Cult. Brain
DOI 10.1007/s40167-017-0050-2

REVIEW ARTICLE

Bias in cross-cultural neuropsychological testing: problems and possible solutions

Alberto Luis Fernández[1] · **Jennifer Abe**[2]

Accepted: 19 April 2017

Abstract Cultural variables exert a powerful effect on test performance. This effect is now largely recognized in the field of neuropsychology, although rather underestimated. This paper has three parts. First, different sources of cross-cultural bias in neuropsychological testing are identified, using the taxonomy proposed by van de Vijver and Tanzer (Eur Rev Appl Psychol 54: 119–135, 2004), specifically, an examination of construct, method and item biases. Second, strategies proposed in the literature to address these biases are reviewed. Finally, a three-level approach to addressing these problems related to bias is proposed. These approaches are hierarchically organized from bottom-to-top: (1) a behavioral approach in the testing situation, (2) test adaptation and, (3) the development of a new generation of neuropsychological tests. Simultaneous test development across multiple cultures is emphasized. Guidelines for the development of these tests are proposed in order to obtain culturally fair and psychometrically robust tests.

Keywords Cross-cultural neuropsychology · Bias · Simultaneous test development · Culture fair tests

✉ Alberto Luis Fernández
neurorehab@onenet.com.ar

Jennifer Abe
jsabe@lmu.edu

[1] Universidad Católica de Córdoba- Cortex Foundation-Universidad Nacional de Córdoba, Chile 279, CP 5000 Córdoba, Argentina

[2] Department of Psychology, Loyola Marymount University, 1 LMU Drive Suite 4700, Los Angeles, CA 90045, USA

Published online: 12 May 2017

Introduction

The influence of culture on cognition poses a tremendous challenge to neuropsychological assessment with culturally diverse populations (Ardila 2007a, b; Nell 2000; Manly 2008), not the least of which includes the difficulty of responding to the wide range of cultural contexts, conditions, and circumstances under which testing may occur around the world. For instance, to what extent is neuropsychological testing available in developing countries? Are appropriate norms available? Are the available tests appropriate for individuals with few years of education or low levels of reading literacy? Some of the subsequent challenges include the interpretation of neuropsychological tests scores when applied to culturally diverse populations, the lack of cross-cultural validity for many neuropsychological tests, as well as the difficulty in analyzing cognitive disturbances in cases of brain pathology in different cultural contexts. Further, neuropsychological assessment itself reflects a culturally-embedded process. That is, culture permeates all aspects of neuropsychological assessment, from the identification and definition of relevant constructs, the development and construction of tests, as well as the clinical interview and observations, process of test administration, and interpretation of performance (Ardila 2007a, b). Thus, a central question concerns how can we recognize the influence of culture on neuropsychological assessment to meaningfully address such highly varied issues?

Growing evidence indicates that culture exerts a strong influence on cognition (Henrich et al. 2010). While basic cognitive processes and characteristics are common to all human beings, these processes and functions can develop in culturally-distinctive ways and may differ in how they are expressed across cultures (Henrich et al. 2010; Uzzell et al. 2007). Cultural influences have been identified for visual perception (Segall et al. 1966; Henrich 2008), analytic versus holistic reasoning (Miyamoto et al. 2003; Norenzayan et al. 2007) spatial reasoning (D'Andrade 1995; Gordon 2004; Majid et al. 2004), brain organization (Chiao and Cheon 2010; Chua et al. 2005), sense of self (Markus and Kitayama 1991; Shweder and Bourne 1984), motivation (Bond and Smith 1996) and personal choice (Iyengar and Lepper 1999; Iyengar and DeVoe 2003), among others.

Moreover, there is also evidence that the influence of culture is present not only in cognition, but also in brain functioning itself. For instance, several studies have demonstrated differences in brain activation of the ventral visual cortex between East Asians and Westerners (Goh et al. 2004, 2007; Gutchess et al. 2006). Although both groups activate similar areas to process visual stimuli, the extent of the activation is different for each group. Differences in brain activation between East Asians and Westerners have also been observed in fronto-parietal regions during the performance of simple attention tasks (Hedden et al. 2008). Chiao et al. (2008) even demonstrated that the amygdala response to fear faces is modulated by culture. Likewise, differential brain activation between East Asians and Westerners is shown in face processing, with the latter activating the fusiform face area bilaterally whereas the former showed more right lateralization of the same region (Goh et al. 2010).

Nonetheless, it is not always clear what we mean by culture, especially when group differences are used as a proxy for cultural differences. How is culture operationalized when identifying these group differences, and how can we deepen our understanding of these cultural influences? Culture represents "… a dynamic process involving worldview and ways of living in a physical and social environment shared by groups, which are passed from generation to generation and may be modified by contacts between cultures, in a particular social, historical, and political context" (Whaley and Davis 2007, p. 564). From this perspective, individuals are viewed as enculturated or socialized within cultural currents or flows (Hermans and Kempen 1998) that are, themselves, historically, geographically, and socially situated. In an age of globalization and significant streams of migrating peoples, the notion of culture has become correspondingly more complex. Hermans and Kempen (1998) argue that, "in an increasingly interconnected world society, the concept of independent, coherent, and stable cultures becomes increasingly irrelevant (p. 1111). Thus, rather than viewing culture as static entities that are "internally homogenous and externally distinctive," the notion of culture may be more fruitfully regarded as "representing particular ways to adapt to and survive in a specific context" (Ardila 2007a, b, p. 24). While the cross-cultural psychology literature has tended to emphasize cultural dichotomies (western vs. non-western, individualistic vs. collectivistic, etc.), a recognition of the complexity of cultural influences on identity, cognition, and behavior is becoming increasingly critical for the further evolution of cross-cultural neuropsychology. The goal of the present paper is to examine specific expressions of cultural influences on cross-cultural neuropsychological assessment with particular attention to test biases, to consider the measurement and assessment approaches that have been developed to respond to such biases, and to identify alternative strategies for increasing the utility of neuropsychological testing across culturally diverse populations.

One important step towards the goal of providing accurate and appropriate neuropsychological testing for culturally diverse groups is to examine the specific ways by which test biases may manifest in cross-cultural neuropsychological assessment. In the present article, the concept of bias, equivalence, and some sources of bias in cross-cultural neuropsychological assessment will be examined following the taxonomy developed by van de Vijver and Tanzer (2004). Here, although "cross-cultural" will be used to represent many types of cross-national cultural differences, including nationality, tribe, ethnicity, language, and religion in keeping with major empirical findings in the literature, we recognize that significant cultural diversity exists within countries and do not wish to minimize the importance of acknowledging and addressing intra-national or within-country cultural differences. Many of the strategies to address cross-cultural biases that are discussed in the present paper may be useful in working with culturally diverse populations within a single country. At the same time, other differences that are not specific to a particular culture are important to examine across cultures, such as the effects of acculturation, bilingualism, age, and exposure to formal education, including literacy levels. These will not be examined at length in this article, as they have been discussed elsewhere, but represent key factors influencing performance

on neuropsychological tests within and across cultures (Nell 2000; Uzzell et al. 2007).

This brief taxonomic review is followed by a discussion of some measurement and assessment strategies proposed by different authors to address these biases, including efforts to adapt existing instruments as well as to develop new neuropsychological instruments that are appropriate for use within specific groups (see the Córdoba Naming Test as an alternative to the Boston Naming Test in Spanish speakers, for example in Fernandez 2013) or across different cultural groups (see the development of the RUDAS for example in Storey et al. 2004). After all, the ultimate goal of cross-cultural neuropsychological assessment is to develop valid and meaningful local measurements of constructs that are equivalent across cultures. Towards this goal then, we need to consider not just the scientific ideal (which tends to minimize the practical challenges of engaging in sound neuropsychological assessment across clinical settings in different cultural contexts), but also identify scientifically-sound practices that are able to address these pragmatic concerns and challenges. Specifically, increasing the cross-cultural validity of neuropsychological tests may require both (a) adapting available instruments so that they attain a similar meaning in a different cultural context (cultural decentering) as well as (b) developing tests for a culture that have meaning and validity with respect to a particular cognitive function (cultural centering).

Next, the paper examines other alternatives for increasing the effectiveness of neuropsychological assessment across cultures. Specifically, the authors propose an alternative to efforts to develop appropriate culture-specific test norms for each group. First, the aspiration to develop appropriate norms for every tested population is not feasible, nor even possible (Ardila 2007a, b). Instead, it is critical to examine and understand factors influencing performance on neuropsychological tests (Ardila 2007a, b) as well as to consider alternatives to the use of norms for comparing individual performance against a range of the performance of similar others. This approach is not meant to replace or supplant the scientific ideal of developing high quality local norms for use with major neuropsychological instruments, but to provide a scientifically-sound set of pragmatic alternatives that may serve as a useful transition until such norms may be established in countries where they are not yet available. In this final section, we propose greater attention to the use of cut-off scores to indicate the presence/absence of significant cognitive impairment in instruments that have been culturally de-centered and developed for cross-cultural use as a useful initial strategy for increasing the efficacy of cross-cultural neuropsychological assessment for underserved populations. This strategy is presented as an alternative to the present emphasis in practice on identifying gradations of cognitive impairment, even in the absence of appropriate norms.

Finally, the paper briefly examines current neuropsychological instruments that have been developed for cross-cultural purposes. This discussion draws greater attention to the particular populations and areas of functioning for which instruments have been developed, the different strategies for developing these instruments, as well as their subsequent strengths and limitations. In summary, the present paper has four sections: (1) an examination of specific sources of test bias in cross-cultural neuropsychological assessment, (2) a discussion of current responses

to these sources of cross-cultural test bias and the need to address cultural differences more meaningfully and pragmatically in the development of neuropsychological tests, (3) a proposed strategy for increasing the utility of neuropsychological tests across culturally diverse, and especially underserved, populations, and (4) a short review of current cross-cultural neuropsychological tests.

What is bias?

According to van de Vijver and Tanzer (2004), "…bias occurs if score differences on the indicators of a particular construct…do not correspond to differences in the underlying trait or ability…" (p. 120). If, for instance, the Trail Making Test were to be administered to two subjects from two different cultures, bias would be said to occur if the differences in the times taken to complete the task were not reflective of differences in the attentional capacity of the subjects from these different cultures, but instead reflective of differences in their respective familiarity with the test stimuli, different administration procedures, or other factors. In contrast, *equivalence* is demonstrated when differences in test performance are reflective of true differences in the underlying trait or ability assessed.

Sources of bias in cross-cultural testing. Neuropsychological examples

As mentioned previously, van de Vijver and Tanzer's (2004) taxonomy may be useful in analyzing bias in cross-cultural neuropsychological testing. Basically, they identified three types of bias: (a) Construct bias; (b) method bias and; (c) item bias. In this section, examples will be given of each of these sources of bias as they may manifest in different neuropsychological cross-cultural testing situations.

