NEUROPSYCHOLOGICAL APPLICATION OF THE INTERNATIONAL TEST COMMISSION'S (ITC) GUIDELINES FOR TRANSLATING AND ADAPTING TESTS

International Neuropsychological Society, Cultural Neuropsychology Special Interest Group,
Assessment Workgroup

Tedd Judd, Juliet Colon, Aparna Dutt, Jonathan Evans, Alexandra Hammond, Mark Hendriks, Tumie Kgolo, Maria Marquine, Lingani Mbakile-Mahlanza, Rune Nielsen, Christopher Nguyen, Shathani Rampa, Yesenia Serrano, Mathew Staios, Busisiwe Zapparoli, & Emily Zhou

2023

The content described in this publication is solely the responsibility of the authors and does not necessarily represent the official views of the International Neuropsychological Society or all Cultural Neuropsychology Special Interest Group members.

Contact information: For information about this work, please contact Tedd Judd at teddjudd@gmail.com.

ACKNOWLEDGMENT OF APPRECIATION

The Cultural Neuropsychology Special Interest Group (CN-SIG), Assessment Workgroup of the International Neuropsychological Society (INS) would like to express our appreciation for the inspiration and support provided by the many neuropsychologists and allied professionals working to translate, adapt, and develop neuropsychological tests, assessments, and other tools to serve unreached populations; the CN-SIG as a whole; INS administration, and especially the ITC for their pioneering work in developing and updating the ITC Guidelines for Translating and Adapting Tests, and for providing feedback on the current document.

We gratefully acknowledge the thoughtful comments of Elaine Ballard, Jamie Berry, Kate Cockcroft, Andy Dawes, Sanne Franzen, Kristin Jacklin, Michel Paradis, Porrselvi A.P., Parisuth Sumranasub, and Sharon Truter, that have been helpful to our process. All errors remain our own.

SUMMARY

ITC Summary:

The second edition of the ITC Guidelines for Translating and Adapting Tests was prepared between 2005 and 2015 to improve upon the first edition, and to respond to advances in testing technology and practices. The 18 guidelines are organised into six categories to facilitate their use: Pre-condition (3), test development (5), confirmation (4), administration (2), scoring and interpretation (2), and documentation (2). For each guideline, an explanation is provided along with suggestions for practice. A checklist is provided to improve the implementation of the guidelines.

CN-SIG Summary:

In this application work, the workgroup has specified how the ITC Guidelines can be applied to neuropsychological measures, where tests may be based not only on the semantic content of the items but upon a wide array of cognitive, perceptual, and praxis skills, novel tasks, and observations. As we strive to reach populations that are ever more diverse in their languages, cultures, educational experiences, behavioural norms, and expectations of the medical and testing encounter, it becomes increasingly important to consider the full social, linguistic, and cultural context of testing. These applications are aimed at guiding neuropsychological test translation, adaptation, and validation of existing tests for other uses, languages, cultures, etc., for consideration in the development of new neuropsychological tests, and other means of assessment where applicable.

Orientation to the organization of this work: To facilitate user-friendliness and access to all applicable material, we have included the original ITC Guidelines content in a document box as the one above. Our application will follow each section.

CONTENTS

ITC BACKGROUND

ITC:

The field of test translation and adaptation methodology has advanced rapidly in the past 25 years or so, with the publication of several books and many new research studies and examples of outstanding test adaptation work (see, for example, van de Vijver & Leung, 1997, 2000; Hambleton, Merenda, & Spielberger, 2005; Grégoire & Hambleton, 2009; Rios & Sireci, 2014). These advances have been necessary because of the growing interest in (1) cross-cultural psychology, (2) large-scale international comparative studies of educational achievement (for example, TIMSS and OECD/PISA), (3) credentialing exams being used world-wide (for example, in the information technology field by companies such as Microsoft and Cisco), and (4) fairness in testing considerations by permitting candidates to choose the language in which assessments are administered to them (for example, university admissions in Israel with candidates being able to take many of their tests in one of six languages).

Technical advances have been made in the areas of qualitative and quantitative approaches for the assessment of construct, method, and item bias in adapted tests and questionnaires, including the uses of complex statistical procedures such as item response theory, structural equation modelling, and generalizability theory (see Hambleton et al., 2005; Byrne, 2008). New translation designs have been advanced by OECD/PISA (see, Grisay, 2003); steps have been offered for completing test adaptation projects (see, for example, Hambleton & Patsula, 1999; exemplary projects are available to guide test adaptation practices - e.g., OECD/PISA and TIMSS projects); and many more advances have been made.

The first edition of the Guidelines (see van de Vijver & Hambleton, 1996; Hambleton, 2005) started from a comparative perspective, which is the purpose of the test adaptation to permit or facilitate comparisons across groups of respondents. The implicit template for which the guidelines were intended used a successive instrument development in a comparative context (the existing instrument has to be adapted for use in a new cultural context). It is becoming increasingly clear, however, that test adaptations have a wider domain of applications. The most important example is the use of a new or existing instrument in a multicultural group, such as clients in counselling who come from different ethnic groups, educational assessment in ethnically diverse groups with a differential mastery of the testing language, and internationally oriented recruitment for management functions in multinational companies. This change in domain of applicability has implications for development, administration, validation, and documentation. For example, possible consequences could be that items of an existing test should be adapted in order to increase its comprehensibility for non-native speakers (e.g., by simplifying the language). Another important extension of the guidelines would be to accommodate simultaneous development (i.e., the combined development of source and target language questionnaires). Large-scale international projects increasingly use simultaneous development in order to avoid the problem that the version developed in one language cannot be translated/adapted to all the languages of the study.

(continued)

ITC (continued):

The first edition of the ITC Guidelines for Translating and Adapting Tests was published by van de Vijver and Hambleton (1996), and by Hambleton (2002), and Hambleton, Merenda and Spielberger (2005). Only minor editorial changes were seen in the publication of the guidelines between 1996 and 2005. In the meantime, many advances have taken place since 1996. First, there have been a number of useful reviews of the ITC Guidelines. These include papers by Jeanrie and Bertrand (1999), Tanzer and Sim (1999), and Hambleton (2002). All the authors highlighted the value of the guidelines but then they offered a series of suggestions for improving them. Hambleton, Merenda, and Spielberger (2005) published the main proceedings of an ITC international conference held in 1999 at Georgetown University in the USA. Several of the chapter authors advanced new paradigms for test adaptations and offered new methodology including Cook and Schmitt-Cascallar (2005), and Sireci (2005). In 2006, the ITC held an international conference in Brussels, Belgium, to focus on the ITC Guidelines for Translating and Adapting Tests. More than 400 persons from over 40 countries focused on the topic of test adaptation and many new methodological ideas were advanced, new guidelines were suggested, and examples of successful implementations were shared. Papers presented in symposia at international meetings from 1996 to 2009 were plentiful (see, for example, Grégoire & Hambleton, 2009) and see Muniz, Elosua, and Hambleton (2013) for an early version of the second edition of the ITC Guidelines in Spanish.

In 2007, the ITC Council formed a six-person committee and assigned them the task of updating the ITC Guidelines to emphasise the new knowledge that was being advanced and the many experiences that were being gained by researchers in the field. These advances include (1) the development of structural equation modelling for identifying factorial equivalence of a test across language groups, (2) expanded approaches for identifying differential item functioning with polytomous response rating scales across language groups, and (3) new adaptation designs being pioneered by international assessment projects such as OECD/PISA and TIMSS. The committee, too, provided presentations and drafts of the new guidelines at international meetings of psychologists in Prague (in 2008) and Oslo (in 2009) and received substantial feedback on them.

The Administration Guidelines section was retained in the second edition, but overlapping guidelines were combined and the total number was reduced from six to two. "Documentation / score interpretations" was the final section in the first edition. In the second edition, we split this into two separate sections - one focused on score scales and interpretations, and the other focused on documentation. In addition, two of the four original guidelines in this section were substantially revised.

(continued)

ITC (continued):

As in the first edition, we want readers to be clear on our distinction between test translation and test adaptation. Test translation is probably the more common term, but adaptation is the broader term and refers to moving the items of a test from one language and culture to another. Test adaptation refers to all of the activities including: deciding whether or not a test in a second language and culture could measure the same construct in the first language; selecting translators; choosing a design for evaluating the work of test translators (e.g., forward and backward translations); choosing any necessary accommodations; modifying the test format; conducting the translation; checking the equivalence of the test in the second language and culture; and conducting other necessary validity studies. Test translation, on the other hand, has a more limited meaning restricted to the actual choosing of language to move the test from one language and culture to another to preserve the linguistic meaning. Test translation is only a part of the adaptation process, but can be, on its own, a very simplistic approach to transporting a test from one language to another with no regard for educational or psychological equivalence.

NEUROPSYCHOLOGICAL APPLICATION INTRODUCTION

Almost from the beginning of the history of psychological testing, tests have been translated, initially mostly among European languages. Over time, psychological tests have been increasingly used in cultural and linguistic contexts that differ considerably from the original population in which the tests were developed. The need for culturally sensitive adaptation of tests, in addition to translation, has been increasingly recognised. Further, technologies have improved for test translation, adaptation, administration, scoring, renorming, revalidation, and end-user interface. In recent years, the International Test Commission (ITC) has taken on the task of consolidating and guiding these developments.

The ITC has member test commissions from most of the national psychological associations of Europe and North America, as well as universities, test publishers, and individual members from at least 63 countries. In 2005, they published their Guidelines for Translating and Adapting Tests (hereafter, The ITC Guidelines) and revised these in a Second Edition in 2017. These Guidelines offer a tremendous service to test developers and users in many fields of psychology, education, employment, and other settings.

Sensitive and accurate test translation and adaptation is not only an issue of scientific accuracy and commercial product quality but also related to the basic human right to health care (United Nations General Assembly, 1948) and an issue of health equity and social justice. Unfortunately, neuropsychology has not always taken advantage of the insights and guidance offered by the ITC Guidelines. Because of its overarching scope addressing broad psychological testing, the ITC guidelines are mostly applicable to verbal, self-report tests. Neuropsychological tests often address abilities and issues that are particular to the linguistic, cultural, and educational context of the intended population. Test purposes, procedures, and item content often need to be cognitively similar rather than semantically similar. Items may need to be adapted by word length, linguistic features, familiarity, or other dimensions rather than literal meaning. In addition, much of the information obtained during neuropsychological assessment is obtained from process, approach to testing, and other aspects of the testing procedures as opposed to only score interpretation. Thus, detailed considerations of these particularities are needed to make the guidelines more practical and clearer for the neuropsychological test user. With the current work, we hope to make the ITC Guidelines readily applicable to neuropsychological measures, more consistently used by neuropsychologists, and more universally relevant to the field.

With these goals in mind, the Assessment Workgroup of the International Neuropsychological Society (INS) Cultural Neuropsychology Special Interest Group (SIG) has developed the current neuropsychological application of the ITC Guidelines. For the convenience of the reader, this document presents the complete Background Section and each of the 18 guidelines from the Second Edition ITC Guidelines followed by neuropsychological application, section by section. This application work has been developed by authors representing at least 10 different nations on all of the continents. We have drawn on literature concerning neuropsychological test development, translation, and adaptation; our own extensive experiences in these areas; and discussions with many colleagues also involved in related work. We have particularly attended to issues arising from the development, translation, and adaptation of languages that are far removed from European languages and for populations with limited or no formal education and limited or no experience with psychological testing. We hope that neuropsychological test developers, translators, and adapters will find these Guidelines and Neuropsychological Application helpful to their work and that editors and reviewers will find them useful in evaluating the adequacies and accuracy of such projects. As with the ITC Guidelines, this is an evolving field, and we anticipate that there will be a need for revisions and new editions and look forward to the refinement and improvement of this application.

