language, poorly-educated people, so average IQ scores DO NOT reflect dominant cultural background.

We want to make three points about the consequences of Pearson's decisions to not adapt their measure of intellectual functioning to the cultural background of the majority of South Africans, and to use a population-matched normative sample in which the disadvantaged group is in the majority. First, this decision seems to reflect the belief of the test developers that the way in which intelligence is manifested is universal across diverse settings, languages, and experiences, can be measured in the same way across diverse groups, and the standard should be that of the dominant culture. Second, as was clearly explained by Victor Nell (2000; 2007), IQ tests in South Africa may not show external bias because the typical criteria or outcomes used for examining bias in IQ tests (school and university success, occupational attainment, and adult income) are embedded in, and defined by, the same dominant culture that is reflected in the test. As Nell states, "the criterion might be as biased as the tests" (2007, p. 67). It is worth noting that even though IQ scores of poorly-educated, African first language Blacks are more likely to be in the average range using population-based norms in South Africa, the rank order of scores across racial, linguistic, educational,

and economic strata will remain the same regardless of the normative standard used. The third point is that the potential “demeaning” and “damaging” (Shuttleworth-Edwards, 2016, pp. XXX) effects, and possible “sinister aspect” (Victor Nell, 2007, p. 68) of setting the average IQ to reflect the performance of an impoverished, segregated, and educationally disadvantaged majority, must not be ignored and can be addressed through empirical investigation using well-powered studies with appropriate recruitment methods. These studies must also adhere to the principles discussed previously about proper use of different normative standards. For example, if the concern is that insurance companies are denying services to brain injured adults, the proper normative comparison group is a demographically matched sample (within group norms) because the question is diagnostic. Scores of brain damaged individuals will be compared to non-brain damaged demographically matched peers, and any difference between these scores will estimate the effect of the injury only, since the normative standard reflects our best estimate of premorbid ability. If the concern is that use of population-based norms leads to “incorrect placement decisions” (Shuttleworth-Edwards, 2016, pp. XXX) based on “inflated” IQ scores, this is a question of bias and test validity that should also be addressed empirically. If IQ scores do not accurately predict clinical outcomes, they lack validity. If their prediction of clinical outcomes is weaker among any subgroup, IQ scores are biased.

A DEARTH OF EMPIRICAL DATA

The previous section highlights the critical importance of adequately powered and well-designed research studies in cross-cultural neuropsychology. It appears that Pearson did not evaluate the external validity of the SA WAIS-IV measures before releasing the test, or if they did, these studies were not made available to readers of the test manual. Dr. Shuttleworth-Edwards describes valiant efforts to collect additional data from African first language blacks on the Wechsler scales, and calls for additional resources to evaluate the utility of these measures in South Africa. Dr. Shuttleworth-Edwards’s manuscript points to two key directions for this work: Goal 1 is the determination of validity of intelligence tests and testing for potential bias of IQ scores across diverse groups in South Africa, and Goal 2 is the collection of demographically appropriate norms for the Wechsler subtests when they are used diagnostically as neuropsychological instruments. Her prior research, described in this article as well as previous studies (A. Shuttleworth-Edwards, Gaylard, Radloff, Laher, & K, 2013; A. Shuttleworth-Edwards, Van der Merwe, van Tonder, & Radloff, 2013; A. B. Shuttleworth-Edwards et al., 2004; A. B. Shuttleworth-Edwards & Van der Merwe, 2015) is consistent with Goal 2; however, as she describes, these samples are too small for widespread use. The study design and recruitment methodology used in her studies are not appropriate for addressing Goal 1. The multidimensional, systematic, and large-scale process of establishing validity of the Wechsler IQ scales in South Africa is beyond the scope of our commentary; however, many examples can be found in the Technical Manuals of the US versions of the WAIS-III and the WAIS-IV, which devote the majority of their pages to description of well-designed validity studies. Numerous manuscripts and reviews have been dedicated to description of large, well-funded studies of predictive utility of IQ tests in the US and other Western countries, but the validity of IQ measures in South Africa

and developing countries remains largely unproven (Sternberg, Grigorenko, & Bundy, 2001).

RESPONSIBILITY OF TEST COMPANIES

The validity and reliability of cognitive tests must be established through a rigorous development process for which the test publisher is responsible. Unfortunately, the presence of normative adjustments in a test manual or a scientific manuscript does not indicate that the measures are appropriate for use for every person or situation, or even that norms have been collected and presented in a way that meets the standard of the field. Collection of nationally representative norms is not sufficient for validation of a measure of intellectual functioning, or for demonstrating the predictive utility of IQ scores. Dr. Shuttleworth-Edwards raises several concerns about the process of test development in South Africa that apply to activities in other non-Western and developing countries.