Construct bias

Construct bias occurs when the construct measured is not equivalent across cultural groups (van de Vijver and Tanzer 2004). One of the most frequently used constructs in neuropsychology is that of intelligence. Thus, it is important to discuss if this construct is equivalent across cultures. The concept itself is controversial since there is not a consensus on what the intelligence is. There are multiple theories and definitions of intelligence (Gardner 2011; Legg and Hutter 2007). The theories of intelligence have different theoretical underpinnings involving psychometrics, cognitive theories, cognitive-contextual theories and biology (Gardner 2011). Nevertheless, for the most part, intelligent behavior has traditionally been defined in Western cultures (see Henrich et al. 2010) and may not be viewed as such in all other cultures. For example, Grigorenko et al. (2001) found that the intelligence concept of Kenyans includes four distinct terms for intelligence: *rieko* (knowledge and skills), *luoro* (respect), *winjo* (comprehension of how to handle real-life problems), and *paro* (initiative). As can be seen, only one of these terms (*rieko*) matches closely with the Western concept of intelligence as measured on most tests,

although *winjo* also appears to relate to the "practical" or "contextual" component of Sternberg's triarchic theory of intelligence (Sternberg 1985). Indeed, Gergen et al. (1996) go so far to state that "the concept of intelligence seems historically constituted to meet the challenges faced by western institutions cultures in gaining control over their constituents" (p. 500). Whether one would go so far as to link the development of the construct of intelligence with the exercise of power is another matter, but it is clear that the construct of intelligence is defined in different ways around the world (Saklofske et al. 2015). Thus, far from appearing to be a universal construct, the notion of intelligence itself appears to reflect its particular cultural context and even historical period in which it was first developed (Saklofske et al. 2015; Shuttleworth-Edwards and Van der Merwe 2016). Nonetheless, from the myriad of theories that abound, one that has gained support from cross-cultural research is the Three-Stratum Factor Analytic Theory or so-called Cattel-Horn-Carroll (CHC) theory of intelligence, which encompasses three levels of cognitive abilities. There is considerable evidence supporting the hypothesis that cognitive processes such as visual processing, short-term memory, quantitative knowledge or auditory processing, among others (included in the second stratum), are universal, i.e., they are present in all human beings independently of the culture (Berry et al. 2011; Lonner 2011; Saklofske et al. 2015; van de Vijver 1997). For example, memory, defined as the ability to store information, seems to be a universal construct. This topic is rarely addressed in a direct way in the cross-cultural neuropsychology literature, with some exceptions, such as Ardila and Keating (2007) who have noted: "Although basic cognitive processes are universal, cultural differences in cognition reside more in the situations in which particular cognitive processes are applied than in the existence of the process in one cultural group and its absence in the other" (p. 109). Many cross-cultural psychologists support the claim that basic cognitive processes are universal and what is influenced by culture is the way they are expressed in different cultural settings (Cole and Packer 2011; Berry et al. 2011; Saklofske et al. 2015; Oyserman et al. 2009; van de Vijver 1997). Mishra (1997) describes four main theoretical models that explain the relationship between cognitive development and culture: general intelligence ("…based on the idea of a unitary cognitive competence, called 'general ability'…", p. 148), genetic epistemology (referring to developmental processes that unfold in a chronological sequence studied across cultures), specific skills [which assume cognitive processes are universal and that "…cultural differences in cognition reside more in the situations to which particular cognitive processes are applied…" (Cole et al. 1971, p. 233)], and cognitive styles (which "…looks for interrelationships (patterns) in cognitive performances, and postulates different patterns of abilities to develop in different ecocultural settings, depending on the demands placed on an individual…" p. 154). He concludes that, "at the theoretical level, the four conceptualizations dealing with the relationship between culture and cognition seem to draw closer to each other. Admission of a 'central processing' mechanism and the perceived probability of 'transfer' of skills from one situation to another in the specific skills approach now brings it closer to other views" (p. 168). All these authors agree on the hypothesis that people share the same basic cognitive processes across cultures and what differentiates these cultures are the "cognitive styles" or "mind-sets",

which are flexible elements that can be shaped by culture. Thus, for example, attention skills can be assessed across cultures in different individuals, but every culture will determine what phenomena should be paid attention to or what elements will be emphasized when paying attention to a stimulus.

There is evidence that supports the universality of basic neuropsychological constructs, as well. Marchant et al. (1995) investigated the notion of universality of handedness. Using an ethological approach of hand use they analyzed the behavior of three traditional (i.e. nonliterate) cultures, the G/wi from Botswana, the Himba from Namibia and the Yanomamö from Venezuela. Although they found that right preference was only visible when applied to tool use, they reported no differences across these cultures, therefore giving support to the notion that handedness does not differ across cultures. Although other comparisons of handedness studies carried out in different cultures yielded remarkable differences, Marchant and McGrew (1999) pointed out that each of these studies are suspected of having a serious methodological bias, making these comparisons questionable. Their methodological observations are: (1) few measures were used to assess hand-use; (2) most subjects were children, which is an inappropriate sample due to the immaturity of the subjects; (3) most of the studies relied upon induced demonstration of the requested tasks and not upon ethological observation; and (4) the tasks administered were not a random sample of manual activity but rather the use of Western tools such as scissors.

However, even when the constructs representing basic cognitive processes such as attention, memory or even handedness may be universal across cultures, some data on other processes, such as color perception, pose a challenge to this notion. Roberson et al. (2000), for example, examined individuals from the Berinmo tribe from Papua-New Guinea, because the Berinmo language has only five words to describe colors. In several studies, they compared color perception between English-speakers and Berinmo-speakers finding that English-speakers had an advantage over Berinmo in the recognition of the colors that were defined in English, but the Berinmo-speakers had an advantage over the other group on the recognition of Berinmo-defined colors. In addition, speakers of both languages experienced the same level of difficulty in learning a new color category which did not exist in either language. Differences in color perceptions related to language differences support the notion that language affects perception, otherwise known as the Sapir-Whorf hypothesis or linguistic relativity hypothesis. According to this hypothesis, our understanding of the world is mediated by language. Although this hypothesis has been the subject of controversy (Kay and Regier 2007), these findings demonstrate that color perception cannot be taken for granted as universal.

Another possible source of bias is that particular tests may not measure the same existing construct across cultures. A good example of this is found in a study by Lee et al. (2000), who found that correlations between the Trail Making Test (TMT) and the Colors Trail Test (CTT), a purportedly TMT analogous culture-free test, were weaker in a Chinese-English bilingual sample than in an English monolingual sample. Whereas correlations for the monolingual sample between Part A of the TMT and Part 1 of the CTT were .72, and .75 between Part B of the TMT and Part 2 of the CTT, these correlations dropped to .25 and .59 respectively in the Chinese-

bilingual sample. Given the fact that these were comparable samples in terms of age and education, the authors question the construct validity of the test in the bilingual sample, suspecting that it might be measuring a different construct in this case. Thus, two important issues need to be analyzed when considering construct bias: the construct itself and the construct validity of the tests.

Cross-cultural psychologists usually resort to the measurement invariance approach to test whether the internal structures of measurement instruments are equivalent between cultural groups (Fischer and Fontaine 2011; Meredith 1993; Meredith and Teresi 2006; Milfont and Fischer 2010). However, although a very valuable tool in cross-cultural testing research, it may not be appropriate for use with neuropsychological tests. This is because all of the measurement invariance methods are based on factor analysis (Fischer and Fontaine 2011; Milfont and Fischer 2010), which are applied to questionnaires or scales that consist of a variety of different items with individual ratings or scores for each item. That is not usually the case for many neuropsychological tests. For example, the Rey Complex Figure Test usually yields one quantitative score that represents the performance of the subject on the whole task. Similarly, the Trail Making Test generates two timed scores, one for each part; for the Stroop test, a single score represents the performance on the whole task. Another example is the Wisconsin Card Sorting Test, which generates several scores, but is not amenable to factorial analysis since these scores do not represent individual items, but rather performance on the entire test. Thus, except for some specific cases, factor analysis cannot be performed with most classic neuropsychological tests which means that the strategy of measurement invariance is difficult to apply. One of the cases in which measurement invariance was successfully employed was with the Spanish and English Neuropsychological Assessment Battery (SENAS), an instrument specifically developed for cross-cultural comparisons using item response theory methods (Mungas et al. 2000; Mungas et al. 2011). Mungas et al. (2011) demonstrated the same factorial structure of the SENAS held across several cultural/linguistic groups (Whites, African Americans, English-speaking Hispanics, and Spanish-speaking Hispanics). Another example of a battery measuring cognitive functions under the IRT umbrella is the Woodcock-Johnson Tests of Cognitive Ability, Third Edition (WJ-III COG) (Woodcock et al. 2001).

In summary, although many neuropsychological constructs seem to be universal, including language, attention, memory, spatial skills and executive functioning, the manner in which they are expressed in different cultures may vary. On the other hand, some constructs, such as intelligence, reasoning or color perception, should not be taken for granted as universal across cultures. Further, it cannot be assumed that a construct assessed by a certain test in the culture for which it was originally designed is the same construct that the same test is assessing in a different culture.

Method bias

Method bias refers to aspects of assessment related to methodological issues and includes sample bias, instrument bias and administration bias.

Sample bias occurs when samples are incomparable on aspects other than the target variable, for instance, differences in size, educational background, age, or selection criteria. An example of how sample bias can affect these comparisons can be observed in the performance of children from Canada, Israel, the U.S., and Ecuador on a neuropsychological battery (Levav et al. 1998). Children from Ecuador had significantly lower scores on all the tests. At the same time, a high proportion of Ecuadorian children suffered from malnutrition and parasitic infections, while the rest of the children were healthy. The authors acknowledged that these unhealthy conditions had an impact on neuropsychological performance. Consequently, no proper comparisons can be made regarding the performance of Ecuadorian children since these samples are not comparable. Another example of sample bias may be found in a study that compared the performance of normative samples from 11 countries on the Trail Making Test (TMT) (Fernandez and Marcopulos 2008), which found large differences in the mean time needed to complete each part of the TMT across the samples. However, samples were not comparable in their composition, especially regarding the number of years of education. This fact precluded other analyses, including exploring the possibility of construct bias, because the samples were not comparable.

The second subtype of method bias is instrument bias. Instrument bias can be caused by different reasons such as differential familiarity with stimulus material or response procedures, or different response styles, e.g. social desirability, extremely scoring, etc. Familiarity with the test materials is very important since it can determine how easily the testee can understand the task he/she is asked to do. One clear example of instrument bias is presented, again, with the TMT. Since this test uses letters from the Latin alphabet it is inappropriate for use in cultures with different alphabets (Chinese, Greek, Japanese, Hebrew, Arabic, etc.). For this reason, Axelrod et al. (2000) developed a Hebrew version of the test using Hebrew characters instead of the Latin characters.

The effects of instrument bias are even more extreme when tests that were developed in Western societies are used in the assessment of people from non-Western cultures. A sample of Aruaco Indians in Colombia, for instance, was not able to complete the block design subtest from the WISC-R within time limits (Ardila and Moreno 2001) an outcome which the authors related to differences in conceptions of time. The Aruaco usage of time is much more flexible than that of the Westerners, so that when they say “later” this may mean in 1 h, several hours, or several days. The authors also observed that three of the 20 individuals assessed were unable to complete the Rey-Osterrieth Copy Figure test because they had never used a pencil before, indicating that the materials and the procedures used to assess constructional praxis were not appropriate for this population. Such findings suggest that the application of standard Western materials and procedures without regard to characteristics of the target population may easily lead to useless and invalid scores. Even worse, such findings may lead to misconclusions regarding the existence of brain damage. Although the Aruaco people, an indigenous group only partially integrated into the Western society, and other indigenous peoples might represent more extreme examples, neuropsychologists frequently face situations

where they have to assess people from very dissimilar cultures from their own in clinical practice.

The final type of method bias to be discussed is related to problems in the administration of tests. In this case, language issues, or other kind of communication problems, may be a source of misunderstanding between the tester and the testee. As a result, scores obtained will not reflect the testee ability, but rather how well he/she understood of what he/she was being asked to do, especially when the test instructions of the tester are not clear to the testee. This might happen if the tester uses advanced vocabulary not commonly used in everyday life or highly technical words that are unknown to the testee. Another important source of this bias is limited language capacity and/or use on the part of the tester or testee. This situation is not rare if the person under evaluation is an immigrant and the tester does not speak the mother tongue of the testee fluently or if he/she has to rely on an interpreter. Sometimes the instructions are hard to translate and the translation may be too literal, which can make them confusing or inappropriate. Changing the administration mode can cause serious differences cross-culturally, as well. Fernandez and Marcopulos (2008) found that the administration of the TMT in the Danish standardization study was different than the other studies. This change in the administration mode might explain why the Danish sample had, on average, shorter times on the test than the other samples.

Shepherd and Leathem (1999) worked with a group of Maori in New Zealand, in another example of administration bias. The Maori have lived in New Zealand since the pre-European colonization times. At present, they represent approximately 15% of the total population (Statistics New Zealand 2013). These investigators administered a survey to 15 individuals who underwent neuropsychological evaluations to determine how they perceived the testing experience. Seven out of the fifteen subjects were Maori and eight were non-Maori. In general, Maori were less satisfied with the service than non-Maori subjects. Some of these respondents expressed that they felt uncomfortable because they had "a feeling of failure as unaware of what level of achievement (expected of me)" (pp. 84). This feedback is reflecting communication problems. Unfortunately, in that paper there is no indication of how these problems might have affected observed performance. Indeed, this issue of misunderstanding test instructions is a pervasive one in the whole field of psychological testing. This issue leads us to a broader discussion about factors affecting test reliability. If the instructions are not clear to the individual under examination, or if the administration procedure is not standardized, then the reliability of the test will be compromised. It is of paramount importance that instructions to the testee are clear but also that the administration conditions are standard (Hogan 2014; Hogan and Tsushima 2016), otherwise, the individual will response according to his/her interpretation of the instructions which may not coincide with the intended interpretation by test developers. The test manual should include an exhaustive description of possible situations that the tester will be able to face during the administration, and the proper reactions to those situations. Although most of the testees will show similar responses, some of them will show a variety of unusual responses. Thus, the manual should instruct the tester on how to react to that in order to secure the reliability. In the specific case of

neuropsychological tests, misunderstandings might stem from language comprehension problems derived from the condition itself (such as in aphasia), which are not related to the instructions or the administrations conditions. However, this situation will compromise reliability, as well. Therefore, neuropsychological tests should be developed in such a way that these kinds of problems can be properly addressed by the tester to ensure an adequate comprehension of instructions by the testee. Ideally, instructions should be so easy to understand that even a patient with language comprehension problems could comprehend them.