NEUROPSYCHOLOGICAL APPLICATION BACKGROUND

The International Test Commission (ITC) Guidelines for Translating and Adapting Tests (Second Edition; 2017) offers a tremendous service to test developers and users in many fields of psychology, education, employment, and other settings. The current ITC neuropsychology application gives further details to make the guidelines more practical and clearer for the neuropsychological user.

The Purpose of the ITC Neuropsychological Application.

The world's linguistic, cultural, age, educational, and neurologic diversity is vast, and developing instruments that can serve that diversity adequately and equitably is a massive undertaking. This ITC neuropsychological application aims to facilitate that undertaking as it relates to neuropsychological tests. They are intended for professionals who develop, adapt, and use neuropsychological tests, in specific to those adapting existing tests to languages and cultures for which the test was not originally developed.

This ITC's neuropsychological application is aspirational and not mandatory; test developers need not comply with all of these recommendations, but we do recommend that they account for their decisions not to comply. We recommend that test developers and editors regard these guidelines and applications as the current consensus on the topic and recommend that these are consulted *prior* to undertaking projects aimed at adapting neuropsychological tests for new languages and cultures. We recommend that they be referred to in publications that involve neuropsychological test translation and/or adaptation.

***The Origins and Development of Neuropsychological Testing*.**

The discipline of neuropsychology and neuropsychological tests emerged from both behavioural neurology and the science and practice of psychology. In recent years, the rapid growth and dispersion of neuropsychology have led to the steady rise of the awareness and influence of distinct factors that contribute to performance on neuropsychological tests. Psychological tests that had been developed for other purposes were investigated for their ability to detect brain damage. This enterprise also led to the development of new tests with a primary neuropsychological design and purpose.

As the field developed, multiple uses of neuropsychological tests emerged. One direction went beyond the initial purpose of detecting brain damage (or "organicity") in general to distinguishing specific types of brain disorders, including localising lesions within the brain. This led to the development of tests that focus on specific cognitive functions and emotional and behavioural manifestations of brain disorders. Testing can be used for varied medical purposes, such as diagnostic evaluation, tracking the evolution of a condition, evaluating the impact of treatments, determining candidacy for treatments, guiding rehabilitation and neuropsychotherapy, etc. The neuropsychological approach has also come to be applied in many other domains, such as education, vocational rehabilitation, disability determination, criminal responsibility, determination of legal competencies, determination of work capacities (especially where public safety is concerned—pilots, drivers, doctors, elected officials), determine disability and access to services and compensations, determine levels of liability, etc. Each of these distinctive applications implies distinctive validities that go well beyond validation based on neurological diagnosis and lesion localization.

Considering how neuropsychological tests are used may have important implications for test development, translation, and adaptation. The purposes of neuropsychological assessments can be broadly categorised into two groups: 1) diagnosing and measuring brain disorders, and 2) making inferences regarding real-life functioning based on neurocognitive strengths and weaknesses.

The advent of neuroimaging techniques has reduced the historical role of neuropsychology in the detection of brain disorders; yet neuropsychological assessments continue to remain relevant for the diagnosis of specific brain disorders that cannot be readily determined with neuroimaging procedures. The predominant model for the detection of underlying brain dysfunction based on neurocognitive tests is the comparison of the test-taker's scores to the scores of a normative sample representative of the test-taker, including intra-individual comparisons across cognitive functions and time. Tracking a person's cognitive functions over time aids in the determination of the trajectory of cognitive change, including gradual worsening, which might be indicative of a progressive brain disorder, or improvement over time, which might be the case after acute insult to the brain. This information by itself, however, is insufficient for diagnostic purposes, and ought to be considered together with the person's history and results of other neurodiagnostic procedures. Additional methods to consider for the detection of brain disorders include pathognomonic signs, self-ratings versus ratings by others, and criterion-based testing.

A second important purpose of neuropsychological testing is the determination of cognitive strengths and weaknesses and their functional correlates in a person's everyday life. While in some cases brain damage might be clear based on history and/or other neurodiagnostic procedures, results of neuroimaging are unable to determine functional impairment or capacity. Neuropsychological tests measure abilities that are important in everyday functioning thus allowing for inferences in this regard. The accuracy with which neuropsychological tools evaluate cognitive strengths and weaknesses has been a key cornerstone of neuropsychological assessment (Lezak et al., 2012).

Clinicians, Researchers, and Other Test Users.

These Guidelines and application are directed to test developers, translators, and adapters, and to the editors and funders who review their work. We hope that this work can also be of use to clinicians, researchers, and other test users in selecting, using, and interpreting tests in diverse language and cultural contexts. Nevertheless, we recognize that busy clinicians often do not have the time or expertise to review all of these guidelines and the major research on a variety of instruments in different languages to select the most appropriate tests and norms for their particular needs. We do not have the space in this document to write explicit guidance for test users on test and norm selection and interpretation. We hope that these Guidelines and application will provide a framework that will allow for review articles and compendia for specific languages that can offer more concise guidance for the busy practitioner. Such reviews are already available, for example, for English (Lezak, et al, 2012; Strauss, Sherman, & Spreen, 2006), Arabic (Fasfous, et al, 2017), Mandarin adult measures (Qi, Sun, & Hong, 2022), and African populations (Shuttleworth-Edwards & Truter, 2023).

There is also a need for further guidance on how to proceed when there are no appropriate instruments available. As will be seen below, using tests that have not been validated for the population and purpose in question is not appropriate and is likely to be unethical in high-stakes medical and legal contexts. This is especially likely to be true for test-takers from languages and cultures remote from the origins of the test and with limited formal and academic education. Nevertheless, there are also clinical situations in which testing of such populations can contribute to care, such as when tests can be used to demonstrate strengths, functional capacities, and rehabilitation resources.

The neuropsychological application below is an attempt to address the special considerations that may be needed in translating and adapting neuropsychological tests and whether they can be validly translated and adapted at all. It considers the nature of the functions that neuropsychological tests measure, the broad range of applications of such tests, and the diverse ways that such abilities and behaviours are manifested across languages, cultures, ages, and types of education.

Cultural Diversity in Neuropsychological Testing.

Psychological and neuropsychological testing have been used for a myriad of purposes in their 100+-year history. Problems with intentional and unintentional racial, ethnic, and linguistic discrimination caused by the test instruments themselves as well as their uses have been well-documented and prominent in that history (Council of National Psychological Associations for the Advancement of Ethnic Minority Interests, 2016) and persist to this day. Some of the most prominent problems have been in their use for eugenics, immigration restriction, military placement, education access, and mental health treatment. For this reason, careful and accurate test translation and adaptation take on even greater importance. This process can only be fully and fairly carried out in a context that also revalidates tests for intended uses and considers the possibility that fair translation and adaptation may not be possible.

As the mission of neuropsychology has broadened, so has the diversity of the populations to which neuropsychological tests are applied. In this process, the field has increasingly encountered that the assumptions underlying the process of neuropsychological testing are not culturally universal. As Matthews (1992) so eloquently described it,

> “A very limited kind of neuropsychology, appropriate to only a fraction of the world’s population, is presented to the rest of the world as if there could be no other kind of neuropsychology, and as if the educational and cultural assumptions on which neuropsychology is based were obviously universals that applied everywhere in the world.”

When these assumptions and understandings vary across cultures, they present challenges to the smooth transfer of instruments from one culture and language to another. This growing awareness of profound cultural differences in cognitive processes has led to greater questioning as to whether valid test translation and adaptation is even possible for some cultures, cognitive functions, and testing uses.

- *Multilingualism in Neuropsychological Testing*. Neuropsychological testing has developed primarily in monolingual populations and/or with assumptions of monolingualism, even though it has been estimated that the majority of the world's population is multilingual. Multilingualism (i.e., having capacities in more than one language) has implications not only for language testing, verbal memory, and comprehension of test instructions but also for cognitive functions in many domains. Multilingualism may be associated with enhanced executive functioning, although these effects appear to be small (Lehtonen, et al., 2018), context-dependent (Blanco-Elorrieta, & Pylkkänen, 2018), and of unclear clinical significance.

  Multilingualism is not unidimensional; multilinguals vary greatly in their language competence in each language in the four major modalities of speaking, oral comprehension, reading, and writing. They may differ further in their vocabularies and competencies in different domains of discourse. For example, many people prefer and are better in their first language for thinking and for discussing feelings and social relationships but prefer and are better in the language they were schooled in for literacy, maths, and academic and work-related discourse. Thus, it may be difficult to identify the best "test language" because two or more languages may be preferentially used depending on the domain of discourse.

  One very common pattern across colonised and formerly colonised populations is that they speak a home language first but then learn a colonial language in school. They may not become literate in their home language. This is the pattern in most of Africa, India, much of the former Soviet Union outside of Russia, across the Arctic, the 40 million Indigenous peoples of Latin America, much of the Pacific Islands, and many parts of China. It is particularly true for Indigenous peoples whose home language may be linguistically quite distant from the colonial language (e.g., uses a different writing system), or whose home language is of a primarily oral culture. For example, there are an estimated 500 million native English speakers in the world, but about 1.5 billion English language learners (those who speak English as a 2nd or more language to varying degrees, Crystal, 2003). Accordingly, 3/4 of the potential users of English language tests would speak English as a second or more language. The teaching of colonial languages and literacies in a transitional or extractive model (rejecting the home language, Serpell, 1993) is viewed by some as linguistic and culture genocide and a crime against humanity (Skutnabb-Kangas & Dunbar, 2010) and so may be resisted in some sectors. The presumption of testing in the colonial language may similarly be viewed with suspicion.

  Test translators/adapters/developers need to consider the multilingualism of their intended test-takers carefully, as performance is also highly influenced by native language reading strategies and native language structure. Care should be taken to avoid unnecessary language burdens that are not critical to the purpose of the test, especially if the test is likely to be given in test-takers second or more languages. They should be familiar with the cognitive and linguistic characteristics of multilingualism and how it is evaluated. They should give explicit instructions regarding specific language competencies in all four modalities needed to take the test and give explicit instructions regarding how to evaluate multilingualism and/or take it into account in test administration and interpretation. In general, if a test is intended for use by multilinguals they should be included in piloting and norming. Depending upon the specifics of the test, its purposes, and intended populations, it may be advisable to look at monolinguals and multilingual separately. Particular care is needed for tasks involving mental mathematics, since most people perform these calculations in the language, they learned maths in, regardless of the language in which the task is presented. Tests that are specific to language or verbal memory also need to be developed with particular attention to the complex dimensions of multilingualism and the specific characteristics of each language (Paradis, 1987).

If neuropsychological assessment has neglected multilingualism, this neglect has been multiplied in regard to plurilingualism. Plurilingualism refers to the skills of the multilingual individual in managing their various language capacities, such as interpretation and translation, knowing which language to use in which circumstances, mixing the use of languages, and metalinguistic awareness (Coste, Moore, & Zarate, 2009). Luk (2022) argues that an understanding of plurilingualism is needed to bring order to the polarised and politicised battleground of bilingual advantages and disadvantages. Although plurilingualism is well-developed in linguistics and has taken its place as a fundamental goal of language education (Council of Europe, 2020), it is barely known in cognitive psychology (Shaharban, Rangaiah & Thirumeni, 2022) and is essentially unknown in neuropsychology. The companion concept of pluriculturalism is similarly underexplored. These concepts suggest that models of testing built on monolingual and monocultural reference populations, assumptions that "best language" can be confidently identified and used for testing, and additive models of multilingualism are overly simplistic and increasingly obsolete.