For-profit test companies are increasingly operating in “markets” where there is a tremendous lack of resources to perform systematic, sustained, and independent research on the products they develop and sell. We suggest that test companies should have more stringent standards for test development, and undergo a more transparent and collaborative process when operating in these environments, not only because of the potential for exploitation that societal inequities present, but also because this will result in a better instrument.

For example, test companies often recruit local professionals to help collect norms for existing measures, offer financial incentives to these testers, and provide unique opportunities for employment for professionals in developing countries. In developing countries, these local testers are unlikely to be neuropsychologists. The power, prestige, and resources of a large company present imbalances that may provide disincentives for local team members to raise concerns about the appropriateness of test items and procedures in the new setting. Therefore, involvement of local professionals in adaptation and norming of cognitive measures does not necessarily imply that it is a collaborative or culturally sensitive process. In fact, the struggles depicted by Dr. Shuttleworth-Edwards in her manuscript suggest that there is a wide disconnect between the needs of clinicians in South Africa and the needs and expectations of well-informed users of the Pearson product.

CONCLUSION

Professor Shuttleworth-Edwards eloquently describes the complex interaction of factors such as educational experience, language, socioeconomic status, and acculturation, their relevance across the lifecourse, and their potential impact on cognitive test performance. She is operating in a context where she confronts extraordinary variability in cultural background in her every day research and practice. Her manuscript is also a troubling description of disconnect between products released by for-profit testing companies, and the needs of well-trained experts who do not have the resources to collect the psychometric and normative data needed for proper use of tests. It is inspiring that Dr. Shuttleworth-Edwards and her colleagues are training a new generation of neuropsychologists to expect rigorous validity

testing and cross-cultural appropriateness of cognitive measures, and to develop research studies to improve the utility of cognitive measures. As a field of neuropsychological scientist-practitioners, and as members of organizations whose goals are to enhance and promote the science of neuropsychology internationally, we should provide opportunities for collaboration, funding and training in non-Western and developing countries, and should partner with local experts to hold test companies accountable to strict, ethical, and comprehensive standards for production of accurate and appropriate measurement of cognitive function.

References

- Busch, RM.; Chelune, GJ.; Suchy, Y.; Attix, DK.; Welsh-Bohmer, KA. *Geriatric Neuropsychology*. New York: Guilford Press; 2006. Using norms in neuropsychological assessment of the elderly; p. 133-157.
- Manly JJ, Echemendia RJ. Race-specific norms: Using the model of hypertension to understand issues of race, culture, and education in neuropsychology. *Archives of Clinical Neuropsychology*. 2007; 22(3):319–325. [PubMed: 17350797]
- Nell, V. *Cross Cultural Neuropsychological Assessment: Theory and Practice*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc., Publishers; 2000.
- Nell V. Environmentalists and Nativists: The IQ Controversy in Cross-Cultural Perspective. *International Handbook of Cross-Cultural Neuropsychology*. 2007:63–92.
- Shuttleworth-Edwards A, Gaylard E, Radloff S, Laher S, K C. WAIS-III test performance in the South African context: extension of a prior cross-cultural normative database. *Psychological assessment in South Africa: Research and applications*. 2013:17–32.
- Shuttleworth-Edwards A, Van der Merwe A, van Tonder P, Radloff S. WISC-IV test performance in the South African context: A collation of cross-cultural norms. *Psychological assessment in South Africa: Current research and practice*. 2013:33–47.
- Shuttleworth-Edwards AB, Kemp RD, Rust AL, Muirhead JG, Hartman NP, Radloff SE. Cross-cultural effects on IQ test performance: A review and preliminary normative indications on WAIS-III test performance. *Journal of Clinical and Experimental Neuropsychology*. 2004; 26(7):903–920. [PubMed: 15742541]
- Shuttleworth-Edwards AB, Van der Merwe AS. WAIS-III AND WISC-IV SOUTH AFRICAN CROSS-CULTURAL NORMATIVE DATA STRATIFIED FOR QUALITY OF EDUCATION. *Minority and Cross-Cultural Aspects of Neuropsychological Assessment: Enduring and Emerging Trends*. 2015; 72
- Sternberg RJ, Grigorenko E, Bundy DA. The predictive value of IQ. *Merrill-Palmer Quarterly*. 2001; 47(1):1–41.
- Wechsler D. *The measure of adult intelligence*. Baltimore, Williams and Wilkins. 1944