In summary, unfamiliar materials or response procedures, as well as communication problems are powerful sources of bias in neuropsychological evaluations leading to serious consequences in the interpretation of test scores.

Item bias

An item is biased when individuals from different cultural groups, who obtained the same score on the construct, have different scores on the item. This has been called differential item functioning (DIF). The DIF concept refers to the fact that the same item has different meanings across cultures (Holland and Wainer 1993). The most widely used statistical method for the detection of DIF is the Mantel–Haenszel procedure, which basically analyzes bias in dichotomously scored items (Dorans and Holland 1993). There are numerous examples of this sort of bias and it is a very frequent source of misunderstandings in research because the same factorial structure across groups can still can mask biased items. Tanzer (1995), for instance, showed that although the factor structure of an academic self-concept test applied to Singaporean and Australian samples of students were similar, substantial differences were found between these samples when specific items were compared.

There are also examples of item bias in neuropsychological testing. One of the clearest examples of item bias in neuropsychology is related to the cross-cultural administration of the Boston Naming Test. This test is comprised of a set of 60 pictures that are hierarchically arranged regarding their naming difficulty level. However, in its many adaptations to different cultures, researchers have frequently found that some items are not culturally appropriate. For example, one item depicts a pretzel, which is unknown in many countries, and another item contains a beaver which only dwells in North America and some parts of Europe and Asia, thus it is unknown to many people in Central and South America, Africa, Oceania and large portions of Europe and Asia. As a result, changes in the original list of items were made in some adapted versions of the BNT. For instance, in the Greek version, four items were replaced (pretzel, doorknocker, stethoscope and scroll) (Patricacou et al. 2007). Cruice et al. (2000) proposed substituting "beaver" and "pretzel" with "platypus" and "pizza" for an Australian version of the test, while in Argentina, the entire order of the items was rearranged since the difficulty level of the original ordering was altered when applied to a local sample (Allegri et al. 1997). Items such as "igloo," "beaver" and "acorn" were found to be much more difficult for the Argentineans than for the North Americans. This fact even affected the cross-cultural concurrent validity of the test (Fernandez and Fulbright 2015).

Another example of item bias is the Famous Face Recognition and Naming Test (Rizzo et al. 2002), developed in Italy to evaluate semantic memory. This test is comprised of 50 pictures of famous people and pictures are presented to the testee whose task is to name the subject of the picture. Among the famous people included are Harrison Ford, Stefi Graf, Benito Mussolini, Rosa Russo Jervolini, Dino Zoff, Diego Armando Maradona, Adolf Hitler, Rosy Bindi, Fiona May, Giuliano Amato, y Rita Levi Montalcini. As can be observed, many of these people might be familiar to Italians, but probably unfamiliar to people from other countries.

One last example of item bias can be found in a cross-cultural study with a calculation and number processing battery (EC301; Dellatolas et al. 2001). The EC301 battery was administered to three samples of European subjects, Italians, French and Germans, with results showing that French individuals made significantly more mistakes than Italians and Germans on several subtests (i.e., spoken verbal counting, enumeration of dots and mental calculation on spoken verbal numbers). The authors speculated that the observed differences could be the result of "special complexities of the French verbal code for numbers (e.g., 70 is soixante-dix, i.e., sixty-ten; 80 is quatre-vingts, i.e., four-twenty)…" (p. 852).[1] According to this explanation, it might be hypothesized that these language complexities might put a higher load on the phonological loop subsystem of the working memory which, in turn, might impair the efficacy of this system.

Language influence

There is one more factor affecting the neuropsychological performance in cross-cultural testing: the language that the testee speaks. Beyond the problems associated with translation, the language in which the test is given can introduce differences when two different language groups are compared. For example, if two groups of individuals are matched in all the significant variables (education, age, sex, etc.), and the test is culturally appropriate for both [equivalent] samples, individuals from group A (Chinese speakers, for example), might be better at a test (remembering digits, for example), than individuals from group B (Spanish speakers, for instance). Strictly speaking, and according to van de Vijver and Tanzer's definition of bias, this kind of influence of language does not constitute testing bias, because this truly reflects differences in ability. Lau and Hoosain (1999), for example, have shown that Chinese speakers are better than Japanese speakers in a mental arithmetic task, and Japanese speakers, in turn, are better than English speakers. They were able to demonstrate that these differences were related to the sound duration of digits, and hence related to working memory. By administering an articulatory suppression task they eliminated these differences between the three groups, thus supporting the hypothesis that shorter sound duration of digits in each language gives an advantage to the subjects of each group. Likewise, when Kempler et al. (1998) administered an animal fluency test to a multi-ethnic group, and after controlling for the effects of

[1] A similar case is that of the Chinese, where multiple-digit numbers are constructed in the following way: first the digit itself, then the place (tens, hundreds or thousands, for example), and finally the next digit. For example, 65 is six tens (and) five ones.

age and education, they found that the Hispanic group produced significantly fewer animal names than the Chinese, White, and Vietnamese groups. They also found that the Vietnamese produced more animal names than the Chinese, White, and Hispanic groups. The most striking difference was found between the Hispanic (12.8 ± 3.9) and Vietnamese (17.3 ± 5.2) sample. A more detailed analysis showed that animal names are longer in Spanish than in Vietnamese. For example, the word "dog" in Spanish is *perro*, while in Vietnamese is *chó*; "rabbit" is *conejo* and *thỏ*, respectively; and "elephant" is *elefante* and *voi*, respectively.

In addition, López et al. (2016) found that bilingual individuals performed differently on a digit-span test, depending on whether they were tested in English or Spanish. Their performance was significantly better in English for the Digit Span Total score (Forward plus Backward scores), but if they analyzed the number of syllables recalled, the performance was better in Spanish. The authors hypothesize that this finding could result from different cognitive strategies used in every language to solve the test.

In summary, it is clear that neuropsychology is not free from the influence of cultural variables. And it is also quite clear that these cultural influences can affect neuropsychological testing from the perception and creation of constructs to the specific items used to test these constructs.

Strategies to address cross-cultural bias

Having considered the evidence of the influence of cultural variables in all stages in neuropsychological testing, what strategies may be used to address these challenges? Below, we describe several strategies suggested by different authors.

Van de Vijver and Tanzer (2004) gave a complete description of possible strategies to address each type of bias in psychological testing. Among the remedies they propose is "decentering," a method used to avoid construct bias-that is, to exclude words or concepts that are clearly specific to one particular language or culture—through the simultaneous development of an instrument in several cultures (see Tanzer 2005). The use of extensive training of administrators, an assessment of response styles and the development of a detailed protocol and instructions are among the strategies proposed to deal with method bias. Finally, in the case of item bias, for instance, they propose the use of psychometric (such as differential item functioning analysis) or judgmental (such as linguistic and psychological analysis) methods of item bias detection. Simultaneous development of an instrument in several cultures can substantially reduce sources of bias involved in test adaptation. Nevertheless, there are some prerequisites that must be fulfilled to achieve this aim (Tanzer 2005), for instance, it is necessary to establish construct equivalence, a system of classification for multicultural/multilingual bias and a taxonomy of strategies to respond to possible sources of bias.

Other investigators (Nell 2000) advocate the development of a behavioral neuropsychology approach, based on the assumption that most neuropsychologists in developing countries—where neuropsychology and health systems are not fully developed—will mostly face cases of diffuse brain damage. Based on his own research and other studies, for instance, Nell demonstrates that most of the accidents affecting the brain worldwide cause diffuse brain injuries (Brown and Nell 1991; Nell and Brown 1991). For example, he mentions that, at the time, traffic accidents caused 73% of all brain injuries among White males in South Africa. In addition, 51% of all brain injuries among Black South Africans were caused by blows with blunt weapons. Both of these conditions usually produce diffuse brain injury rather than highly specific lesions. In the behavioral neuropsychology approach, the use of testing is avoided and instead, the extent of brain injury is evaluated according to a comprehensive description of the behavior and functioning of individuals in four domains: arousal, personality, thinking, and physical functioning. Nell (2000) proposes the standardization of a common core neuropsychological battery, and based on Luria's core principle of "bringing the zone of proximal development into testing," to provide testees with enough practice trials before the administration of the actual test in order to facilitate their familiarization with the testing materials and response procedures. Nell specifically suggested the use of the WAIS-III and the Wechsler Memory Scale as a core battery based on their excellent psychometric properties and the use of practice items, proposing the addition of other tests such as the Trail Making, Tower of London, or Token Test as supplementary instruments.

Using the assumption that there are no universal constructs, Caetano (2007) promotes a holistic-qualitative approach to neuropsychological assessment and rehabilitation. Her assumption is heavily based on the use of intelligence testing which, in her view, represents a prototype of the bias that can be encountered when psychometric tests are used cross-culturally. In this approach, the use of Luria's Neuropsychological Investigation (LNI, Christensen 1975) is proposed for use at the level of impairment analysis, and the use of observational methods is used at the level of functional limitations.

In interpreting test results, the development of appropriate normative data for different cultural groups help make the neuropsychological tests more suitable for use with these groups (Nelson and Pontón 2007). When such normative data are unavailable, Nelson and Pontón (2007) suggest a qualitative approach to the neuropsychological assessment, using the LNI based on the evidence of its efficacy (Christensen and Caetano 1999). However, this evidence may be specific to particular instances; the cases presented by Christensen and Caetano (1999) are mainly patients with focal lesions so the validity of this approach in cases of diffuse lesions or subtle deficits remains uncertain. Besides, very little work has been done to validate the LNI or test its reliability among individuals with diffuse lesions or subtle deficits. Finally, Ardila proposes a number of strategies to address these biases (Ardila 1995, 2005, 2007a, b; Ardila and Keating 2007; Ardila et al. 2006). First, he proposes that psychometric testing should be performed in those societies with a strong tradition in the use of these tests; otherwise a qualitative approach is preferred. Second, regarding test development he advocates the adaptation of tests, stating that"…neuropsychological tests must be adapted (i.e., redeveloped; not just

translated)…" (Ardila 1995, p. 148), and then "…re-developing cognitive tests according to the cultural conditions" (Ardila 2005, p. 192). Nevertheless, he concludes that "culture-free cognitive tests are simply impossible" (Ardila and Keating 2007, p. 120). Third, with respect to the use of culture-specific norms, his suggestions have evolved over time, from a call for "…the normalization of current basic neuropsychological instruments in different cultural contexts, somehow representing a sufficient broad sample of the humankind species (e.g., Amazonian Indians, Eskimos, Australian aboriginals, etc.)…" (Ardila 1995, p. 148; see also Ardila 2005), to an acknowledgment that obtaining norms for every language and/or cultural group is an unrealistic endeavor given the enormous number of languages (estimated at 6800) and cultural groups (several thousands) existing in the world (Ardila 2007a, b). As a consequence, he asserts that an understanding of the variables that affect test performance is the most important issue (Ardila and Keating 2007; Ardila et al. 2006). Although the authors do not elaborate more on this issue, the rationale behind their argument seems to imply that once these critical variables affecting test performance are identified, the tests may be normed in one cultural group and then used with other cultural groups, provided that the populations are comparable with respect to these identified critical variables.