- *Illiteracy, Innumeracy, and Formal Education.* Formal schooling and learning literacy and maths contribute to the development of many different underlying cognitive abilities that are often measured by neuropsychological tests (Ardila, et al., 2010). The impact of years of schooling on neuropsychological test performance is not linear but shows the greatest effects for the first few years of schooling (Ostrosky, et al., 1998). It is not surprising that acquiring literacy improves phonemic awareness, phonemic verbal fluency, non-word repetition, and syntactic comprehension, and that this has implications for aphasia testing (Tsegaye, De Bleser, & Iribarren, 2011). The acquisition of literacy also improves verbal and visual memory, phonological awareness, executive functioning, and visuospatial and visuomotor skills (Ardila, et al., 2010), although some of these effects may be due to schooling more generally and not exclusively due to literacy. Many of these effects are not just a matter of degree or test norms, but rather, of acquired skills including reading, writing, maths, drawing, strategies for problem-solving, abstract thinking, etc. For instance, Nielsen and Jørgensen (2013) found that both quantitative and qualitative aspects of visuoconstructional drawing performances in illiterate Turkish immigrants in Denmark were impossible to discern from the performances commonly found in people with neurocognitive disorders.

  Literacy also shows variability in its manifestations according to the writing system and the individual's life circumstances. There is no single, fixed, and agreed test or definition of illiteracy. The operationalization of the construct of illiteracy needs to be selected according to the purposes of the test and the characteristics of the target population and writing system. For example, in ideographic writing systems such as Mandarin Chinese (Han Zi), its Japanese derivative, Kanji, and its Korean derivative, Hanja, one's reading vocabulary will extend only as far as the characters one has learned and is likely to be smaller than one's spoken vocabulary. So, literacy may be limited even with schooling. With transparent orthographies such as Spanish, Italian, Shona, and Turkish, learning to read can be relatively fast and many learn informally without schooling. With opaque orthographies such as English and French, learning literacy is more likely to be slow and difficult so some people are functionally illiterate even after several years of schooling.

Literacy skills and ease also vary by the diglossia of a language and writing system, that is, the difference between the written and spoken form of the language. Additionally, many educational systems use a language of instruction that is not the student's home language, and this typically delays the acquisition of literacy. Because of these factors, years of schooling have a very variable relationship with acquired literacy and also with the acquisition of cognitive skills related to literacy such as metalinguistic awareness and linguistic mediation of reasoning and problem-solving.

Diglossia impacts not only the learning and ease of literacy but also the pragmatics of testing. Test-takers with limited literacy skills and habits may have difficulty comprehending and relating to diglossic written test instructions and content even if it is read to them. On the other hand, diglossic written materials sometimes cross dialect boundaries for those who are sufficiently literate. For example, Arabic and Chinese each have many spoken dialects that may be mutually unintelligible, but their written forms may be comprehended reasonably well across some dialects (e.g., areas using the Mandarin written system, Hanzi, but speaking various Chinese dialects of Guan Hua (Mandarin) such as Sichuan Hua and Nanjing Hua).

Once learned, literacy skills tend to be retained, but if not used regularly they may become dysfluent and impractical. Many individuals also show large differences between their reading and their writing skills, a distinction that also varies by language and writing system. Multilingual people may be literate in some languages and illiterate in others.

All of these factors imply that the collection of cognitive skills that are developed through formal schooling and the acquisition of literacy are likely to be variable depending upon the nature of schooling; characteristics of the language and of the writing system; multilingualism; and the cognitive, literary, and life habits of the individual after schooling and the acquisition of aspects of literacy.

With these considerations in mind, it is particularly important that test translators/adapters/developers consider from the start whether they intend their tests to be used with illiterate populations. They need to plan, pilot, norm, and validate their instruments accordingly. In such circumstances, it is also important to pay attention to how illiteracy is determined. Test developers should take particular care to avoid lumping those who are potentially illiterate into the same normative group as those who are literate, for example, by assuming that years of education is an accurate marker for literacy or by lumping together into one normative group those with fewer than a certain number of years of education. These considerations go well beyond considering whether tests require the test-takers to read.

Many neuropsychological tests presume basic abilities in arithmetic to measure other functions such as attention, speed of processing, reasoning, executive functions, and performance validity. Those who are illiterate may also be innumerate, without precise arithmetic abilities beyond low single digits. However, even some of those who have learned functional arithmetic outside of formal schooling may not have learned it in the same systematic and multifunctional ways that come with formal schooling. For such individuals, irregular arithmetic skills may be mistaken for impairments in attention, reasoning, etc. Test translators/adapters/developers need to consider from the start whether they intend their tests to be used with such populations and adapt their methodologies and instructions accordingly.

People with low education or lack of formal education may use several elementary strategies to perform adequately in neurocognitive assessments. Using fingers to count in a serial subtraction test or repeating words after the examiner during word list learning trials are examples of such everyday strategies. A novice clinician may misinterpret them and may even withhold clients from applying such strategies, which might hinder their performance (Dutt, Evans, & Fernandez, 2022).

To avoid overestimation of cognitive impairment in the illiterate population and also in some low-educated groups, it may be preferable to develop functional assessment tools rather than tests for many cognitive domains.

Domains of Neuropsychological Testing.

The different domains measured by neuropsychological tests have distinctive methods of inference. Culture and language impact each domain and its measurement in different ways. For these reasons, the distinctiveness of test adaptation and translation needs to be described for each domain separately. Here are those descriptions commonly used by neuropsychologists (for a more detailed description of cognitive domains of function, please refer to Lezak et al, 2012):

- *Sensory/Perceptual*. While sensory functions are not expected to vary greatly as a function of culture or language, the understanding and responses to instructions may vary. The primary focus of test translation and adaptation is to preserve accurate measurement of sensory functions rather than preserving the semantic equivalence of instructions. This may require elaborated explanations of the purpose of the exam and expectations for some populations. Perceptual functions may have greater cultural influences than sensory functions because cultural learning is involved in the interpretation of sensations. The familiarity of stimuli influences how they are recognized and interpreted (for example, in the common perceptual tests of fingertip number writing perception, the naming of felt objects and shapes, perception of line drawings, perception of the visual representation of 3 dimensions in 2 dimensions, phonemic discrimination, and odour identification). Furthermore, although basic sensory/perceptual functions may not differ between cultures, performance on tests may be influenced by cultural differences in how/what people attend to stimuli in the world.

- *Motor/Praxis*: As with sensory functions, basic motor functions are not expected to vary greatly as a function of culture or language, but the understanding and responses to instructions may vary. Coordinated movement, speed of movement, absolute strength, and praxis are all likely to show variation by culture and various experiences such as sports training, musical training, types of employment, and general cultural expectations. Pathologies of movement such as tremors and tics may vary culturally in the degree to which they are voluntarily suppressed, the degree to which they are regarded as problematic, and the causes attributed to them.

- *Attention***:** Language-based measures of attention such as Digit Span, mental maths, and overlearned series are closely tied to phonemic characteristics of the words and learning histories. This domain involves the ability to attend to specific stimuli, over a prolonged period, without being deterred by other external cues. It also involves regulating and monitoring actions that lead to the execution of a plan, minimising errors, and achieving goals. Test translation and adaptation need to place priority on reproducing cognitively similar tasks relative to language and culture rather than semantically similar tasks. For example, capacity in Digit Span and mental maths is relative to the word lengths for the names of the numbers in the language in which it is carried out. Welsh digit names are longer than English digit names, so Welsh-dominant Welsh-English bilinguals can repeat more digits forward in English than in Welsh (Ellis, 1992). Visual attention is also shaped by culture and language, for example, the direction of scanning of the visual field is influenced by the direction of reading. Again, cognitive equivalence is the priority of test adaptation.

- *Information Processing Speed*: In this domain of cognition, sometimes called the speed of information processing, participants are expected to use multiple cognitive processes to reach a goal. Speed, working memory, and the ability to store, retrieve, and use information are paramount, with compromised functioning suggestive of compromised information processing and impaired brain functioning. Traditional neuropsychological tests in this domain usually rely on speed as a primary metric. However, cultures differ greatly in the emphasis and value that they place on speeded mental and physical tasks and work in education, employment, and other life domains. Speed-accuracy trade-off values also show cultural differences, undermining the implicit assumptions of traditional neuropsychological tests in this domain.

- *Memory:* As with verbal working memory, verbal short-term memory is dependent upon language characteristics. However, what is considered to be accurate memory may vary by culture. For example, most word list learning tasks count only exact word recall as accurate, while some cultures may consider synonym recall being accurate. Similar issues may arise with the recall of stories and drawings. Practice in memorising and studying and other memory strategies may also vary by culture, so word lists with implicit categories may be obvious in some cultures and not in others. Again, cognitive rather than semantic equivalence is the priority of test adaptation.

  For multilinguals where testing may be needed in more than one of their languages it is especially important NOT to use translated versions of the same verbal comprehension or memory materials (word lists, word pairs, stories) because there is likely to be substantial semantic memory carry-over from testing in one language to testing in another language. Rather, different languages should follow similar construction principles but with intentionally non-overlapping content (Paradis, 2011).

  Visual-spatial memory is often measured through drawing tasks. Drawing is a learned skill that varies by education, practice, and culture. Even visual recognition memory tasks are likely to vary by culture depending upon the familiarity not only of the specific stimuli, but also by the familiarity of the style or form of the stimuli, such as the differences in form between non-verbal stimuli that resemble the writing of English, Arabic, and Chinese.

- *Executive Functions*: Many tests of executive functions are designed to be sufficiently familiar as a category of activity that the test-taker can infer expectations but sufficiently unfamiliar so that they need to develop somewhat novel strategies to be successful. The target of an almost-familiar task is an elusive one across cultures. Such tests may resemble card games or games of chance or logic whose distribution and familiarity vary from culture to culture. Individuals who lack cognitive flexibility, for example, usually act on routine and find it difficult to adjust when changes in behaviour are called for. Impairment often leads to perseverative patterns recurring. Perseveration could, however, be because of other factors having to do with the understanding of instructions, rather than cognitive inflexibility. Translating and adapting executive function tests is particularly challenging. It calls for construct expertise in the target culture, including familiarity with types of tasks and puzzles and problem-solving strategies that may not be explicitly known or studied in a target culture and an understanding of the general level of familiarity with such activities.

  Executive functions are notoriously context dependent. Even if an executive function task has been well-translated and adapted, correlates well with other executive function tests, and shows diagnostic validity, this does not guarantee that it will accurately predict adaptive behaviours in the real world.

  Executive functions are also measured by behavioural rating scales. For acquired disorders, cultural differences can be mitigated somewhat by comparing their current presentation to their presentation before the onset of the disorder. Overall, there are wide cultural and subcultural differences in expectations for the exercise of executive functions. Behavioural rating scales need to be relevant to the target culture and may call for considerable adaptation of content to reflect not only culturally typical activities and behaviours, but also values.

- *Language:* Neuropsychologists are often called upon to evaluate acquired disorders of language (aphasia, alexia, agraphia, anomia) and disorders of language acquisition (specific learning disorders). Aphasia tests have many subtests designed to measure specific language functions such as aspects of reading, writing, articulation, etc. Languages differ intrinsically in many features, including phonology, vocabulary, semantics, syntax, and prosody, so the translation of the semantic content of a test may not capture the linguistic features being tested (Ivanova & Hallowell, 2013). A feature critical to one language (such as linguistic tone or phonetic analysis in reading) may be minimal or absent in another language. "A direct or literal translation of an existing test is never appropriate (Paradis, 1987) because there is not a one-to-one match between words and syntactic structures across any two languages (even similar languages with common origins, such as the Latin languages)" (Ivanova & Hallowell, 2013). The biased nature of some pictorial representations used in language tests is also of concern in neuropsychological assessment. Furthermore, recorded audio tasks in an unfamiliar accent can likely have an impact on performance on language-based tasks.

- *Visual-Spatial/Perceptual Skills:* A persisting false myth of neuropsychology is that non-verbal tasks are culture-fair. Recognizing line drawings and other conventions of graphic representation such as the representation of 3 dimensions with 2 dimensional drawings is a learned skill. Likewise, drawing, and constructing figures from sticks, blocks, materials/manipulables are visual-motor skills that are learned through practice. Mental rotation is a learned skill. Whether or not rotations of figures are considered the "same" figure varies with cultures and contexts. Tests that are intended to detect loss of visual-spatial/perceptual skills due to brain disorders should use materials and activities that are typically well-learned in the target population, as determined by field study and piloting.

- *Social Cognition:* Although there is arguably universality to certain emotional facial expressions, there is considerable cultural variation in the contexts in which such emotions are expressed. Aside from very basic emotions, there is much cultural subtlety and variation in how feelings, opinions, judgments, relationships, etc. are expressed. Such variations may be associated with different languages (e.g., Italian speakers express X judgment with Y tone of voice and Z facial expression while Vietnamese speakers express it with P tone of voice and Q facial expression). But such variations are not likely to be as closely tied to specific languages as are specific vocabularies, for example. Many actors and comedians derive much of their art from exploiting regional and national differences in emotional expressions within the same language. It is challenging to identify population clusters of forms of social cognition and communication to identify appropriate normative groups. Tests that concern perception and expression of emotion and other forms of social communication and cognition are still in relatively early stages of development. While the semantic content of such tests will carry varying degrees of weight when they are translated, the "translation" of the emotional and social messages also needs to be considered.

  Beyond the expression and perception of emotions and opinions, social cognition shows great cultural variation, again according to social groupings that may not follow the divisions of different languages. For example, behaviours that are considered polite, rude, aggressive, flirtatious, and friendly may vary widely among English speakers from Canada, Guyana, Scotland, India, Nigeria, and Australia. The appropriateness of eye contact is well-known to show considerable cultural variation. "Social lying" to make someone else feel comfortable is common in many cultures but varies regarding when it is appropriate. Social cognition is an exciting but challenging field of development in neuropsychological measurement. Test developers in this domain should take care to specify to the degree possible the social groups and populations for which the test is intended, recognizing that language and/or nationality are likely to be insufficient specifiers.

- *Behavioural Regulation During Testing*: Conventional testing is formal, in an isolated environment, and involves a relationship between the tester and testee, usually two individuals who have previously never been in contact with one another. This interaction has been studied as contributing to testing and performance score differences as this arrangement may be unconventional in some social-group-oriented societies.

- *Emotional Status:* The ITC Guidelines have been extensively applied to conventional tests and scales of emotions and personality, tests that are frequently heavily dependent upon the semantic content of the items. These Guidelines likewise apply to neuropsychological tests and scales of emotions and personality that are dependent upon semantic content. Similarly, however, although items may be translated and adapted to accurately reflect specific behaviours, emotions, and tendencies, they may not necessarily reflect the varied cultural values and belief systems attached to such descriptions. For example, behavioural changes due to dementia such as hallucinations and confabulations may have similar manifestations in different cultures but may be regarded as quite aberrant and problematic in one culture but acceptable in another. In another example, the hyperkinetic activity of Attention Deficit Hyperactive Disorder may be viewed and dealt with very differently in a Chinese public school and a Zambian Montessori school. Inventory items that may be acceptable in some cultures may be considered too personal, offensive, or objectionable in other cultures, for example, items concerning sexuality, religion, family relationships, drug and alcohol use, death, and mental illness. The nature of the validation of a measure of such behaviours may therefore depend upon whether the intended application is for diagnosis or for identifying targets of professional intervention.

- *General Adaptive Abilities:* Neuropsychological testing includes the measurement of adaptive abilities and disabilities, that is, the performance of functions in everyday life in domains such as self-care, medication management, employment, financial management, academic achievement, transportation, social relationships, and recreation. General scales of functioning aim to measure broadly across all domains of functioning to determine the overall impact of a neurological disability on life. There are many applications of such measurement such as rehabilitation goals, educational goals and strategies, guardianship, disability pensions, and legal and criminal responsibility. Such scales are dependent upon presumptions regarding what activities and abilities are typical and how they are valued. For example, many such scales use as their metric the degree of personal independence in performing the function, yet the degree to which personal independence is valued shows great cultural variation. For acquired brain disabilities, cultural differences can be mitigated, to some extent, for such scales by comparing pre-disability functioning to current functioning. Nevertheless, cultural variations in attitudes towards disability, rehabilitation, and life roles will also impact such comparisons. Full validations of such adapted scales will also need to take such cultural variations in typical activities and values into account.

- *Specific Adaptive Abilities:* Neuropsychological assessment also often addresses specific life skills through ratings and direct measurement of performance. Validation of such measurements needs to be relative to the domain of application. For example, financial capacities in dementia need to be evaluated relative to the evaluee's usual financial skills and habits (cash vs. check vs. credit card vs. online, etc.) and contextual vulnerabilities (scams, fraud, undue influence). The cognitive and English language capacities of commercial airline pilots, on the other hand, appropriately have a global standard. No single standard and strategy will serve equally well for all test translation and adaptation projects for specific adaptive abilities. The ways that culture will need to be considered will vary project by project concerning the goals of application and means of validation.

- *Premorbid Functioning.* Neuropsychology has developed several methods for estimating premorbid cognitive functioning to provide a basis for refining determinations of impaired cognitive functioning. These methods frequently rely on demographic variables such as education, occupation, and gender. These variables vary in their relationship to cognitive test performance across different languages and cultures. Premorbid cognitive functioning is also estimated by performance measures such as certain reading tasks, especially the reading aloud of irregularly spelled words. These types of tasks are very dependent upon the nature of the writing system. A word that is spelled irregularly in one language may not be spelled irregularly in another language. Semantic translation of these tests is not appropriate. Test developers have, instead, tried to find words or reading functions that are similar in various languages.

- *Behavioural presentation and Symptom and Performance Validity:* A test performance is a performance (in the theatrical, communicative, or social construction sense). There are considerable inter-individual and inter-cultural differences in how symptoms and disabilities are performed. How this is presented to providers and on tests is one small part of this. For example, everyone experiences pain, but there are many ways of showing, communicating, or performing pain. The last two decades have seen tremendous growth in the measurement of performance and symptom validity in neuropsychology with both free-standing instruments and embedded measures. Despite this growth, however, there remains considerable disagreement regarding the interpretation of many of the performances that fall outside of the normal range on such measures. Full validation of such measures in diverse cultures must await a clearer understanding of how disability and related states are regarded and "performed" in the target cultures and such encounters with professionals.

Cultural Adaptation/Development Strategies in Neuropsychological Assessment.

Within neuropsychology, the need for test translation and adaptation technologies has also been driven by the need for clinical and research instruments that are cross-culturally valid to provide accurate diagnoses and other clinical services and accurate research. This has become especially pertinent for the study of diseases that show significant ethnic and geographic epidemiological variation, such as dementia, multiple sclerosis, HIV, malnutrition, and tropical diseases. Nevertheless, when faced with such a need, the test translator/adaptor/developer is faced with several broad choices. It is worth noting, that although the ITC Guidelines and our commentary primarily address test adaptation and translation, there may be situations where test development may be the most appropriate option. All of these options are thus reflected in the following Assessment Adaptation/Development Typology:

- *Same test.* Translate and slightly adapt a test so that it can provide psychometric equivalence in the source and target languages to allow direct, quantitative comparisons across populations with what is considered to be "the same" test in both content and function. This is the primary focus of the ITC Guidelines (e.g., WAIS Block Design subtest in different languages, which uses all of the same items, but the order of item difficulty has been found to differ across languages, so the order of item administration has been adapted).

- *Test adapted version.* Translate and adapt a test to such a degree that it can serve similar functions in the target language but without psychometric equivalence or comparability so that it is considered to be a version of the test (e.g., WAIS Information subtest in different languages, in which a minority of the items refer to historical figures specific to the country of testing).

- *Test family.* Construct a test in the target language that follows the same design and principles or paradigm of the model test but with substantially new and culturally relevant content and procedures. Such a test can serve similar functions in the target language without psychometric equivalence or comparability. It can be considered to belong to a family of tests (e.g., the Hong Kong Verbal Learning Test is part of the Word List Learning "family" of tests).

- *Indigenous test version.* Construct a test in the target language/cultural context with a substantially new design and principles and culturally relevant content and procedures, but to measure the same construct(s) as in the original language/context. This may be regarded as an “indigenous” test, constructed from indigenous activities and materials (e.g., the “Which Car Test” for northern Australian Aborigines to assess executive functions and cognitive flexibility by using a social judgment problem specific to their cultural norms; Rock & Price, 2019).

- *Indigenous measure*. Construct a test in the target language with a substantially new design and principles and culturally relevant content and procedures, to measure indigenous construct(s). This may be regarded as an “indigenous” test, constructed from indigenous activities and materials (e.g., measures of traditional beliefs especially regarding health conditions and treatments). This category may also apply to measures of adaptive functions that may be used in neuropsychology and apply only to specific cultural contexts, for example, measures of competence to stand trial are only applicable within a specific criminal justice system.

The above typology is more of a continuum than distinctive categories. However, when the purpose is to demonstrate equivalence across languages (same test), there is a large collection of procedures to be followed to demonstrate test equivalence across languages. Some of these procedures are not necessary for uses that do not include direct comparisons of populations. However, even when a test has been demonstrated to be equivalent across populations, the populations may differ, and so separate norms are generally needed for neuropsychological purposes.

Furthermore, there are hybrids and variations in the above categories. For example:

- *The etic/emic model* combines measures of imported etic constructs with indigenous emic constructs within the same instrument (Cheung, 2012).

- *The multicultural model* involves commencing test development to measure constructs across several languages and cultures, rather than starting first in one language and later translating and adapting. Some have referred to this as the Universal Design model, but testing itself is a non-universal cultural enterprise, so a truly universal test is not possible. It is possible, however, to develop tests that function reasonably reliably across many languages and cultures, provided that those languages and cultures share the underlying assumptions and constructs of the test, (e.g., European Cross-Cultural Neuropsychological Test Battery; Nielsen, et al., 2018).

- *The centering model* involves developing a new version of the original test in the original language during the process of translating/adapting the test so that the new version is culturally and linguistically appropriate across the target languages.

It is not possible to give full guidance on how to choose an assessment strategy from this array of options. Such decisions will depend upon the anticipated purposes of the assessment tools and educated judgments regarding what strategies are most likely to produce successful, reliable, valid, and acceptable instruments. These Guidelines and Applications may assist project teams in making such choices by allowing them to anticipate steps needed and complications that may be encountered.

Selecting tests for translation, adaptation, or development.

As a starting place, we suggest that the development team should not start at the point of deciding what and how to adapt a specific test, but rather should begin by carefully considering what they wish to accomplish with their project, what constructs they wish to measure, in what populations, and for what purposes. Once they have clarified their purposes, the team can select the most promising strategies and typologies to accomplish their purposes. It is rarely possible to anticipate all the uses to which a test might be put in the future, but a reasonable effort to anticipate major uses may help to avoid wasted efforts and misleading and even harmful tools and applications.

By focusing on the purpose of the project rather than simply an accurate translation of the test, the development team may be able to extend their work past the foundations set forth by the ITC Test Adaptation and Translation Guidelines process. Specifically, the background section of the Guidelines acknowledges that test adaptation includes "deciding whether or not a test in a second language and culture could measure the same construct in the first language." But the Guidelines themselves do not address this process. They proceed after a decision has been made to translate and adapt a specific test. Often such a decision may be based upon the desire of the test designer and/or manufacturer to expand their research and/or market. Or the decision may be based on a hypothesis that the test will function similarly in the new language/culture/market. Our application addresses the adaptation decision in PC-2 and elsewhere.