A three-level solution

First level: the behavioral neuropsychological approach

In our view, these strategies are categorized at three different hierarchical levels. At the most basic level, a *behavioral neuropsychology approach* seems to be the most appropriate option. Along with other authors (Ardila 2005; Nell 2000; Nelson and Pontón 2007), we concur that this option is most appropriate for those contexts where health systems are not developed, and/or there is a lack of resources or infrastructure for the practice of more comprehensive neuropsychological examinations, or any other conditions under which the use of tests could be seriously misleading. An example of the latter might be the administration of a current test to an immigrant who does not speak, or barely speaks, the interviewer's language, with few years of education and who comes from a non-Western country. In the case of the assessment of non-literate individuals or those with very few years of education, for instance, semantic memory might be evaluated by asking the person to recall oral traditions, and procedural memory by watching them perform some everyday tasks. This approach, however, involves some important risks. First, it is very subjective. Neuropsychological assessment is always a question of measurement and without the use of objective tests, the testing itself becomes a subjective measurement. Judging if the performance of a given patient is within the expected or "normal" range can be very difficult and the interviewer can make substantial mistakes if he/she does not have a clear frame of reference with which to judge what would be considered normal performance. For example, how much forgetting of the oral tradition is necessary to be considered a performance below normal? Saying that the patient is not impaired, or saying that he/she has a low, moderate or severe

impairment is a quantitative measure, but is more subjective in nature, and as such, prone to mistakes. Second, qualitative assessments demand highly trained clinicians (Witsken et al. 2008). Because of the difficulty of subjective measurement, only very experienced clinicians, who have developed an internal measure of normal performance through the repeated observation of normal and clinical cases, can perform an adequate assessment. In this instance, then, how many cases are necessary to develop this ability? How many years does it take that training? Third, inter-rater reliability is challenged. Would two different clinicians, within this approach, make the same diagnosis? Would they arrive at the same conclusions? Since there is no standardized method in the data collection they might arrive at different conclusions based on the method by which the data were collected. It is well known within the neuropsychology field that different measures can elicit different behaviors even when they are intended to measure the same construct. For example, perseverative responses are often seen with graphic tests and not with verbal tests and vice versa (Witsken et al. 2008; Lezak et al. 2012). Likewise, in a qualitative assessment one clinician could ask a patient to perform certain everyday activities, while a different clinician could ask for others. These everyday activities might not be equally impaired and, as a consequence, lead both clinicians to different conclusions. Fifth, as others have noted (Holtz 2011), this kind of approach is not sensitive to subtle deficits in cognition. For instance, forgetting is a normal experience in everyday life (Baddeley 1999), but a deeper forgetfulness is an early sign of a degenerative disease (Cullum and Lacritz 2009). Thus, how can a clinician detect these early signs without a clear measurement reference? How many times a day and what information must a person forget in order to affirm that he/she has early signs of dementia? This picture becomes even more complex when we consider the different educational levels of the patients. How much information loss for a person with low educational level represents impairment and how much for an individual with a high educational level? This lack of sensitivity ultimately undermines the validity of the process.

Nevertheless, despite all the weaknesses of this approach it would seem to represent a better alternative than using biased neuropsychological tests when we are working with Non-WEIRD (or non-Western, Educated, Industrialized, Rich, Democratic, or WEIRD) populations (Henrich et al. 2010). Just as placing the hand on the forehead of a feverish individual might be more effective to estimate the temperature than the use of a defective thermometer, so it may be better to use a behavioral neuropsychological approach than to use neuropsychological tests inappropriately.

Second level: test adaptation

At an intermediate level, the adaptation of current neuropsychological tests represents a preferred strategy. There are several findings suggesting the effectiveness of this approach. The experience with the adaptation of several tests (Porteus Mazes Test, Trail Making Test, Mattis Dementia Rating Scale, Stroop Test, Controlled Oral Word Association Test, among others) in different countries, has yielded positive results (Böhm et al. 2005; Fernandez and Bendersky 2004;

Fernandez et al. 2002, 2004; Fernandez and Scheffel 2003; Kim and Kang 1999; Kabir and Herlitz 2000; Konstantinopoulou et al. 2011; Marino et al. 2001; Messinis et al. 2011; Nampijja et al. 2010). The major limitation of the adaptation approach, however, is that it represents an endless task. Given the enormous number of cultures and languages existing in the world, adaptation of current tests implies a task of huge proportions that simply is not feasible. In addition, comparisons between the performances of subjects would be very arduous since establishing the equivalence between several versions or adaptations of an instrument would be difficult to achieve. Finally, the process of adapting and norming tests is not feasible in many developing countries simply due to a lack of resources, financially and professionally (Ruffieux et al. 2010). Therefore, although the adaptation and norming of instruments represents a potentially higher testing standard than the behavioral neuropsychological approach, an alternative response is needed for the future. Adapting a test would then only be the preferred option when the third level is not possible.

Third level: a new generation of tests

Psychometrically and culturally improved tests

Lastly, the third level involves the simultaneous development of completely new neuropsychological tests. We agree with Tanzer (2005) that this method could be highly advantageous, since it avoids all the problems of adaptation. These tests should be developed, based on current knowledge of cross-cultural testing in order to make them appropriate for use across a variety of cultures. Moreover, the development of these tests would allow the raising of a new and improved generation of neuropsychological tests. In addition to issues raised regarding the lack of cultural fairness, many of the current tests have psychometric weaknesses such as multifactorial composition, inadequate extension, asymmetrical normative score distribution or inappropriateness for certain population groups such as for individuals with low levels of education. A multifactorial composition is a very common and undesirable feature in many neuropsychological tests since it is a confounding factor when interpreting scores. For example, it appears that the Raven's Standard Progressive Matrices (SPM) test and the Advanced Progressive Matrices (APM) tap different cognitive functions and, at least, two factors, perceptual and analytical, have been found (Mackintosh and Bennett 2005). A similar situation can be described with the Wisconsin Card Sorting Test (WCST), which, although it is considered an executive functioning test, encompasses three different cognitive functions: the ability to shift set, problem solving/hypothesis testing, and response maintenance (Greve et al. 2005). Therefore, a low score could stem from impairment in different cognitive processes which would lead to very different neuropsychological interpretations. Moreover, despite being an excellent test for some purposes, it has been called "a one shot test" since it can only be applied once. That is, because the WCST implies that the patient has to discover the sorting and shifting cards principle, once he/she has solved the principle, it cannot be applied again (Lezak et al. 2012; Paolo et al. 1996). Another example of a test

with psychometric difficulties is the Boston Naming Test, which has been seriously questioned because of its ceiling effect and multiple psychometric flaws when applied outside the U.S.A, even after an adaptation process (Fernandez and Fulbright 2015; Harry and Crowe 2014). For its part, the Stroop Test, a widely used attention test, cannot be administered to people with reading difficulties or very low literacy because of its heavy emphasis on reading abilities.

Besides, many tests have been developed for research purposes and then transferred into the clinical setting. As a consequence, practically speaking they are often overly extensive and artificial for use in a clinical setting. These characteristics make them unsuitable for many situations such as bedside assessments or assessments in contexts with significant time constraints, such as in public hospitals in developing countries. In addition, many patients find it difficult to understand the relationship between the test and his/her problems which undermines the face validity of the instrument, and conspires against the cooperation of patients because of their tendency to become fatigued in taking the tests. Examples of this include the Paced Auditory Serial-Addition Task (Gronwall 1977), or the WCST.

A new generation of tests could encompass characteristics of both culture fairness and psychometric robustness. Tanzer (2005) depicts a series of problems that can be encountered when developing simultaneous tests and the remedies that can be applied. These result in a set of basic principles that may be useful in the development of simultaneous tests. First of all, these tests should contain culturally fair items. The content and format answer of the items should be easily understood by people no matter where they are from, whether Europe, Asia, Africa, America (North, Central, South), or Oceania. As such, they should avoid cultural/linguistic particularities. Thus, instead of using the alphabet as stimuli, for instance, it would be preferable to use numbers, or colors, since the alphabet is different in many languages (Latin, Chinese, Greek, Arabic, Cyrillic, Japanese, and so forth). In fact, this is precisely the way that the CTT (using colors and numbers) was developed, as a cross-cultural alternative to the TMT (it uses numbers and letters) (Maj et al. 1993).

Tests applicable to individuals of different educational levels

A new generation of tests should be developed in such a way that they can be applied to people with different educational levels. Formal education is one of the most powerful variables influencing neuropsychological test performance (Ardila and Rosselli 2007), probably even more so than culture (Ostrosky-Solís et al. 2004). The Wechsler intelligence scales might be a model for such an approach, in that each subscale is designed to be applied to children of different ages, and thus start with very easy items and progress to more difficult ones. The test is also discontinued after several consecutive wrong answers. In this way, individuals with low levels of education may take the same test that individuals with high education, but the latter individuals would probably be administered more items than those with low levels of education. If the test format does not allow such an approach, then appropriate items should be identified, that can be understood by every educational group. For example, instead of school related items (e.g., letters,

mathematical operations), everyday objects and operations (e.g., counting objects, bills, receipts and so forth) should be used as test items.

The assessment of language functions

Although language assessment is a challenging issue for the construction of neuropsychological tests due to the huge variability in the length of words, syntax, meanings, and the use of words across languages, there are some language tests that have demonstrated cross-cultural validity. For example, although the number of items in the Animal Generation Task (Kempler et al. 1998) differed between some languages (Spanish, English, Chinese and Vietnamese), the construct remained stable across them, with, 83% of all subjects able to provide more than 10 animal names in the allotted time period, providing evidence of the cross-cultural appropriateness of the instrument (Ardila et al. 2006). Also, Ardila (2007b) proposed a cross-linguistic naming test which has been recently tested with three different cultural groups (Gálvez-Lara et al. 2015). Although the test was administered to small samples and two of them spoke the same language (Spanish), initial results appear promising.

Short but reliable tests

The appropriateness of the implementation of a new generation of neuropsychological tests is a very important issue, especially if these tests are assumed to apply across diverse populations and cultural groups. Neuropsychological services are poorly developed in many parts of the developing world and most of those populations cannot afford the luxury of a 10–12 h-long testing session. Therefore, these new generation of tests should be designed in order to maximize the time/information ratio, i.e., to obtain the maximum information possible within the shortest time. Certainly, psychometric theorists recommend longer tests in order to improve reliability, but short tests such as the Mini-Mental State Examination (MMSE), the Montreal Cognitive Assessment (MoCA) or the Short Portable mental Status Questionnaire have demonstrated good reliability coefficients (Lehser and Whelihan 1986; Nasreddine et al. 2005). Across several studies, for instance, the MMSE has demonstrated, a reliability coefficient between .80 and .95 (Tombaugh and McIntyre 1992), while the MoCA showed high internal consistency (Cronbach's alpha = .83) and test–retest stability (r = .92) coefficients. Nevertheless, Cullen et al. (2007) have acknowledged that "an administration time of more than 10 min appears to be an unavoidable cost of achieving sufficiently robust statistical performance while covering key domains" (p. 795).

Administration competencies

A question arises that if the screening measures are short and relatively simple, would it be appropriate and/or eventually expected that non-professionals would adminster these instruments? In response, it is important to note that even when short tests can be developed, the combination of different short tests, each one

targeted at one cognitive function, might result in a battery that will take more than 5–10 min to administer. Thus, a well trained professional in neuropsychological testing should administer these tests rather than lay persons. Besides, although these tests might be considered a screening battery in many settings, in many other settings, the administration of this short battery might be the only opportunity to test a patient at all (for example, in contexts where neuropsychology is not well developed). This fact demands that emphasis is placed on obtaining the maximum quality of information possible through the application of this battery. Moreover, the qualitative information that emerges in how the patient approaches the testing situation and the behaviors exhibited during the session can only be assessed by a trained neuropsychologist. Consequently, in our opinion, neuropsychologists or well trained psychometricians are best suited to administer these instruments.

Ecological validity

Although the ecological validity of the neuropsychological tests is an issue of great controversy, an effort should be made to ensure that new tests demonstrate greater ecological validity. The main purpose of neuropsychological tests is to assess the status of different cognitive functions; it is clear that the prediction of a patient's functioning outside the testing room depends on many other variables that a neuropsychological test does not, and cannot, address (Sadek and van Gorp 2010). However, the inclusion of items that resemble the activities of daily living might give the examiner a better idea of how the patient could perform that activity in his/her daily activities than if he/she completes tasks that are artificial and restricted to laboratory settings. Besides, ecologically-valid items provide tests with more face validity, which leads to better cooperation of the patient with the testing process. For example, instead of performing school-like mathematical operations patients could be asked to count real objects,or to add receipts or shopping items.

About the inclusion of practice items

The value of including practice items in maximal performance tests has been emphasized by some authors, although their approach is different. This is especially important for non test-wise populations or individuals with low levels of education. Nell (2000) proposes the incorporation of an extended practice period with test stimuli before the actual administration of the test, along with a guided learning experience, that is, a coaching procedure with explicit instructions about how the test works and what is expected from the examinee. Nell states that "…the period of extended practice…will bring clients to the asymptotic performance before the test proper begins" (p. 176). Test-wise individuals should receive the standard administration described in the manual. Tanzer (2005), however, suggests the inclusion of hidden warm-up items at the beginning of the test, and excluding those items from later scoring. Nonetheless, in his view, the number of hidden warm-up items will depend on the findings of local studies that should be completed for every culture, considering that the number of necessary items may vary across cultures.