A common tendency in the history of testing is that when psychology or related testing disciplines encounter a new domain of application, the initial tendency is to apply existing tests to that domain to see how well they function. Frequently, they do not function very well and there is a need to either revise the tests or develop new tests specifically targeted at that purpose. This type of sequence can be seen, for example, as the fields of neuropsychology, forensic psychology, school psychology, health psychology, industrial psychology, and cultural psychology each developed. Learning from this tendency, it is to be expected that even if a translation and adaptation of a test may serve one set of functions well in the target language, it may not serve the entire range of functions it is used for in the language of origin.

If test developers first stop and consider the goals of their project before engaging in the test development and adaptation process, they will have the opportunity to reconsider the field of applications, constructs, and validities of a test. Tests often show drift from their original purpose and understanding as they are applied to new domains. An adaptation/translation project is an opportunity to consider whether the test is still serving its original purposes and if its current uses remain appropriate. It is an opportunity to hypothesize which uses are likely to be valid in the target language and to plan the project in such a way as to address those potential uses.

Yet, there are no clear guidelines regarding when to use each of the above strategies in test development or adaptation projects. Such guidance would be useful in avoiding disappointing projects that fail to achieve test validity in a target language and culture. Below we will make a beginning attempt at such guidance.

Cost.

Neuropsychological testing is well developed in most major European languages and cultures. These languages and cultures have the advantages of relative similarity among them making test translation and adaptation relatively easy. Test translation and adaptation get more difficult when moving beyond the North Atlantic countries and Australia to countries where European languages are colonial languages, such as the rest of the Americas, much of Africa, and parts of South Asia. In these settings, European languages may be established but spoken as second or third languages and cultures are often markedly different. Moving beyond that are languages and cultures that are quite different from European origins where test translation and adaptation is quite challenging. Often these are settings in which there is very limited funding for research, health care, and other domains of use of neuropsychological tests. Faced with such situations, research teams will need to make pragmatic decisions concerning which of the recommendations in these applications are of greatest priority for their particular circumstances and goals. We hope that such teams sharing their experiences will assist future projects in their prioritising. We also hope that these considerations will be recognized by international organisations and funders. We recognize that the considerable costs of test development and the potential benefits must be weighed against other health and non-health funding priorities.

Pre-Condition Guidelines

PC-1 (1) Obtain the necessary permission from the holder of the intellectual property rights relating to the test before carrying out any adaptation.

ITC:

Explanation. *Intellectual property rights refer to a set of rights people have over their own creations, inventions, or products. These protect the interest of creators by giving them moral and economic rights over their own creations. According to the World Intellectual Property Organisation (www.wipo.int), "intellectual property relates to items of information or knowledge which can be incorporated in tangible objects at the same time in an unlimited number of copies at different locations anywhere in the world."*

There are two branches of intellectual property: Industrial property and copyright. The first one refers to patents protecting inventions, industrial designs, trademarks, and commercial names. Copyright refers to artistic and technology-based creations. The creator (the author) has specific rights over their creation (e.g., prevention of some distortions when it is copied or adapted). Other rights (e.g., making copies) can be exercised by other persons (e.g., a publisher) who have obtained a licence from the author or copyright holder. For many tests, as with other written works, copyright is assigned by the author to the publisher or distributor.

As educational and psychological tests are clearly creations of the human mind, they are covered by intellectual property rights. Most of the time the copyright does not refer to specific content of items (e.g., no one has rights on items such as "1+1 = ..." or "I feel sad"), but to the original organisation of the test (structure of the scales, scoring system, organisation of the material, etc.). Consequently, mimicking an existing test, i.e., keeping the structure of the original test and its scoring system but creating new items, is a breach of the original intellectual property rights. When authorised to carry out an adaptation, the test developer should respect the original characteristics of the test (structure, material, format, scoring . . .), unless an agreement from the holder of the intellectual property allows modifications of these characteristics.

Suggestions for practice*. Test developers should respect any copyright law and agreements that exist for the original test. They should have a signed agreement from the intellectual property owner (i.e., the author or the publisher) before starting a test adaptation. The agreement should specify the modifications in the adapted test that will be acceptable regarding the characteristics of the original*

Neuropsychological Application

Explanation. As both scientists and practitioners, neuropsychologists have a major commitment to the accuracy of measurement, test validity, and linguistic and cultural appropriateness. These values may manifest themselves somewhat differently depending upon the anticipated uses of a translated/adapted test. For example, translation/adaptation strategies may differ if the goal is to make direct comparisons between populations versus measuring a construct as accurately as possible in the target language/culture/population. However, these scientific measurement goals in whatever form they take should take priority over the goals of preserving intellectual property rights. This may, at times, place the goals and interests of test translators/adaptors in conflict with those of copyright holders.

Suggestions for practice.

- *Accuracy first.* Test developers and adapters should strive to maintain construct validity, collaborate with original test developers, and formulate a clear and transparent process throughout the translation/adaptation process. Test integrity and validity should be prioritised above considerations of intellectual property rights and naming. Specifically, the priority should be on accurate measurement in the target population for the anticipated test purposes rather than the commercial or reputational interests of the holders of the intellectual property rights. For example, there have been many attempts to translate word memory tests that have proven disappointing in their translated validity. Consequently, many have preferred to generate new and more appropriate word lists in the target language, while still following known or similar word list memory test paradigms. This is not only our scientific and clinical obligation but is arguably also in the long-term best interests of the holders of the intellectual property rights so that their intellectual property does not deteriorate in value and reputation.

- *Collaboration with test authors.* Getting permission from the original test developer is an important part of the adaptation process, regardless of whether the test is free to use or not. Specifically, the author(s) of the original test made certain considerations in the item selection, item order, and scoring criteria. Including the original test developers in the adaptation process – by discussing why the test may need to be adapted for the target population – can provide insight regarding how much adaptation may be required, as well as any limitations to adaptation. The test adapters can also add a rationale in the adapted test manual explaining why the original test was not suitable for the intended population.

- *Naming, description, and credit.* The adaptation process should be considered an iterative rather than a linear approach. A few issues should be discussed with the original test developer and/or publisher. These include issues such as the name of the revised test, description of the test, and authorship. The question of how much adaptation is allowed for a test to be considered a version of the original test needs to be considered. However, all adaptations necessary to ensure that the test is suitable for its intended purpose in the new cultural context must be made. For this reason, it is important, wherever possible, to avoid agreements specifying a limit on the number of items that may be adapted, as the adaptation and piloting process (see TD-5) may reveal empirically that a culturally or linguistically appropriate adaptation goes beyond the agreement.

  Furthermore, potential changes in the design of stimuli, test instructions, quantitative metrics, or testing/scoring procedures should also be discussed. For example, if the adaptation requires a re-shuffling of the order of test items for it to make sense to test-takers in the target population, would that effectively alter the core aspects of the test and its interpretation? In some tests, the item order is determined by item difficulty, and this can change during the adaptation process because of various factors such as semantic and cognitive equivalence. The Test Adaptation Typology described in the Application Background comments are a continuum. We are not international intellectual property rights lawyers and cannot fully resolve such issues here. The critical components of neuropsychological tests often lie in their design and procedures rather than the content. For this reason, intellectual property rights may resemble patents more than copyrights. Seeking such agreements in advance based on existing precedents may help to facilitate such processes.

Overall, the aims of the process should be shared with the developer from the outset, with the possibility of an ongoing collaborative venture if more changes are indicated. An agreement on the intellectual property rights of the adapted test is necessary to acknowledge the work of the test developers to make the test suitable for a new target population. This is also needed for naming the test. The claims of institutions regarding intellectual property rights also need to be taken into consideration.

PC-2 (2) Evaluate that the amount of overlap in the definition and content of the construct measured by the test and the item content in the populations of interest is sufficient for the intended use (or uses) of the scores.

ITC:

Explanation. *This guideline requires that what is assessed should be understood in the same way across language and cultural groups, and this is the foundation of valid cross-cultural comparisons. At this stage in the process, the test or instrument has not even been adapted so compilation of previous empirical evidence with similar tests, and judgements of construct-item match and suitability for the language groups involved in the study would be desirable. Ultimately, however, this important guideline must be assessed with empirical data along the lines of evidence required in C-2 (10). The goal of any analyses is not to establish the structure of a test, though that is a by-product of any analyses, but to confirm the equivalence of the structure of the test across multiple language versions.*

Suggestions for Practice. *Individuals who are experts with respect to the construct measured, and who are familiar with the cultural groups being tested, should be recruited to evaluate the legitimacy of the construct measured in each of the cultural/linguistic groups. They can try and answer the following question: Does the construct make sense in the cultures of both groups? We have seen many times in educational testing, for example, that a committee has judged the construct measured by a test to lack meaning or have diminished meaning in a second culture (for example, quality of life, depression or intelligence). Methods such as focus groups, interviews and surveys can be used to obtain structured information about the degree of construct overlap.*

Neuropsychological Application

Explanation. Neuropsychological constructs are not necessarily universal. As the first point of consideration, test adapters should ensure that the construct measured by the test exists in the vocabulary of the target population. If the construct does not exist, is it practical to attempt to measure it in the intended population? An example of a frequently measured construct in neuropsychology is intelligence. Research has established that there is not a clear consensus on what intelligence is. For instance, the western explanation of intelligence meets only one of 4 distinct terms of intelligence in Kenya, where intelligence is made up of *rieko* (i.e., knowledge and skills), *luoro* (i.e., respect), *winjo* (i.e., comprehension of how to handle real-life problems), and *paro* (i.e., initiative) (Fernadez & Abe, 2018). This supports the notion of the influence of culture on this purportedly universal construct.

While some neuropsychological constructs are deemed universal (e.g., working memory), it has been suggested that culture tends to influence the way these constructs are expressed. By exploring and understanding these constructs, with the aim to adapt their measurement within different cultures without assuming universality, test adapters would likely be a step closer to assessing construct equivalence among tests purporting to measure similar domains (e.g., comparing measures of attention and executive function in one culture). Translation and adaptation projects should therefore be built considering constructs and their uses. This principle applies to all domains of neuropsychological constructs, diagnoses, cognitive and behavioural domains, symptoms, syndromes, and adaptive abilities.

The test adapters should assess whether the construct has the same qualities across cultures. For example, some cultures explain and understand their experiences somatically, whereas others explain and understand their experiences using conventional psychological terminology.

Examining the item-content equivalence is another critical component of the test adaptation process. For example, in some cultures an umbrella is most used for protection from the rain; however, in countries such as Botswana, it is commonly used as protection from the sun. So, although the “umbrella” does exist across cultures, how it is used varies which may affect responses in some tests such as abstract reasoning tests that require respondents to identify relationships between objects. Therefore, adapters must ask themselves whether the items are examining the same cognitive process as the original intentions of the test developers. For example, the silhouette task, a subtest of the Visual Object and Space Perception battery (Warrington & James, 1991), requires test-takers to identify an object from its silhouette. This task is supposed to measure perception, however, familiarity with the test items and familiarity with the silhouette representation of the items will influence the test-takers’ ability to correctly match the silhouette to their cognitive representation of that item (Chatterjee, 2021).

In summary, test adapters should assess whether the construct has the same qualities across cultures/countries.

Suggestions for practice: Establishing construct equivalence requires experts in more than just languages. The test adapters should be knowledgeable in the culture of the target population as well as the intended uses of the test.

- *Construct equivalence (see C-2)* should be established early in the adaptation process, including considerations for diagnostic validity and application to adaptive behaviour. In other words, adapters should ensure that the test captures the construct in a way to assist with diagnosis and other purposes of neuropsychological assessment in the context for which is intended. A starting point would therefore be to investigate the construct of interest in the target culture/population by involving language and cultural experts of the population of interest (e.g., academic, professional, and local informants) and exploring how this construct is currently understood and measured. Another pivotal step towards the development of accurate and appropriate cross-cultural neuropsychological testing is to examine the specific ways by which test biases may likely develop. Once the construct is well understood from the perspective of the target population, a step towards either adapting an existing test or developing a new one would be based on a comprehensive understanding of the construct.

- *Intended purposes.* Even if construct equivalence is achieved, the adapted test may not serve the intended purposes. This concerns the current and historical uses of a test – the adapters should establish whether the test is valid for what they want to use it for in the intended population. Neuropsychological tests can be used to make determinations about individual functioning in the real world; therefore, the ecological validity of tests comes into question. Test adapters, especially those who intend to use the test to infer individual functioning, should be mindful of whether the adapted test can still serve the same function by adding measures to assess predictive validity in the validation process. Moreover, they should determine whether the intended uses are relevant to the intended population, including any possible negative consequences. When possible, adapters should also try to anticipate new uses for the test in the target population to consider whether it may serve those purposes. For example, the Trail Making Test (Reitan, 1955) was developed to detect brain damage, but has since been used in many countries to help to determine the ability to drive a vehicle (Vaucher, et al, 2014). If the Trail Making Test is translated and adapted to a new language and validated for detecting brain damage, that will not guarantee that it has a similar predictive ability for driving capacity in the new population. Such validation will need to be carried out separately. Otherwise, evaluees could be unfairly deprived of the privilege of driving.

PC-3 (3) Minimise the influence of any cultural and linguistic differences that are irrelevant to the intended uses of the test in the populations of interest.

ITC:

Explanation. *The cultural and linguistic characteristics irrelevant to the variables that the test is intended to measure should be identified at the early stage of the project. They can be related to the item format, material (e.g., use of computer, pictures or ideograms…), time limits, etc.*

An approach to the problem has been to assess the 'linguistic and cultural distance' between the source and target language and cultural groups. Assessment of linguistic and cultural distance might include considerations of differences in language, family structure, religion, lifestyle, and values (van de Vijver & Leung, 1997).

This guideline relies mainly on qualitative methods and specialists familiar with the research on specific cultural and language differences. It places special pressure on the selection of test translators and requires that translators be native to the target language and culture, since knowing the target language only is not sufficient for identifying possible sources of method bias. For example, in the Chinese-American comparative study of eighth-grade mathematics achievement carried out by Hambleton, Yu, and Slater (1999), format and test length problems were identified, along with a host of cultural features associated with the eighth-grade mathematics test.

Suggestions for practice. *This is a difficult guideline to address with empirical data at any time. It is especially difficult at the early stages of test adaptation. At the same time, qualitative evidence can often be collected:*

> *Either by observation, interview, focus group, or survey, determine motivational levels of participants, their understanding of the instructions, their experience with psychological tests, the speediness associated with test administration, familiarity with the rating scales, and cultural differences (but even these comparisons could be problematic because of cultural differences in understanding the variables themselves). When collecting such research data from participants is problematic, obtain as much information as possible from the translators. Some of this work could be done prior to any progress with the test adaptation.*
>
> *It may be possible to control for these 'nuisance variables' in any subsequent empirical analysis once the test has been adapted and is ready for validation studies via the use of analysis of covariance or other analyses, that match participants across language/cultural groups on variables such as motivational level or familiarity with a particular rating scale (e.g., Johnson,*

Neuropsychological Application

Explanation. It can be difficult for test developers/adaptors to determine which cultural or linguistic factors may be irrelevant, particularly if they themselves are not native to the language or culture of the population of interest. In essence, test developers may be blinded to these differences because “you don’t know what you don’t know.” Addressing this issue therefore requires a carefully thought out and deliberate approach. Translators can certainly be one source of information during the test adaptation process, but, by definition, translators are literate and therefore have a certain level of education. Neuropsychology often focuses on populations at the extremes of cognitive abilities. So, for example, translators may not be able to identify nuisance variables that would be most pertinent to test-takers who have no formal education. Test adaptation and translation projects are increasingly designed to extend neuropsychological capacities to linguistic and cultural populations that have been neglected or excluded.

Effects of irrelevant differences should be minimised early in the project, but such differences will likely continue to emerge throughout the process. Test adapters should be vigilant in finding such differences throughout the process and incorporate searches at each stage. Minimising such differences will be most important for uses that involve direct comparisons across populations. Direct review, surveys, piloting, debriefing of test-takers and administrators, and focus groups may all be helpful, but such participants may not always be able to directly articulate what the irrelevant factors are. This may require inference, hypotheses, and testing hypotheses.

Suggestions for practice.

- *Constructs and Uses*. To implement this guideline fully it is necessary to go beyond the content of the source test to examine 1) the nature of the construct the test is intended to measure and 2) the nature of the intended uses of the test in the target language/culture. This is an important stage for determining the feasibility, the strategies, and the goals of the project. This early stage of the project is the best time to determine whether the most feasible strategy is to develop 1) the same test, 2) an adapted version, 3) a test family version, or 4) an indigenous test version (see Application Background section). This stage may call for expertise and professionals who have knowledge and skills that go beyond translation, language, and test construction (see TD-1). The nature of the intended construct in the target culture may best be understood through ethnographic studies and the use of appropriate experts. Intended test uses may best be understood by consulting intended test users (i.e., test administrators, those who refer for or request test results, and test-takers). Some of this may need to take place before any translation is attempted. This type of consideration should also continue during the review and piloting of the first translation/adaptation drafts (see TD-5).

- *Test Technologies.* Consider how test technologies and techniques show wide cultural variation in their familiarity and usefulness. For example, languages such as Japanese and Gaelic do not have words for yes and no, and so responses to polar questions need to be reformatted. This may mean choosing between the positive and the negative form of a statement (e.g., Choose: “Are you depressed?” or “Are you not depressed?” instead of: “Are you depressed? Choose yes or no”). Many cultures are not familiar with multiple choice or Likert scales. For example, Chinese culture values the doctrine of the mean and avoids extreme responses, Chinese individuals may tend to choose neutral responses on Likert Scales. Previous research also found that Chinese and Japanese are more likely to choose midpoints when requested to admit positive emotions compared to European Americans (Wang, et al., 2008). Many Western neuropsychological tests are speeded, but cultures vary in the degree to which they value speed and in their speed/accuracy trade-offs. Similarly, cultures vary in willingness to guess at answers and risk wrong answers. Cultures may vary in what they regard as repeating or remembering “the same” word or story so synonyms or paraphrases may or may not be considered correct. Cultures also differ in the knowledge of cardinal directions. For example, some Australian aboriginal languages do not typically refer to something as to a person’s left or in front of them, but rather to their west or north. Cultures differ in their familiarity with and names for geometric shapes. Different cultures may make different presumptions regarding whether rotations and mirror images of drawings are considered “the same thing.”

- *Intracultural Variability*. Within cultures, there may be age and education variations in familiarity with test materials and techniques. Many countries have had important changes in their education policies, language policies, and even writing systems within their current lifetimes, and this will greatly influence what materials and skills are familiar to different generations. For example, it is nearly universal that younger generations are more familiar and facile with computers, smartphones, software, and other information technologies than older generations. Access to information technologies also shows great variation along many dimensions. Likewise, there can be great intracultural variation in familiarity and facility with writing implements, writing systems (print versus cursive; typing versus dictation; Hangul versus Hanja; simplified Chinese versus pinyin; direction of reading, use of emojis and abbreviations), perception of line drawings and representation of three dimensions, and familiarity with test materials such as blocks and abstract geometric forms. Because of these intracultural variations, it is necessary to consult the full range of intended test users early to minimize the influence of irrelevant cultural and linguistic differences and to address these through piloting (see TD-5).

- *Test Item Content.* Once issues of test constructs, uses, and technologies have been addressed, differences in irrelevant cultural and linguistic content can also be addressed at the item level. Generally, item content should have similar levels of difficulty and familiarity relative to the culture and across all subcultures of anticipated application. For example, one US neuropsychological test of speech articulation included repetition of "Methodist Episcopal." This item would most likely be easier for a Methodist Episcopal than for a Vajrayana Buddhist. On naming tests familiarity, age of acquisition, and frequency need to be considered both for the word or vocabulary and for the visual images. Items should be reviewed for ambiguities and unintended meanings, including potentially offensive items or items that may suggest affiliation with a particular religious, ethnic, gender, social, or political group or perspective. Focus groups and pilot studies should include the full range of test users as much as possible (see TD-5). Qualitative feedback is important.

- *Translators as Cultural Experts.* Translators can be one very useful source of information to assist with this process, especially regarding features of language, dialects, and idioms as their areas of professional expertise. Nevertheless, they should not be relied upon as the only source, since they may not be aware of or able to articulate cultural differences and may not be aware of how such differences impact the construct under study. Furthermore, the power differential between translators and researchers may make it difficult for them to raise such issues. For example, the first posted Vietnamese translation of the Montreal Cognitive Assessment asked test-takers to say all the words they could that began with the letter F, but Vietnamese does not have a letter F. It appears likely that the translator knew this but, because of the power differential, it may be possible that they did not feel able to tell the test designers. This was only rectified through feedback from test users.

  Another challenge with engaging professional translators as cultural experts emerges when translators use formal vocabulary rather than the more commonly spoken language. This poses a problem for test-takers who are not familiar with the formal words and are not proficient in other languages. Local community members can be engaged as part of a team of cultural experts to ensure that the preliminary work applies to the local setting.

TD-1 (4) Ensure that the translation and adaptation processes consider linguistic, psychological, and cultural differences in the intended populations through the choice of experts with relevant expertise.

ITC:

Explanation. *This has been, over the years, one of the most impactful guidelines because there is considerable evidence suggesting that it has been influential in getting testing agencies to look for translators with qualifications beyond knowledge of the two languages involved in the test adaptation (see, for example, Grisay, 2003). Knowledge of the cultures, and at least general knowledge of the subject matter and test construction, has become part of the selection criteria for translators. Also, this guideline appears to have been influential in encouraging agencies translating and adapting tests to use at least two translators in various designs (e.g., forward and backward translation designs). The old practice of relying on a single translator for all decisions, however well qualified that person might be, has been eliminated from the list of acceptable practices today.*

Knowledge/expertise in the target culture results from using translators who are native in the target language and are living in the target locale, with the former being essential and the latter highly desirable. A native speaker of the target language will not only produce an accurate translation, but also one that reads fluently and appears indigenous. Living in the target locale will ensure up-to-date knowledge of the current language use.

Our definition of an "expert", then, is a person or a team with sufficient combined knowledge of (1) the languages involved, (2) the cultures, (3) the content of the test, and (4) general principles of testing, to produce a professional quality translation/adaptation of a test. In practice it may be effective to use teams of people with different qualifications (for example, translators with and without expertise in the specific subject, a test expert, etc.) in order to identify areas that others may overlook. In all cases, knowledge of general principles of testing, in addition to knowledge of the test content, should form part of the training that translators receive.

Suggestions for practice. *We would suggest the following:*

> *Choose translators who are native speakers of the target language and have an in-depth knowledge of culture into which a test is being adapted, preferably living in the target locale.* ***A common mistake is to identify persons as translators who know the language, but not very well the culture,*** *because an in-depth knowledge of the culture is often essential for maintaining cultural equivalence. Having the cultural knowledge will identify cultural references (e.g., cricket, Eiffel Tower, President Lincoln, kangaroo etc.), with which local participants may be unfamiliar.*
>
> *Choose translators, if possible, with experience in the content of the test, and with knowledge of assessment principles (e.g., with multiple-choice items, the correct answer should be about as long and no longer or shorter than other answer choices; grammatical cues should not be helpful in locating the correct answer; and, in true-false items, true statements should not be notably longer than false statements).*
>
> *Translators with knowledge of test development principles may be nearly impossible, in practice, to find, and thus it would be essential to provide training for translators to provide them with the item writing principles for the formats with which they will be working. Without the training, sometimes overly conscientious translators will introduce sources of error, which can lower the validity of a translated test. For example, sometimes a translator might add a clarifying remark to ensure an intended correct answer is in fact the correct answer. In doing so, the translator may make the item easier than was intended, or the longer correct answer choice may provide a clue*

Neuropsychological Applications

Explanation: Adapting a translation to the target locale is called the “localization” of the translation (Harris, 2022). Using translators living in the target locale will also contribute to an understanding of how the testing process and procedures might be viewed locally, although translators should not be the only source of such perspectives. Localization is especially important for languages that are widely diffused and show local variation (pluricentric languages), such as English, Spanish, French, Arabic, Chinese, Swahili, and Quechua.