Although the rationale for suggesting the inclusion of practice items before the test is sound, the issue is controversial and its application very difficult. Several problems arise with these proposals. First, giving extended practice to some subjects and not to others can introduce an advantage to the former. Nell (2000) assumes that individuals with at least 12 years of formal education are test-wise; we cannot be sure, however, that the level of comprehension of a task to be completed would be equivalent for someone with 12 years compared to that of an individual with 18 years of formal education. Besides, how much more test-wise is someone with 12 years of education than someone with 11 years? The latter person might benefit much more with the practice than the former, thus introducing a serious bias in the measurement process. Although Nell warns that the extended practice should be utilized even with some individuals who have more than 11 years of school (e.g., "…the victims of a failed educational system…", p. 177), it is very difficult to estimate the level of test wiseness of any given individual. Second, how many extended practice items would be considered necessary in each case? Tanzer proposes the use of a different number according to the findings of local studies. Nevertheless, adapting the warming-up period to each particular cultural context would lead us to the same problem as we have with normative data: it is an endless task. For a cross-cultural test, whose format and administration procedure should be standardized across cultures and educational groups, it must be clear how many warm-up or practice items have to be administered and how to score those items. Third, administering several items to the patient could result in a rather long test. As stated above, short tests are preferable in order to make them practical for use across many settings where time and financial resources are scant. Fourth, the question arises as to when the scoring of a patient's performance should actually begin. Although studies regarding this topic are rather scarce, some evidence seems to support the idea that the fundamental increase in scores due to practice after repeated administrations occurs between the first and the second test interval (Beglinger et al. 2005; Collie et al. 2003; Falleti et al. 2006). Moreover, this effect has also been found in the administration of a computerized neuropsychological test battery to a sample of Indigenous Australians adolescents (Dingwall et al. 2009). This effect, however, does not manifest equally across different domains. (Beglinger et al. 2005; Collie et al. 2003; Duff et al. 2012; Falleti et al. 2006). To sum up, then, the issue of the inclusion of extended practice items has a sound basis, but its effective application is very difficult to achieve; therefore it remains an unresolved issue that demands creative and sophisticated solutions.

Who should develop these new tests?

An important issue regarding this new generation of tests is who should develop them. Should they be developed by research groups at universities, research at clinical facilities, or by commercial publishing companies? In regard to this topic, there is always a tension between scientific and commercial issues. Therefore, the advantages and disadvantages of each option should be considered. Developing tests is costly in human terms and financial resources. The main advantage of a test developed by non-profit working groups is that the resulting tool can be freely

distributed and therefore it is more easily accessible to all professionals, especially those in the developing world. However, the test will not be marketed, thus becoming an unknown tool for many neuropsychologists. Moreover, the human and financial cost of developing tests will probably make this endeavor feasible only to research groups working in high income countries.

In contrast, a test developed by a commercial publishing company will be marketed and its use promoted, but two main financial issues arise: (a) companies might not be interested in developing tests that will probably will be highly demanded in the developing world where they have less opportunities to make profits; (b) the cost of the test might be prohibitive for middle and low income countries. All things considered, it seems that two options emerge here: a) these tests should be developed by non-profit research groups at universities and/or research facilities that then make an effective effort to promote them or; b) they should be developed by commercial publishing companies that could find the way to make profit of them at a reasonable price for countries of the developing world.

The use of modern technologies

With the profusion of new technologies these days it is worth to wonder how these new generation of tests might benefit from them. Elements such as tablets, cell phones and applications are progressively being introduced in the testing field. While the introduction of these technologies has advantages such as the reduction of error at the scoring level, the transportability, the simplicity to develop multiple forms of a test or the multisensory possibilities for the stimuli design, it also has disadvantages. Less educated people or people living in low-tech societies might find difficult to understand how these technology operates, thus, hindering their performance. As a result, the introduction of technological devices in the development of these new tests should be carefully considered on an individual basis considering the advantages and disadvantages in each case.

New psychometric models

The vast majority of the current neuropsychological tests are developed under the classical test theory (CTT). Nevertheless, a more recent theoretical development in the psychometrics field has become more and more popular in the last several years, specifically, item response theory (IRT). The IRT represents some advantages over the CTT. Hambleton and Jones (1993) summarize them in the following points: (1) "item statistics that are independent of the groups from which they were estimated; (2) scores describing examinee proficiency that are not dependent on test difficulty; (3) test models that provide a basis for matching test items to ability levels and; (4) test models that do not require strict parallel tests for assessing reliability" (p. 44). As can be observed, IRT could offer some tools that can be very useful to approach cross-cultural issues, such as the item independence from the estimation group. Nonetheless, the use of IRT has also some disadvantages such as: (1) the need for large samples (usually over 500 cases), very complex mathematical analysis, difficulties in model parameter estimation and the need for "strict goodness-of-fit

studies to ensure a good fit of model to the test data" (Hambleton and Jones 1993) (p. 40). The issue regarding large samples is of concern in the neuropsychological testing field because more of these tests are administered individually. As a result, collecting more than 500 cases is a very difficult goal to achieve. Yet, approaches like TestMyBrain have found a way to collect massive amount of cases quite easily. TestMyBrain is a battery of Internet-based tests that any individual can take on-line (Germine et al. 2012). By using this method the authors have been able to collect thousands of cases (more than one hundred thousand subjects have completed the Famous Faces subtest up to March 2017). This is an imaginative way to solve the problem of large data collection. Nonetheless, due to the technological nature of this battery, only individuals with Internet access (not to be taken for granted in the developing world) can take such tests. Besides, the tests should be available in numerous languages in order to make them accessible for individuals of multiple cultures (TestMyBrain, for example, is only available in English). Moreover, designing computerized self-administered tests for some cognitive functions is very challenging, such as the case of tests that require oral responses.

The IRT approach has very attractive advantages but it does not seem to be by chance that so few neuropsychological tests have been developed using this theory. It is quite clear that IRT is more appropriate for group administration tests and designing group administration neuropsychological tests have proven very challenging so far.

Universal cut-off scores instead of infinite normative data

Finally, an alternative to the problem of local normative data is critically needed. We propose to address this issue from a different, pragmatically-based approach. Although normative data are important in order to accurately determine the level of performance of a given subject, the central objective of the neuropsychological assessment is to determine if the subject is cognitively impaired or not. In line with this goal, it may be more useful to identify the critical cut-off score for a given test than to focus on specifying the exact performance level for a subject. For example, once the score of an individual being assessed falls within the unimpaired range, the task of determining whether the subject's performance is within the "above average" versus "average" level is a secondary priority. The main concern is to determine that the subject is not impaired. A similar line of reasoning can be applied if the subject's performance is within the impaired range—that is, it is more important initially to identify the presence of impairment, than the precise level of impairment. Thus, the argument here is NOT that the precise positioning of the subject's performance is not important, indeed it is very valuable, but that if a test can be developed in such a way that can be validated and used cross-culturally by using a cross-cultural cut-off score, then abandoning the precise performance scaling represents a reasonable price to pay for the development of such an instrument. The use of cut-off scores for screening is a common approach in neuropsychology. Numerous studies have implemented cut-off scores instead of developing norms for the rapid screening of neuropsychological impairment (Alegret et al. 2013; Konstantinopoulou et al. 2011; Kwak et al. 2010; Sacktor et al.

2005; Zuo et al. 2016). This approach has been specially used in situations in which developing norms is very difficult because of the limited available resources. For example, Sacktor et al. (2005) developed the International HIV dementia scale which consists of three short subtests. They administered the scale to two samples, one from the U.S.A and the other from Uganda. The authors were able to find a cut-off score that yielded an 80% sensitivity in both samples, and 57% specificity in the U.S.A sample and 55% in the Uganda sample.

In accordance with Ardila (2007a, b), in our view, adapting and norming a whole battery of neuropsychological tests for all the cultures and languages is an unrealistic endeavor. Yet, developing a test with a cut-off score that can be validated for multiple languages and cultures is possible. The validity generalization approach in psychometrics uses meta-analysis to correlate a test score and a criterion and may be useful in this regard. Namely, by gathering several criterion validity studies and applying meta-analysis to the results obtained in these studies, it would be theoretically possible to find a coefficient that summarizes the relationship between the test score and the criterion variable. Thus, by accumulating enough validity evidence on the ability of the test to predict the behavior of the subject in a given situation, it may be considered reasonable to use the test in a new situation without the need to conduct a local validity study (Murphy 2003; Society for Industrial and Organizational Psychology 2003). This approach has been mainly used in the industrial psychology field. Thus, it would be reasonable and necessary to run a concurrent validity study of such a cross-cultural test based on the use of a cut-off score that could allow us to classify subjects as cognitively impaired or non-impaired. After this, replicating this study in a sample of different cultures would allow us to obtain enough evidence to run a meta-analytic study of the sensitivity and specificity of the test (Cleophas and Zwinderman 2009; Hasselblad and Hedges 1995). This procedure would allow a cross-cultural cut-off score or a cut-off score range to be identified. Consequently, the validated cut-off score or cut-off score range could be used in other cultures where the test has not been validated based on the results of the validity generalization study.

Certainly, this approach has its pros and cons. The main advantages would be: (a) to have tests that can be easily translated into different languages without costly adaptation processes; (b) neuropsychological services would be possible in environments where otherwise it would be very difficult to establish; (c) the development of neuropsychological services would have impact on public health, since earlier and more accurate diagnosis would allow better targeted treatment options and consequently better health conditions for the population of that society, and; (d) neuropsychology as a clinical and academic discipline would emerge invigorated with greater accessibility and renewed relevance across more cultural contexts and practice settings. Yet, there are also disadvantages in such an approach. First, the use of a cut-off score might lead to diagnostic mistakes. Even when the sensitivity and specificity of these tests might be maximized, there would always be a certain proportion of patients who would be incorrectly identified. The lack of additional tests that might help avoid or limit this problem is a definite shortcoming. However, this might be compensated for by improving the clinical training of the practitioners who could find other sources of data (behavioral observation, school

and work records, family interviews, etc.), to supplement the information gathered with the tests. Second, not having the opportunity of accurately grading the performance of the patient may lead to under- or overestimating his/her difficulties. Again, training and experience can help ameliorate this negative outcome. Finally, the possibilities of performing research that can address some theoretical issues with these instruments are reduced, mainly due to its simplicity, i.e., because these instruments would not test any cognitive function in deep (they are meant to be short), they would not be appropriate to investigate some cognitive processes in detail. For example, a memory test with these characteristics probably would not describe in detail the different cognitive processes and stages of memory as a test like the California Verbal Learning Test would do (Delis et al. 2000).

Even when tests like these might not be possible for every cognitive function, at least they would cover a greater spectrum than what is currently possible. For those particular aspects of functioning that could not be assessed with such a test (e.g., some language functions, for example), such functions could be addressed by developing/adapting a specific test for that language/culture. In that way, efforts needed to develop a whole assessment battery for a particular cultural context could be reduced to developing just a few key tests. Although changes might be necessary in order to adapt them to some cultures, such changes would be relatively minor, or at least, much less than what would be necessary with the neuropsychological tests currently in use.

Current cross-cultural neuropsychological tests and batteries

Although this proposal seems ambitious, there is evidence that such an approach is feasible. Indeed, some researchers have already made some specific attempts in order to develop cross-cultural, or culture-fair, neuropsychological tests (see Table 1 for a list of cross-cultural neuropsychological tests). One noteworthy characteristic of these tests is that most of them have been designed to assess dementia in the elderly; there are few if any, tests that have been cross-culturally developed to assess general neuropsychological functioning in adult or child populations. In the following paragraphs, some of these instruments will be reviewed.

Dick et al. (2002) describe the Cross-Cultural Neuropsychological Test Battery (CCNB), which demonstrated appropriate validity indices. However, no further publications were found by these authors on the use of this battery. Only two subtests of the CCNB seem to have survived, the Cognitive Abilities Screening Instrument (CASI) (Teng et al. 1994), and the Common Objects Memory Test (COMT) (Kempler et al. 2010). The CASI is an instrument that was developed in order to be used in cross-cultural dementia epidemiological studies. Its short form has been used in studies across different languages, including English (Teng et al. 1994), Japanese (Teng et al. 1994), and Portuguese (Rezende et al. 2013). There is also a Chinese version of the extended form of the CASI (Tsai et al. 2007). The short form of the CASI includes four sections: repetition of three words, temporal orientation, verbal fluency (four-legged animals in 30 s) and recall (3 words). The

Table 1 List of cross-cultural neuropsychological tests and batteries

Test/battery	Characteristics	Reference
Culture-fair assessment of neurocognitive abilities (CANA)	Dementia screening in multi-ethnic populations	Amin et al. (2003)
Cognitive abilities screening instrument (CASI)	Dementia screening in cross-cultural epidemiological studies	Teng et al. (1994)
Cross-cultural cognitive examination (CCCE)	Assessment of dementia in multi-ethnic populations	Glosser et al. (1993)
Cross-cultural neuropsychological test battery (CCNTB)	Assessment of dementia in multi-ethnic populations	Dick et al. (2002)
Consortium to establish a registry for Alzheimer's disease (CERAD)	Dementia screening in cross-cultural epidemiological studies	Maj et al. (1993)
Community screening interview for dementia (CSI 'D')	Dementia screening in cross-cultural epidemiological studies	Hall et al. (2000)
Common objects memory test (COMT)	Assessment of memory in multi-ethnic populations	Kempler et al. (2010)
Repeatable Battery for the Assessment of Neuropsychological Status (RBANS)	Assessment of dementia in multi-ethnic populations	Randolph (1998)
Rowland Universal Dementia Assessment (RUDAS)	Dementia screening in multi-ethnic populations	Storey et al. (2004)
Spanish and English Neuropsychological Assessment Scales (SENAS)	Assessment of elderly in Spanish and English	Mungas et al. (2004).
WHO Neurobehavioral Core Test Battery	Assessment of neurotoxic effects in cognition in cross-cultural studies	Johnson et al. (1987)

CASI gathers two important characteristics for a cross-cultural test: brevity and cultural fairness. Unfortunately, some of the items do not seem very appropriate for low educated individuals and, indeed, education has a significant effect on its score (Damasceno et al. 2005; Teng et al. 1994). This makes it less appropriate for environments where most of the individuals have a low educational level, and demands norms or different cut-off scores according to the educational level.