For neuropsychology, the definition of an “expert” should include not only knowledge of 1) the languages involved, 2) the cultures, 3) the content of the test, and 4) general principles of testing, but also 5) knowledge of the constructs of the test and their measurement. For example, take the item, “How are these two things alike, Orange-Banana?” A translator who is an expert in the content of the test and in Spanish localization would know to translate “banana” as “guineo” for the Caribbean but as “banano” for other parts of Latin America. But if “orange” had been translated” as “anaranjado” (the colour orange rather than the fruit orange), this would back-translate fine, it could look OK to a content expert, but it would take a construct expert to recognize that this item misses the intended construct (the similarity of being fruit. It would be like asking, “How are these two things alike? blue – banana”). Therefore, including persons with expertise in general principles of testing would be insufficient. This process should also require expertise in test constructs.

In addition to item localization, whole subtests can be similarly impacted. For example, some semantic fluency tests instruct test-takers to “Tell me all the names of different kinds of vegetables that you can think of.” A translator who is an expert in the *content* of the test may not find an issue when the word “verduras" can prompt Spanish-speaking test-takers from Puerto Rico to confine their responses to the names of starch or root vegetables (i.e., yuca, yautía, potatoes), rather than also including the intended broader category of vegetables (cucumbers, okra, peppers, etc.), creating an inherently more difficult task (Bender, García, & Barr, 2010). Including persons with expertise in general principles of testing would be inadequate; it also requires expertise in *constructs* and their measurement, as well as knowledge of intracultural differences.

Suggestions for practice: When translation is appropriate, test materials should also be adapted to be relevant to the test use settings. Therefore, the project team should include members with cultural, content, and testing expertise in the target language/population. For example, a neuropsychologist from the target population may not know the source language well enough to be a translator but may be a valued team member because they bring cultural, content, and testing expertise to the target language/population and can adapt the translation appropriately. In other words, while the original ITC Guideline above recommends addressing this issue by training translators in test development principles, we go beyond, recommending working with professionals with neuropsychological test expertise and expertise in the language and culture. Translations should preferably be conducted by a team consisting of subject matter experts and researchers working in areas related to the functions or constructs of interest. In particular, teams for tests of language functions and aphasia should include linguists and aphasia experts such as speech-language pathologists (Ballard, Charters, & Taumoefolau, 2018; Ivanova & Hallowell, 2013; Paradis, 1987).

For individually administered tests with complex instructions, it may be useful to include target test administrators in the translation team. This will allow for pilot studies with those who will administer the tests, which can help to refine and validate the translation of the instructions. It may also be helpful to include translators who are experienced interpreters to implement a protocol described in TD-2.

Finally, it may be useful to include other test users and stakeholders either as experts, as focus groups, or in surveys. For example, the stakeholders in a screening test for dementia may include psychologists, several medical specialties, nurses, medical paraprofessionals, and research assistants who may be called upon to administer the screener, researchers, aging and disability administrators, people with dementia, families, and caregivers. The stakeholders in a test for learning disabilities may include psychologists, speech-language pathologists, several medical specialties, teachers, educational administrators, parents, and students.

TD-2 (5) Use appropriate translation designs and procedures to maximise the suitability of the test adaptation in the intended populations.

ITC:

***Explanation**. This guideline requires that decisions made by translators or groups of translators maximise suitability of the adapted version to the intended population. This means the language should feel natural and acceptable; focusing on functional rather than on literal equivalence. Popular translation designs to achieve these goals are forward translations and backward translations. Brislin (1986) and Hambleton and Patsula (1999) provide full discussions of the two designs, including their definitions, strengths and weaknesses. But it should be noted that both designs have flaws, and so rarely would these two designs provide sufficient evidence to validate a translated and adapted test. The main drawback of the backward translation design is that, if this design is implemented in its narrowest form, no review of the target language version of the test is ever done. The design too often results in a target language version of the test which maximises the ease of back translation, but sometimes produces a rather awkward target language version of the test.*

A double-translation and reconciliation procedure is aimed to address the shortcomings and risks of relying on idiosyncrasies of single translations. In this approach, a third independent translator or an expert panel identifies and resolves any discrepancies between alternative forward translations, and reconciles them into a single version. In large-scale cross-cultural assessment programmes such as PISA, two different language versions (for example, English and French), may be used as separate sources for translation, which are then reconciled into a single target language version (Grisay, 2003). This approach offers important advantages, such as possible discrepancies are identified and reviewed directly in the target language. In addition, using more than one source language helps minimise the impact of cultural characteristics of the source.

Differences in the language structure can cause problems in test translation. For instance, in a well-known scale developed by Rotter and Rafferty (1950) in English, examinees are required to fill in the blanks in incomplete item format such as, "I like….."; "I regret….."; "I can't .....". *However, the same format is inappropriate in the Turkish language, where the object of a sentence must come before the verb and subject. The use of incomplete sentences as in the English version, therefore, would change the answering behaviour completely since the Turkish students should first look at the end of the statement before they fill out the beginning.*

In any alternative solutions to this problem, the translated (i.e., target language) version will be somehow different than the source language version in terms of format specifications.

***Suggestions for practice.** The compilation of judgemental data from reviewers seems especially valuable for checking that this guideline is met:*

> *Use the rating scales advanced by Brislin (1986), Jeanrie and Bertrand (1999), or Hambleton and Zenisky (2010). Hambleton and Zenisky provide an empirically validated list of 25 different features of a translated test that should be checked during the adaptation process. Sample questions from the Hambleton and Zenisky (2010) include* "Is the language of the translated item of comparable difficulty and commonality with respect to the words in the item in the source language version?" *and* "Does the translation introduce changes in the text (omissions, substitutions, or additions) that might influence the difficulty of the test item in the two language versions?"
>
> *Use multiple translation designs if practically feasible. For example, a backward translation design can be used to double-check the target version created through double-translation and reconciliation by an expert panel.*

ITC (continued):

If a test is intended to be used cross-culturally, consider simultaneous / concurrent development of multiple language versions of the test from the start in order to avoid future problems with translating/adapting the source version. More information on concurrent test development can be found, for example, in Solano-Flores, Trumbull, and Nelson-Barber (2002). At the very least, design the source version that enables future translations and avoids potential problems as much as possible; specifically, avoiding cultural references, idiosyncratic item and response formats, etc.

Considering the syntax differences across languages, using formats that rely on the rigid structure of sentences should be avoided in large-scale international assessments and probably with psychological tests, too, because of the translation problems that may arise.

Neuropsychological Applications

Explanation: To achieve construct and measurement equivalence, test translation/adaptation strategies may need to vary markedly among test instructions, test type (e.g., language/aphasia, memory, attention, executive functions), inventories, etc. The ITC Guidelines focus on questionnaires and inventories where translation techniques are designed to preserve the semantic content of tests as well as the construct equivalence as mediated by that semantic content. For many neuropsychological tests, the construct is not mediated by the semantic content of the test. Rather, many tests use language-mediated tasks to measure phonetic, grammatical, and lexical language skills; working memory; verbal memory; and executive functions. For these types of tests, translation techniques must prioritise construct equivalence over semantic equivalence to achieve a successful adaptation. At times, semantic content may be entirely irrelevant. Such translation goals may be unfamiliar to professional translators.

Suggestions for Practice:

- *Test Instructions*. Different language and cultural populations may differ dramatically in their test-taking experiences and presuppositions (Ardila, 2005). Neuropsychological tests are quite varied in test instructions and the expectations of the tasks. For these reasons, the translation of test instructions, even when semantically accurate, may not be adequate to assure construct and measurement equivalence. Generally, the ideal translation/adaptation of test instructions will result in test-takers in the target language having the same understanding and purpose in taking the test as test-takers in the source language. This may call for elaborated instructions or making explicit instructions that are implicit or assumed in the source language (Ardila, 2005). For example, the original instructions for the Symbol Search subtest of the WAIS assume that the test-taker will realise that mirror images and rotations do not “count” as being “the same" symbol. This implicit assumption may need to be made explicit for some cultures.

  One possible translation/adaptation strategy for such situations is to select team members who are native to the target language and culture and who are trained in testing. These team members develop instructions in the target language that are congruent and appropriate for that language and culture. Such instructions are then piloted. Criteria for the understanding of instructions need to be specific to each test. Problem-solving and piloting are repeated until a satisfactory result is achieved. While initial piloting may be done with a population of convenience, it is important that final piloting include the full range of possible test-takers where comprehension difficulties might be anticipated. This might include (depending on the test's purposes) elderly, very young, low formal education, hard of hearing, cognitively impaired, speakers of different dialects of the target language, and non-native speakers of the target language, for example.

Once the instructions to the test-takers are completed, a similar process will be needed for writing the instructions to the test *administrators*. Here, however, the piloting needs to be different. The target audience will be only those professionals who have sufficient training to be administering tests. If the intention is that the test user can learn to use the test from the manual only, without direct instruction, then the translated, adapted instructions should be piloted from the written instructions only. Again, the criteria for success are that both the test-takers and test administrators have the same understanding and purpose in taking the test as test-takers and administrators in the source language.

- *Language Tests.* Aphasia tests and other tests designed to measure specific language functions generally are not directly translatable but require adaptation or reconstruction in the target language (Ivanova & Hallowell, 2013). Fortunately, the major features of this process have long been available. For example, the *Bilingual Aphasia Test* (Paradis & Libben, 1987) phonemic discrimination subtest involves pointing at one picture out of 4 that represents the correct answer. The translation instructions state, "16 sets of 4 words in the target language that differ only in their initial consonantal sound (such as *man, pan*, *van,* and *fan* in English) are selected. . . . These words must be picturable. Hence, generally, they are concrete nouns, action verbs, and occasionally adjectives." Clearly, a semantically-accurate translation would, in most instances, defeat the purpose of such a subtest (with the above example, *homme, casserole, fourgonnette,* and *ventilateur* in French).

  These instructions are for bilingual aphasia tests, and many of the measures are sign tests based primarily on the expectation that most native speakers can do most of the items, rather than being based on a psychometric distribution of abilities across a broad normal range. Nevertheless, these translation/adaptation/development principles have been further refined to include controls for phonemic complexity, articulatory difficulty, word frequency and familiarity, age of acquisition of lexical items, morphological length and complexity, specific syntactic structures, syntactic complexity, verbal stimulus length, cultural relevance, verb tense, aspect, mood, noun case and gender, and type of script (Ivanova & Hallowell, 2013).

  - Confrontation naming tests often have broader use than aphasia tests and call for distinctive translation/adaptation strategies (Ballard, et al., 2018). This includes item selection and refinement, image creation, and determining naming difficulty.

  - Academic tests of language functions often serve both diagnostic and educational purposes. Such translation/adaptation projects generally face all the considerations of aphasia testing when looking at similar language functions and therefore can be partially guided by the same sources cited above. But they also often have the additional constraint or guidance of being criterion-referenced to educational curricula, which may vary by school system jurisdiction. The local educational testing subculture may also influence what may be the most culturally-appropriate testing formats. Literacy functions, in particular, show huge differences across languages in cognitive processes, ease of acquisition, and typical performance depending upon the transparency of the orthography and the type of orthographic system (abjad, abugida, syllabary, alphabet, semanto-phonetic, Omniglot.com). Important constructs in one system may be minor or missing in another. This is a domain in which test translators/adapters need to be particularly careful to evaluate whether their project is meaningful or viable at all.