The COMT is a visual/verbal memory test that was designed to bypass the difficulties observed with word list tests. It is comprised of ten color pictures of objects familiar across cultures that are shown to the examinee in three consecutive trials. The subject must name the objects and after each trial, must recall the objects that he/she saw. There is also a 5-min delayed recall trial and a recognition trial immediately after the delayed recall. It has been administered to five culturally and linguistically distinct populations: Caucasian and African-American English speakers, as well as native Chinese, Spanish, and Vietnamese speakers. The recall score of the COMT was found to adequately differentiate individuals with dementia from healthy controls. The COMT seems to be a very good option for cross-cultural

memory testing since its stimuli are ecologically appropriate and suitable for different cultural settings, it is short, it does not need a translation and it has little education effect. Moreover, it has also good validity properties. The authors have relied on a standardization approach, but by increasing the demented patients in each group, a common cut-off score might be found for these groups. This, in turn, may lead to a validity generalization of the cut-off score.

The Repeatable Battery for the Assessment of Neuropsychological Status (RBANS) (Randolph 1998) is another brief instrument (30 min approximately) that evaluates four cognitive functions: attention, language, visuospatial/constructive ability and memory, and was designed for testing adults between 12 to 89 years of age. Although this is not designed as a cross-culturally battery, it has been adapted to other languages (Armenian, Chinese, Hungarian, Russian and Spanish) and it is currently in use as a cross-cultural tool (Azizian et al. 2011; De la Torre et al. 2014; Lim et al. 2010). Among the most important features of RBANS as a cross-cultural test are its brevity and its simplicity for the adaptation process. Besides, it contains verbal tasks that seem to be quite amenable to adaptation. However, these elements are not universal, therefore they require and adaptation rather than just a translation. Other issues are the low ecological validity of some stimuli and the high influence of educational level (Garcia et al. 2008; Gontkovsky et al. 2002).

The Spanish and English Neuropsychological Assessment Scales (SENAS) (Mungas et al. 2004; Mungas et al. 2005; Mungas et al. 2000; Mungas et al. (2005), is a battery developed using item response theory methods. This battery was developed from the beginning in two versions: Spanish and English. This battery is targeted for use with older adults, and takes between 2 and 4 h to administer (Mungas 2006). It does not have a commercial distribution. The SENAS has very good psychometric properties and demonstrated measurement invariance between Spanish and English-speakers (Mungas et al. 2011). Nonetheless, its limitations include the length of the battery and its limited use for older adults. Although developed as a cross-cultural test, it has been specifically designed for English and Spanish speakers, and its appropriateness for other cultural groups remains to be demonstrated.

The most promising cross-cultural neuropsychological test is unquestionably the Rowland Universal Dementia Assessment (RUDAS) (Storey et al. 2004). This is a very brief test that usually takes no more than 10 min to administer. With a score ranging from 0 to 30 it encompasses six domains: registration. visuospatial orientation, praxis, visuoconstructional drawing, judgment, memory recall and language. A very interesting feature of this instrument is that it was designed using the simultaneous development methodology and utilized culture and health advisory groups to select culture-fair items. Moreover, the test has been translated into different languages (Arabic, Danish, Iranian, Malayalam, Thai) and applied in several studies involving distinct multicultural samples resulting in a pooled sensitivity of 77.2% and a pooled specificity of 85.9% (Naqvi et al. 2015). All of its reliability indices (test–retest, inter-rater, and internal consistency) are very high (Naqvi et al. 2015). Notably, there is little effect of language, immigration status, and education on the RUDAS score (Naqvi et al. 2015). The RUDAS gathers all of the desirable features of a cross-cultural test: brevity, cultural fairness, good

psychometric properties, and little effect of education levels. Moreover, some data indicate that it is associated with the functional performance of clients with suspected dementia, which suggests evidence of ecological validity (Joliffe et al. 2015).

Unfortunately, most of these tests, despite some good initial results, have been little used. There is a paucity of publications including these tests, and efforts to publish findings in the cross-cultural neuropsychological field seem to have been discontinued for most of them.

In summary, there is evidence supporting the notion of the influence of cultural variables in neuropsychological testing. This influence can be observed in many aspects such as the definition of constructs, the construct validity of the instruments, the materials and response formats used in testing, as well as the items in the tests.

Various strategies have been proposed to avoid the biases caused by the influence of cultural variables. These remedies involve a hierarchy of solutions such as simultaneous development of new neuropsychological tests, adaptation and normalization of existing neuropsychological tests, as well as the development of a behavioral neuropsychology that avoids the use of tests altogether. The selection of the strategies to be used should be carefully evaluated in every situation in order to identify and justify the most appropriate approach for the occasion. In some cases, there will be only one option, whereas in others more choices may co-exist. In this paper, the authors advocate the simultaneous development of a new generation of neuropsychological tests. In order to achieve this goal, a series of recommendations are offered. There is enough theoretical and empirical evidence supporting the development of these tests, and although a great deal of effort would be necessary to develop such tests, the envisioned results would utterly justify this effort.

References

Alegret, M., Espinosa, A., Valero, S., Vinyes-Junqué, S., Ruiz, A., Hernández, I., et al. (2013). Cut-off scores of a brief neuropsychological battery (NBACE) for Spanish individual adults older than 44 years old. *PLoS ONE, 8*(10), 1–8.

Allegri, R. F., Mangone, C. A., Fernández Villavicencio, A., Rymberg, S., Taragano, F. E., & Baumann, D. (1997). Spanish Boston naming test norms. *The Clinical Neuropsychologist, 11,* 416–420.

Amin, K., Dill, R., & Thomas, S. M. (2003). *The CANA Test administration and scoring manual.* Phoenix, AZ: Scandinavian Graphics LLC.

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

Ardila, A. (2005). Cultural values underlying psychometric cognitive testing. *Neuropsychology Review, 15*(4), 185–195.

Ardila, A. (2007a). The impact of culture on neuropsychological test performance. In B. P. Uzzell, M. Pontón, & A. Ardila (Eds.), *International handbook of cross-cultural neuropsychology* (pp. 23–44). Mahwah: Lawrence Erlbaum Associates.

Ardila, A. (2007b). Toward the development of a cross-linguistic naming test. *Archives of Clinical Neuropsychology, 22,* 297–307.

Ardila, A., & Keating, K. (2007). Cognitive abilities in different cultural contexts. In B. P. Uzzell, M. Pontón, & A. Ardila (Eds.), *International handbook of cross-cultural neuropsychology* (pp. 109–125). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Ardila, A., & Moreno, S. (2001). Neuropsychological test performance in Arauco Indians: An exploratory study. *Journal of the International Neuropsychological Society, 7,* 510–515.

Ardila, A., Ostrosky-Solis, F., & Bernal, B. (2006). Cognitive testing toward the future: The example of semantic verbal fluency (ANIMALS). *International Journal of Psychology, 41*(5), 324–332.

Ardila, A., & Rosselli, M. (2007). Illiterates and cognition: The impact of education. In B. P. Uzzell, M. Pontón, & A. Ardila (Eds.), *International handbook of cross-cultural neuropsychology* (pp. 181–198). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Axelrod, B. N., Aharon-Peretz, J., Tomer, R., & Fisher, T. (2000). Creating interpretation guidelines for the Hebrew trail making test. *Applied Neuropsychology*, *7*(3), 186–188.

Azizian, A., Yeghiyan, M., Ishkhanyan, B., Manukyan, Y., & Khandanyan, L. (2011). Clinical validity of the Repeatable Battery for the Assessment of Neuropsychological Status among patients with schizophrenia in the Republic of Armenia. *Archives of Clinical Neuropsychology, 26*(2), 89–97.

Baddeley, A. (1999). *Human memory. Theory and practice* (Revised ed.). Hove: Psychology Press.

Beglinger, L. J., Gaydos, B., Tangphao-Daniels, O., Duff, K., Kareken, D. A., Crawford, J. F., et al. (2005). Practice effects and the use of alternate forms in serial neuropsychological testing. *Archives of Clinical Neuropsychology, 20,* 517–529.

Berry, J., Poortinga, Y., Breugelmans, S., Chasiotis, A., & Sam, D. (2011). *Cross-cultural psychology. Research and applications* (3rd ed.). Cambridge, UK: Cambridge University Press.

Böhm, P., Peña-Casanova, J., Gramunt, N., Manero, R., Terrón, C., & Quiñones-Ubeda, S. (2005). Spanish version of the Memory Impairment Screen (MIS): Normative data and discriminant validity. *Neurología, 20*(8), 402–411.

Bond, R., & Smith, P. B. (1996). Culture and conformity: A meta-analysis of studies using Asch's (1952b, 1956) line judgment task. *Psychological Bulletin, 119*(1), 111–137.

Brown, D. S. O., & Nell, V. (1991). The epidemiology of traumatic brain injury in Johannesburg: I. Methodological issues in a developing country context. *Social Science and Medicine, 33,* 283–287.

Caetano, C. (2007). Qualitative assessment within and across cultures. In B. P. Uzzell, M. Pontón, & A. Ardila (Eds.), *International handbook of cross-cultural neuropsychology* (pp. 93–108). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Chiao, J. Y., & Cheon, B. K. (2010). The weirdest brains in the world. *Behavioral and Brain Sciences, 33,* 61–135.

Chiao, J. Y., Iidaka, T., Gordon, H. L., Nogawa, J., Bar, M., Aminoff, E., et al. (2008). Cultural specificity in amygdala response to fear faces. *Journal of Cognitive Neuroscience, 20,* 2167–2174.

Christensen, A. L. (1975). *Luria's neuropsychological investigation: Manual and test materials* (1st ed.). New York: Spectrum.

Christensen, A. L., & Caetano, C. (1999). Luriás neuropsychological evaluation in the Nordic countries. *Neuropsychology Review, 9*(2), 71–78.

Chua, H. F., Boland, J. E., & Nisbett, R. E. (2005). Cultural variation in eye movements during scene perception. *Proceedings of the National Academy of Sciences USA, 102,* 12629–12633.

Cleophas, T., & Zwinderman, A. (2009). Meta-analyses of diagnostic studies. *Clinical Chemistry and Laboratory Medicine, 47*(11), 1351–1354.

Cole, M., & Packer, M. (2011). Culture and cognition. In K. D. Keith (Ed.), *Crosscultural psychology: Contemporary themes and perspectives* (pp. 133–159). Malden, MA: Wiley-Blackwell.

Cole, M., Gay, J., Glick, J., & Sharp, D. (1971). *The cultural context of learning and thinking*. New York: Basic Books.

Collie, A., Maruff, P., Darby, D. G., & McStephen, M. (2003). The effects of practice on the cognitive test performance of neurologically normal individuals assessed at brief test-retestintervals. *Journal of the International Neuropsychological Society, 9,* 419–428.

Cruice, M. N., Worrall, L. E., & Hickson, L. M. H. (2000). Boston Naming Test results for healthy older Australians: A longitudinal and cross sectional study. *Aphasiology, 14,* 143–155.

Cullen, B., O'Neill, B., Evans, J. J., Coen, R. F., & Lawlor, B. A. (2007). A review of screening tests for cognitive impairment. *Journal of Neurology, Neurosurgery and Psychiatry, 78,* 790–799.

Cullum, C. M., & Lacritz, L. H. (2009). Neuropsychological assessment in dementia. In M. F. Weiner & A. M. Lipton (Eds.), *Textbook of Alzheimer disease and other dementias* (pp. 85–103). Washington, DC: American Psychiatric Publishing.

D'Andrade, R. G. (1995). *The development of cognitive anthropology*. Cambridge, UK: Cambridge University Press.

Damasceno, A., Delicio, A. M., Mazo, D. F. C., Zullo, J. F. F., Scherer, P., Ng, R. T. Y., et al. (2005). Validation of the Brazilian Version of mini-test CASI-S. *Arquivos de Neuropsychiatria, 63,* 416–421.

De la Torre, G. G., Suárez-Llorens, A., Caballero, F. J., Ramallo, M. A., Randolph, C., Lleó, A., et al. (2014). Norms and reliability for the Spanish version of the Repeatable Battery for the Assessment of Neuropsychological Status (RBANS) Form A. *Journal of Clinical and Experimental Neuropsychology, 36*(10), 1023–1030.

Delis, D. C., Kramer, J. H., Kaplan, E., & Ober, B. A. (2000). *California verbal learning test—second edition, adult version*. San Antonio, TX: The Psychological Corporation.

Dellatolas, G., Deloche, G., Basso, A., & Claros-Salinas, D. (2001). Assessment of calculation and number processing using the EC301 battery: Cross-cultural normative data and application to left- and right-brain damaged patients. *Journal of the International Neuropsychological Society, 7*(7), 840–859.

Dick, M., Teng, E. L., Kempler, D., Davis, D., & Taussig, I. M. (2002). The cross-cultural neuropsychological test battery (CCNB): Effects of age, education, ethnicity, and cognitive status on performance. In F. R. Ferraro (Ed.), *Minority and cross-cultural aspects of neuropsychological assessment* (pp. 17–41). Lisse: Swets & Zeitlinger.

Dingwall, K. M., Lewis, M. S., Maruff, P., & Cairney, S. (2009). Reliability of repeated cognitive testing in healthy Indigenous Australian adolescents. *Australian Psychologist, 44*(4), 224–234.

Dorans, N. J., & Holland, P. W. (1993). In P. W. Holland & H. Wainer *Differential item functioning* (pp. 31–66). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Duff, K., Callister, C., Dennett, K., & Tometich, D. (2012). Practice effects: A unique cognitive variable. *The Clinical Neuropsychologist, 26*(7), 1117–1127.

Falleti, M. G., Maruff, P., Collie, A., & Darby, D. G. (2006). Practice effects associated with the repeated assessment of cognitive function using the CogState battery at 10-minute, one week and one month test-retest intervals. *Journal of Clinical and Experimental Neuropsychology, 28,* 1095–1112.

Fernandez, A. L. (2013). Development of a confrontation naming test for Spanish-speakers: The Cordoba naming test. *The Clinical Neuropsychologist, 27*(7), 1179–1198.

Fernandez, A. L., & Bendersky, V. (2004). Valores normativos para el Test de Stroop en una muestra de hispano parlantes [Normative data for the Stroop Test from a sample of Spanish-speakers]. *Psicodiagnosticar, 13,* 63–72.

Fernandez, A. L., & Fulbright, R. L. (2015). Construct and concurrent validity of the Spanish adaptation of the Boston naming test. *Applied Neuropsychology: Adult, 22*(5), 355–362.

Fernandez, A. L., & Marcopulos, B. (2008). A comparison of normative data for the trail making test from several countries: Equivalence of norms and considerations for interpretation. *Scandinavian Journal of Psychology, 49,* 239–246.

Fernandez, A. L., & Scheffel, D. L. (2003). A study on the criterion validity of the Mattis Dementia Rating Scale. *International Journal of Testing*, *3*(1), 49–58.

Fernandez, A. L., Marino, J. C., & Alderete, A. M. (2002). Estandarización y validez conceptual del Test del Trazo (trail making test) en una muestra de adultos argentinos (Normative data and construct validity of the Trail Making Test in a sample of Argentinian adults). *Revista Neurológica Argentina, 27,* 83–88.

Fernandez, A. L., Marino, J. C., & Alderete, A. M. (2004). Valores normativos en la prueba de Fluidez Verbal-Animales sobre una muestra de 251 adultos argentinos [Normative data in the Verbal Fluency-Animals test from a sample of 251 Argentinian adults]. *Revista Argentina de Neuropsicología, 4,* 12–22.

Fischer, R., & Fontaine, J. R. J. (2011). Methods for investigating structural equivalence. In D. Matsumoto & F. J. R. Van de Vijver (Eds.), *Cross-cultural research methods*. New York: Oxford University Press.

Gálvez-Lara, M., Moriana, J., Vilar-López, R., Fasfous, A., Hidalgo-Ruzzante, N., & Pérez-García, M. (2015). Validation of the cross-linguistic naming test: A naming test for different cultures? A preliminary study in the Spanish population. *Journal of Clinical and Experimental Neuropsychology, 37*(1), 102–112.

Garcia, C., Leahy, B., Corradi, K., & Forchetti, C. (2008). Component structure of the Repeatable Battery for the Assessment of Neuropsychological Status in dementia. *Archives of Clinical Neuropsycholy, 23,* 63–72.

Gardner, M. K. (2011). Theories of intelligence. In M. A. Bray & T. J. Kehle (Eds.) *The Oxford Handbook of School Psychology* (pp. 79–100). Oxford: Oxford University Press.

Gergen, K. J., Gulerce, A., Lock, A., & Misra, G. (1996). Psychological science in cultural context. *American Psychologist, 51,* 496–503.

Germine, L., Nakayama, K., Duchaine, B., Chabris, C., Chatterjee, G., & Wilmer, J. (2012). Is the web as good as the lab? Comparable performance from web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin & Review, 19*(5), 847–857.

Glosser, G., Wolfe, N., Albert, M. L., Lavine, L., Steele, J. C., Calne, D. B., et al. (1993). Cross-cultural cognitive examination: validation of a dementia screening instrument for neuroepidemiological research. *Journal of the American Geriatrics Society, 41*(9), 931–939.

Goh, J. O., Chee, M. W., Tan, J. C., Venkatraman, V., Hebrank, A., Leshikar, E. D., et al. (2007). Age and culture modulate object processing and object-scene binding in the ventral visual area. *Cognitive, Affective, & Behavioral Neuroscience, 7,* 44–52.

Goh, J. O., Leshikar, E. D., Sutton, B. P., Tan, J. C., Sim, S. K. Y., Hebrank, A. C., et al. (2010). Culture differences in neural processing of faces and houses in the ventral visual cortex. *SCAN, 5,* 227–235.

Goh, J. O., Siong, S. C., Park, D. C., Gutchess, A. H., Hebrank, A., & Chee, M. W. (2004). Cortical areas involved in object, background, and object-background processing revealed with functional magnetic resonance adaptation. *Journal of Neuroscience, 24,* 10223–10228.

Gontkovsky, S. T., Mold, J. W., & Beatty, W. W. (2002). Age and educational influences on RBANS index scores in a nondemented geriatric sample. *The Clinical Neuropsychologist, 16,* 258–263.

Gordon, P. (2004). Numerical cognition without words: Evidence from Amazonia. *Science, 306*(5695), 496–499.

Greve, K., Stickle, T., Love, J., Bianchini, K., & Stanford, M. (2005). Latent structure of the Wisconsin Card Sorting Test: A confirmatory factor analytic study. *Archives of Archives of clinical neuropsychology, 20,* 355–364.

Grigorenko, E. L., Geissler, P. W., Prince, R., Okatcha, F., Nokes, C., Kenny, D. A., et al. (2001). The organization of Luo conceptions of intelligence: A study of implicit theories in a Kenyan village. *International Journal of Behavioral Development, 25,* 367–378.

Gronwall, D. (1977). Paced auditory serial-addition task: A measure of recovery from concussion. *Perceptual and Motor Skills, 44,* 367–373.

Gutchess, A. H., Welsh, R. C., Boduroglu, A., & Park, D. C. (2006). Cultural differences in neural function associated with object processing. *Cognitive, Affective, & Behavioral Neuroscience, 6,* 102–109.

Hall, K. S., Gao, S., Emsley, C. L., Ogunniyi, A. O., Morgan, O., & Hendrie, H. C. (2000). Community screening interview for dementia (CSI 'D'); performance in five disparate study sites. *International Journal of Geriatric Psychiatry, 15*(6), 521–531.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice, 12*(3), 38–47.

Harry, A., & Crowe, S. F. (2014). Is the Boston naming test still fit for purpose? *The Clinical Neuropsychologist, 28*(3), 486–504.

Hasselblad, V., & Hedges, L. (1995). Meta-analysis of screening and diagnostic tests. *Psychological Bulletin, 17*(1), 167–178.

Hedden, T., Ketay, S., Aron, A., Markus, H. R., & Gabrieli, J. D. (2008). Cultural influences on neural substrates of attentional control. *Psychological Science, 19,* 12–17.

Henrich, J. (2008). A cultural species. In M. Brown (Ed.), *Explaining culture scientifically* (pp. 184–210). Seattle, WA: University of Washington Press.

Henrich, J., Hein, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33,* 61–135.

Hermans, H. J. M., & Kempen, H. J. G. (1998). Moving cultures: The perilous problems of cultural dichotomies in a globalizing society. *American Psychologist, 53*(10), 1111–1120.

Hogan, T. P. (2014). *Psychological testing: A practical introduction* (3rd ed.). Hoboken, NJ: Wiley.

Hogan, T. P., & Tsushima, W. T. (2016). Psychometrics and Testing. In J. C. Norcross, G. R. VandenBos, and D. K. Freedheim (Editors-in-Chief) *APA handbook of clinical psychology. Applications and Methods* (Vol. 3). American Psychological Association (APA).

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Holtz, J. L. (2011). *Applied clinical neuropsychology: An introduction*. New York: Springer.

Iyengar, S. S., & DeVoe, S. E. (2003). Rethinking the value of choice: Considering cultural mediators of intrinsic motivation. In V. Murphy-Berman & J. Berman (Eds.), *Cross-cultural differences in perspectives on the self*. Lincoln: University of Nebraska Press.

Iyengar, S. S., & Lepper, M. R. (1999). Rethinking the value of choice: A cultural perspective on intrinsic motivation. *Journal of Personality and Social Psychology, 76*(3), 349–366.

Johnson, B. L., Baker, E., EI Batawi, M., Gilioli, R., Hanninen, H., Seppalainen, A., et al. (1987). *Prevention of neurotoxic illness in working populations*. New York: Wiley.

Joliffe, L., Brown, T., & Fielding, L. (2015). Are clients' performances on the Rowland Universal Dementia Assessment Scale associated with their functional performance? A preliminary investigation. *British Journal of Occupational Therapy, 78*(1), 16–23.

Kabir, Z. N., & Herlitz, A. (2000). The Bangla adaptation of Mini-mental State Examination (BAMSE): An instrument to assess cognitive function in illiterate and literate individuals. *International Journal of Geriatric Psychiatry, 15*(5), 441–450.

Kay, P., & Regier, T. (2007). Color naming universals: The case of Berinmo. *Cognition, 102*(2), 289–298.

Kempler, D., Teng, E. L., Dick, M., & Taussig, I. (1998). The effects of age, education, and ethnicity on verbal fluency. *Journal of the International Neuropsychological Society, 4,* 531–538.

Kempler, D., Teng, E. L., Taussig, M., & Dick, M. B. (2010). The common objects memory test (COMT): a simple test with cross-cultural applicability. *Journal of the International Neuropsychological Society, 16*(3), 537–545.

Kim, J. K., & Kang, Y. (1999). Normative study of the Korean-California verbal learning test (K–CVLT). *The Clinical Neuropsychologist, 13,* 365–369.

Konstantinopoulou, E., Kosmidis, M., Ioannidis, P., Kiosseoglou, G., Karacostas, D., & Taskos, N. (2011). Adaptation of Addenbrooke's cognitive examination-revised for the Greek population. *European Journal of Neurology, 18*(3), 442–447.

Kwak, Y. T., Yang, Y., & Kim, G. W. (2010). Korean Addenbrooke's cognitive examination revised (K-ACER) for differential diagnosis of Alzheimer's disease and subcortical ischemic vascular dementia. *Geriatrics & Gerontology International, 10,* 295–301.

Lau, C. W., & Hoosain, R. (1999). Working memory and language difference in sound duration: a comparison of mental arithmetic in Chinese, Japanese, and English. Psychologia. *An International Journal of Psychology in the Orient, 42,* 139–144.

Lee, T., Cheung, C., Chan, J., & Chan, C. (2000). Trail making across languages. *Journal of Clinical and Experimental Neuropsychology, 22*(6), 772–778.

Legg, S., & Hutter, M. (2007). A Collection of Definitions of Intelligence. In *Proceedings of the 2007 conference on Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms: Proceedings of the AGI Workshop 2006*. Amsterdam: IOS Press Amsterdam.

Lehser, E., & Whelihan, W. (1986). Reliability of mental status instruments administered to nursing home residents. *Journal of Consulting and Clinical Psychology, 54*(5), 726–727.

Levav, M., Mirsky, A. F., French, L. M., & Bartko, J. J. (1998). Multinational neuropsychological testing: Performance of children and adults. *Journal of Clinical and Experimental Neuropsychology, 20,* 658–672.

Lezak, M., Howieson, D., Bigler, E., & Tranel, D. (2012). *Neuropsychological assessment* (5th ed.). New York: Oxford University Press.

Lim, M. L., Collinson, S. L., Feng, L., & Ng, T. P. (2010). Cross-cultural application of the Repeatable Battery for the Assessment of Neuropsychological Status (RBANS): performances of elderly Chinese Singaporeans. *The Clinical Neuropsychologist, 24*(5), 811–826.

Lonner, W. J. (2011). The continuing challenge of discovering psychological 'order' across cultures. In F. J. R. van de Vijver, A. Chasiotis, & S. M. Breugelmans (Eds.), *Fundamental questions of cross-cultural psychology* (pp. 64–94). Cambridge, UK: Cambridge University Press.

López, E., Steiner, A. J., Hardy, D. J., IsHak, W. W., & Anderson, W. B. (2016). Discrepancies between bilinguals' performance on the Spanish and English versions of the WAIS Digit Span task: Cross-cultural implications. *Applied Neuropsychology: Adult*, *19*, 1–10.

Mackintosh, N., & Bennett, E. S. (2005). What do Raven's Matrices measure? An analysis in terms of sex differences. *Intelligence, 33,* 663–674.

Maj, M., D'Elia, L., Satz, P., Janssen, R., Zaudig, M., Uchiyama, C., et al. (1993). Evaluation of two new neuropsychological tests designed to minimize cultural bias in the assessment of HIV-1 seropositive persons: A WHO study. *Archives of Clinical Neuropsychology, 8*(2), 123–135.

Majid, A., Bowerman, M., Kita, S., Haun, D. B. M., & Levinson, S. C. (2004). Can language restructure cognition? The case for space. *TRENDS in Cognitive Sciences, 8*(3), 108–114.

Manly, J. (2008). Critical issues in cultural neuropsychology: Profit from diversity. *Neuropsychology Review, 18*(3), 179–183.

Marchant, L. F., & McGrew, W. C. (1999). Human handedness: An ethological perspective. *Human Evolution, 13*(3–4), 221–228.

Marchant, L. F., McGrew, W. C., & Eibl-Eibesfeldt, I. (1995). Is human handedness universal? Ethological analysis from three traditional cultures. *Ethology, 101,* 239–258.

Marino, J. C., Fernandez, A. L., & Alderete, A. M. (2001). Valores normativos y validez conceptual del Test de Laberintos de Porteus en una muestra de adultos argentinos (Normative data and construct validity of the Porteus Mazes Test in a sample of Argentinian adults). *Revista Neurológica Argentina, 26,* 102–107.

Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review, 98*(2), 224–253.

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

Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care, 44,* S69–S77.

Messinis, L., Malegiannaki, A. C., Christodoulou, T., Panagiotopoulos, V., & Papathanasopoulos, P. (2011). Color trails test: Normative data and criterion validity for the Greek adult population. *Archives of Clinical Neuropsychology, 26*(4), 322–330.

Milfont, T. L., & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of Psychological Research, 3*(1), 111–121.

Mishra, R. C. (1997). Cognition and cognitive development. In J. W. Berry, P. R. Dasen, & T. S. Saraswathi (Eds.), *Handbook of cross-cultural psychology* (2nd ed., Vol. 2, pp. 143–176)., Basic Processes and Human Development Boston: Allyn & Bacon.

Miyamoto, Y., Nisbett, R. E., & Masuda, T. (2003). Culture and the physical environment: Holistic versus analytic perceptual affordances. *Psychological Science, 17*(2), 113–119.

Mungas, D. (2006). Neuropsychological assessment of Hispanics elders. In G. Yeo & D. Gallagher-Thompson (Eds.), *Ethnicity and the dementias* (2nd ed.). Florence, US: Routledge.

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. (2005a). 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. M. (2000). Development of psychometrically matched English and Spanish neuropsychological tests for older persons. *Neuropsychology, 14,* 209–223.

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

Mungas, D., Widaman, K. F., Reed, B. R., & Tomaszewski Farias, S. (2011). Measurement invariance of neuropsychological tests in diverse older persons. *Neuropsychology*, *25*(2), 260–269.

Murphy, K. (2003). The logic of validity generalization. In K. Murphy (Ed.), *Validity generalization. A critical review*. (pp. 1–30). Mahwah: Lawrence Erlbaum Associates.

Nampijja, M., Apule, B., Lule, S., Akurut, H., Muhangi, L., Elliott, A. M., et al. (2010). Adaptation of Western measures of cognition for assessing 5-year-old semi-urban Ugandan children. *The British Journal of Educational Psychology, 80,* 15–30.

Naqvi, R. M., Haider, S., Tomlinson, G., & Alibhai, S. (2015). Cognitive assessments in multicultural populations using the Rowland Universal Dementia Assessment Scale: A systematic review and meta-analysis. *Canadian Medical Association Journal, 187*(5), 169–176.

Nasreddine, Z., Phillips, N., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., et al. (2005). The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment. *Journal of the American Geriatric Society, 53,* 695–699.

Nell, V. (2000). *Cross-cultural neuropsychological assessment: Theory and practice*. Mahwah, NJ: Lawrence Erlbaum Associates.

Nell, V., & Brown, D. S. O. (1991). The epidemiology of traumatic brain injury in Johannesburg: II. Morbidity, mortality and etiology. *Social Science and Medicine, 33,* 289–296.

Nelson, N. W., & Pontón, M. O. (2007) The art of clinical neuropsychology. In B. P. Uzzell, M. Pontón, & A. Ardila, (Eds.), *International handbook of cross-cultural neuropsychology* (pp. 45–62). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Norenzayan, A., Choi, I., & Peng, K. (2007). Perception and cognition. In S. Kitayama & D. Cohen (Eds.), *Handbook of cultural psychology* (pp. 569–594). New York: Guilford Press.

Ostrosky-Solís, F., Ramírez, M., Lozano, A., Picasso, H., & Vélez, A. (2004). Culture or education? Neuropsychological test performance of a Maya indigenous population. *International Journal of Psychology, 39*(1), 36–46.

Oyserman, D., Sorensen, N., Reber, R., & Chen, S. X. (2009). Connecting andseparating mind-sets: Culture as situated cognition. *Journal of Personalityand Social Psychology, 97,* 217–235.

Paolo, A. M., Axelrod, B. N., & Tröster, A. I. (1996). Test-retest stability of the Wisconsin card sorting test. *Assessment, 3*(2), 137–143.

Patricacou, A., Psallida, E., Pring, T., & Dipper, L. (2007). The Boston naming test in Greek: Normative data and the effects of age and education on naming. *Aphasiology, 21,* 1157–1170.

Randolph, C. (1998). *Repeatable battery for the assessment of neuropsychological status, manual*. San Antonio, TX: Psychological Corporation.

Rezende, G. P., Cecato, J., & Martinelli, J. E. (2013). Cognitive abilities screening instrument-short form, mini-mental state examination and functional activities questionnaire in the illiterate elderly. *Dementia & Neuropsychologia, 7*(4), 410–415.

Rizzo, S., Venneri, S., & Papagno, S. (2002). Famous face recognition and naming test: a normative study. *Neurological Sciences, 23*(4), 153–159.

Roberson, D., Davies, I., & Davidoff, J. (2000). Color categories are not universal: Replications and new evidence from a stone-age culture. *Journal of Experimental Psychology: General, 129*(3), 369–398.

Ruffieux, N., Njamnshi, A., Mayer, E., Sztajzel, R., Eta, S., Doh, R., et al. (2010). Neuropsychology in Cameroon: First normative data for cognitive tests among school-aged children. *Child Neuropsychology, 16*(1), 1–19.

Sacktor, N. C., Wong, M., Nakasujja, N., Skolasky, R. L., Selnes, O. A., Musisi, S., et al. (2005). The international HIV dementia scale: A new rapid screening test for HIV dementia. *AIDS, 19,* 1367–1374.

Sadek, J., & van Gorp, W. (2010). The prediction of vocational functioning from neuropsychological performance. In T. Marcotte & I. Grant (Eds.), *Neuropsychology of everyday functioning* (pp. 113–135). New York: The Guilford Press.

Saklofske, D. H., van de Vijver, F. J. R., Oakland, T., Mpofu, E., & Suzuki, L. A. (2015). Intelligence and culture: History and assessment. In S. Goldstein et al. (Eds.), *Handbook of intelligence: Evolutionary theory, historical perspective, and current concepts* (pp. 1–22). New York: Springer.

Segall, M. H., Campbell, D. T., & Herskovitz, M. J. (1966). *The influence of culture on visual perception*. Indianapolis, IN: Bobbs-Merrill.

Shepherd, I., & Leathem, J. (1999). Factors affecting performance in cross-cultural neuropsychology: From a New Zealand bicultural perspective. *Journal of the International Neuropsychological Society, 5*(1), 83–84.

Shuttleworth-Edwards, A. B., & van der Merwe, A. S. (2016). WAIS-III and WAIS-IV South African cross-cultural normative data stratified for quality of education. In F. R. Ferraro (Ed.), *Minority and cross-cultural aspects of neuropsychological assessment* (2nd ed., pp. 72–96). New York: Taylor & Francis.

Shweder, R. A., & Bourne, E. J. (1984). Does the concept of the person vary crossculturally? In R. A Shweder & R. A. LeVine (Eds.), *Culture Theory* (pp. 158–199). Cambridge: Cambridge University Press.

Society for Industrial and Organizational Psychology. (2003). *Principles for the validation and use of personnel selection procedures* (4th ed.). Ohio: Bowling Green.

Statistics New Zealand. (2013). Census QuickStats about Māori. Retrieved from http://www.stats.govt.nz/.

Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of intelligence*. Cambridge: Cambridge University Press.

Storey, J. E., Rowland, J. T., Conforti, D. A., & Dickson, H. G. (2004). The Rowland Universal Dementia Assessment Scale (RUDAS): A multicultural cognitive assessment scale. *International Psychogeriatrics, 16,* 13–31.

Tanzer, N. K. (1995). Cross-cultural bias in Likert-type inventories: Perfect matching structures and still biased? *European Journal of Psychological Assessment, 11*(3), 194–201.

Tanzer, N. (2005). Developing test for use in multiple languages and cultures: A plea for simultaneous development. In R. M. Hambleton (Ed.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 235–263). Mahwah, NJ: Erlbaum.

Teng, E. L., Hasegawa, K., Homma, A., Imai, Y., Larson, E., Graves, A., et al. (1994). The cognitive abilities screening instrument (CASI): A practical test for cross-cultural epidemiological studies of dementia. *International Psychogeriatrics, 6,* 45–58.

Tombaugh, T., & McIntyre, N. (1992). The mini-mental state examination: A comprehensive review. *Journal of the American Geriatrics Society, 40*(9), 922–935.

Tsai, R. C., Lin, K. N., Wang, H. J., & Liu, H. C. (2007). Evaluating the uses of the total score and the domain scores in the cognitive abilities screening instrument, Chinese Version (CASI C-2.0): Results of confirmatory factor analysis. *International Psychogeriatrics, 19*(6), 1051–1063.

Uzzell, B. P., Pontón, M., & Ardila, A. (Eds.). (2007). *International handbook of cross-cultural psychology*. Mahwah, NJ: Lawrence Erlbaum Associates.

van de Vijver, F. J. R. (1997). Meta-analysis of cross-culturalcomparisons of cognitive test performance. *Journal of Cross-Cultural Psychology, 28,* 678–709.

van de Vijver, F. J. R., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology, 54,* 119–135.

Whaley, A. L., & Davis, K. E. (2007). Cultural competence and evidence-based practice in mental health services: A complementary perspective. *American Psychologist, 62*(6), 563–574.

Witsken, D. E., D'Amato, R. C., & Hartlage, L. C. (2008). Understanding the past, present, and future of clinical neuropsychology. *Essentials of neuropsychological assessment: Treatment planning for rehabilitation* (pp. 3–29). New York: Springer.

Woodcock, W. R., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III tests of cognitive abilities*. Itasca, IL: Riverside.

Zuo, L., Dong, Y., Zhu, R., Jin, Z., Li, Z., Wang, Y., et al. (2016). Screening for cognitive impairment with the Montreal Cognitive Assessment in Chinese patients with acute mild stroke and transient ischaemic attack: A validation study. *British Medical Journal Open, 6,* e011310. doi:10.1136/bmjopen-2016-011310.