- *Attention Tests.* Neuropsychologists often measure attention through various verbal tasks involving verbal working memory, mental maths, and verbal vigilance. Cross-cultural and cross-linguistic research has found that verbal working memory task performance is influenced by word length (Chan & Elliott, 2011), speech rates (Chincotta & Underwood, 1996), and literacy (Rosselli & Ardila, 2003). Thus, while the classic Digit Span task was once thought to measure primarily the storage of bits of semantic information with a capacity of 7 plus or minus 2 (Miller, 1956), it has become clear that verbal working memory is mediated by a phonological loop (Baddeley & Hitch, 1974) that varies by language. Thus, while digits are near universals with similar semantics and are convenient for task adaptations, translation/adaptation strategies for digit and mental maths tasks also need to account for phonemic, visual, and cultural/educational characteristics of digits across languages when targeting construct and measurement equivalence. Similarly, phonemic discriminability needs to be taken into account in verbal vigilance tasks. For example, the spoken letter names for /b/ and /v/ are indistinguishable in some dialects of Spanish. Likewise, the number of items to discriminate among may be of importance, for example, Hawaiian has 18 letters, while Thai has 59. So, while digit and letter tasks may translate readily, especially among closely related languages, translation/adaptation projects also need to pay close attention to language-related factors that can affect construct and measurement equivalence.

- *Memory Tests.* Verbal episodic memory is often assessed by using either word lists, story memory, or paired associate learning tasks. These tasks are impacted by factors such as age, education, language, and culture. Level and quality of education have also been noted to impact performance on memory tasks, with highly educated individuals displaying superior performances due to strategic encoding, chunking techniques, and prior experience with testing situations. Linguistic effects (i.e., syllabic word length) have also been noted to have an impact on memory encoding strategies and should also be taken into consideration when adapting or translating test material.

  Adapting a verbal memory task for use within non-English-speaking groups has proved challenging. Direct translation may be inadequate as the stimuli may be less familiar in the culture to which the test is being adapted than in the culture of the test’s origin Nell, 2000). This is particularly problematic with verbal memory tests, as word frequency in each language may differ considerably, changing the level of difficulty of the test from one language to another. These issues have also been noted across English-speaking nations. For example, Barker-Collo and collaborators (2002) administered an American list-learning test and a version modified to reflect New Zealand-relevant content and found that New Zealanders performed significantly more poorly on the American version of the task than on the culturally relevant, New Zealand version, even though both versions of the task were in English. This highlights the nontrivial influence of cultural factors above and beyond language.

Several suggestions have been made regarding methods for adapting neuropsychological tests to non-English-speaking groups. These include substituting culturally appropriate items for unfamiliar stimuli to ensure tests are fair, unbiased, and incorporate familiar stimuli. Translating existing list-learning tests from English to a target language, that at face value appears ecologically valid (i.e., Hopkins Verbal Learning Test–Revised, form 5; Professions, Foods, Sports (Brandt, & Benedict, (2001), has resulted in success (Vicente et al., 2020). In cases where new list learning tests/content have been developed for use in non-English speaking nations, the following methods employed in the development of established tests have resulted in the production of clinically viable tests. For example, the Greek Verbal Learning Test (GVLT) (Vlahou, et al., 2013), which was based on the structure and format of the CVLT, incorporated culturally relevant stimuli familiar to Greek nationals. Likewise, several translation/adaptation projects have found cultural relevance rather than semantic accuracy is fundamental to the success of story memory tasks.

While achieving semantic accuracy may be possible in certain respects, it should not be the primary goal in test adaptation. In verbal working memory and verbal memory, the outcome should be to have materials and tasks that are understandable in the target culture, an appropriate psychometric range of results, measure the intended cognitive constructs, and are clinically valid. Factors such as word length, familiarity with both the material and the task, and validities are critical (Messinis, Nasios, Mougias, et al, 2016).

- *Executive Functions Tests.* Culture has major influences on performance on executive function tests (addressed in the Application Background section). One major executive function task particularly susceptible to language effects is verbal fluency. Performance on letter fluency is influenced by demographic characteristics, with age and education being consistently identified as significant contributors to test performance (Cohen & Stanczak, 2000; Kempler et al., 1998; Tombaugh et al., 1999). The F-A-S condition has been reported to be comparable and appropriate for use across a range of Indo-European languages, reflecting a similar ratio of word frequency in languages such as English, Spanish, and Portuguese. In contrast, other studies have observed that F-A-S is not appropriate for clinical use in a range of other languages due to cross-linguistic differences and variable low word frequency rates. As a result, several cultures have chosen alternative letter fluency stimuli, including Greek (X-A-Σ; Kosmidis et al., 2004), Farsi (P-M-K; Ghasemian-Shirvan et al., 2018), and Hebrew (B-G-S; Kave', 2005). A further factor can be the number of syllables in each language. For instance, Kosmidis and colleagues (2004) observed that healthy Greek nationals (across all education bands; low, medium, high) generated fewer words relative to English (Tombaugh et al., 1999) and Spanish (Acevedo et al., 2000) groups matched for age and education. This discrepancy was attributed to the higher prevalence of polysyllabic words in Greek, as well as a decreased familiarity with such testing procedures. Letter fluency is likely to be influenced by the writing system and is almost meaningless for Chinese. It is also influenced by literacy and phonemic awareness. Translation/adaptation strategies for letter fluency tasks need to be adapted to the specific features of the target language and the target constructs. Semantic verbal fluency tasks need to consider the salience and familiarity of the semantic categories chosen within the target language, culture, and educational system.

- *Translation localization.* Some professional translators are capable of localising their translations when requested to regions where specific dialects or idiomatic features prevail. This need has emerged within many classic English-language neuropsychological tests. For example, regional flora, fauna, items, and foods in the Boston Naming Test (Kaplan, Goodglass, & Weintrab, 1983) are not well-recognized in other parts of the anglosphere. The North American "wrench" and "sweater" are "spanner" and "jumper" in Australia. Even words that are recognized in other regions may be less familiar, so the prime minister, archbishop, and haddock used familiarly in some British tests come across as more novel in the US. These types of differences have been found to have a significant impact within the anglosphere, such as with the New Zealand memory word list example above.

  This localizationist difficulty is best addressed at the stage of test design. Designers can strive towards universal or standard English, similarly understood and familiar throughout the anglosphere. Similarly, there are relatively standard versions of other major languages such as Spanish and Arabic. These versions are best elicited through language experts in these respective languages. In the case of developed but localised tests, decentering can be attempted in revisions. Translation strategies may also call for prioritising standard language versions over translation precision.

- *Back Translation. Back translation is generally an obsolete practice that produces overly literal translations that often capture neither the intended meaning nor the intended construct in the target language (Behr, 2017; Ozolins, et al., 2020; Son, 2018). This simple example shows how semantic equivalence and back translation can be inappropriate for some neuropsychological testing. Writing the following sentence is sometimes used to check that someone can form all the letters of the English alphabet: "The quick brown fox jumps over the lazy dog." This sentence translates into Spanish as: "El veloz zorro marrón salta sobre el perro perezoso." This back translates perfectly to English, preserving the semantic content, but the Spanish version is missing 13 letters of the Spanish alphabet. The meaning is preserved, but the construct is seriously under-measured. Translating and back-translating this sentence to Chinese, 敏捷的棕色狐狸跳过了懒狗, may come close to preserving the semantic content, but the task only demonstrates the participants' ability to write some Chinese characters at different difficulty levels rather than the ability to form letters. The construct of the ability to form the letters of the alphabet is close to meaningless for Chinese.*

TD-3 (6) Provide evidence that the test instructions and item content have similar meaning for all intended populations.

ITC:

Explanation. The evidence demanded by the guideline can be gathered through a variety of strategies (see, for example, van de Vijver and Tanzer, 1997). These strategies include (1) use of reviewers native to local culture and language; (2) use of samples of bilingual respondents; (3) use of local surveys to evaluate the test; and (4) use of non-standard test administrations to increase acceptability and validity.

Conducting a small try-out of the adapted version of the test is a good idea. The small try-out can employ not just test administration and data analysis, but also, and most importantly, interviews with the administrators and the examinees to obtain their criticisms of the test itself. Other designs using content experts from different language backgrounds, or bilingual content experts, are also possible. For example, bilingual content experts could be asked to rate the similarity of the difficulty of the item formats and content of the two tests. Cognitive interviewing is another method that is showing promise (Levin, et al., 2009).

Suggestions for practice. Several suggestions were offered above for addressing this guideline. For example,

> Use reviewers native to local culture and language to evaluate the test translation/adaptation.
>
> Use samples of bilingual respondents to provide some suggestions about the equivalence of the two versions of the test, both on test instructions and test items.
>
> Use local surveys to evaluate the test. These small-scale try-outs can be very valuable. Be sure to interview the administrator and the respondents following the test administration because often administrator and respondent comments are more valuable than the respondents' actual responses to the items in the test.
>
> Use adapted test administrations to increase acceptability and validity. Following similar test instructions makes no sense if they will be misunderstood by respondents in the second

Neuropsychological Application

Explanation: The use of a piloting procedure to ensure that the adapted test maintains validity for use in the intended population is necessary for a successful adaptation process. The inclusion of feedback from administrators and examinees is useful to ensure that not only do the potential users of the test understand the purpose but also that it serves its intended purpose. Furthermore, obtaining feedback from test-takers is also valuable because it includes the target population and gets their views on relevance and acceptability. For example, with the clock drawing test – the phrasing of the instructions “10 after 11.” In some contexts, such as Botswana, the same time would be stated as “10 past 11.” This is a minor difference but could lead some test-takers to indicate “10 minutes to 11” rather than the intended “10 after 11.”

- *Ecological relevance*. Ecological relevance refers to whether a construct/test item has relevance to everyday life (or the knowledge or skills) of people within the intended population.

Therefore, it is necessary to assess if test items maintain ecological relevance during the piloting process. For example, using a picture of a “beaver” to assess naming ability may not be relevant for people living in Asia just as it may not be relevant to assess naming ability in people from the US using a picture of a “jackfruit”.

Also, even when a translated test item assesses the same construct and achieves semantic equivalence, this may not result in a construct/item that is relevant to the population – which may affect performance and interpretation. Is it applied the same way across populations? Is it experienced the same way across populations? An example of this is the word “reluctant” when used in vocabulary tests. While this is a common term for English speakers, a mere translation to the Spanish word “reacio” would not be appropriate because it is a word used much less frequently in the Spanish language and, therefore, it may negatively impact scores on this subtest. Another similar example is the comprehension subtest. While a translation to Spanish would maintain the construct and semantic equivalence, the requested information may not be emphasised in academic curricula and individuals may perform lower due to limited exposure.

It is also worthwhile to consider that populations are not homogenous due to differences in cultural practices, acculturation, and other socio-contextual factors that will impact cognitive representations when interpreting performance on neuropsychological tests.

- *Translation Process.* Test adapters should be aware of the certain properties of language that can influence the understanding and interpretation of test items and interpretation. These properties include prototypicality, word frequency, and item difficulty. For example, nine of the twelve naming items on the *Addenbrooke’s Cognitive Examination* (Bak & Mioshi, 2007) from Britain were found to be insufficiently familiar with the Arabic translation for Saudi Arabia (Al Salman, 2013). Another example may be the US and UK versions of the MMSE. The orientation item: “Which State are we in?” in the original US MMSE was adapted to “Which county are we in?” in the UK version (due to the lack of states in the UK). However, this proved to be a more difficult item than states, which is most likely due to differences in the relevance of states and counties to people in the UK and the US. In these examples, the language is English for both target populations, however, because of differences in how the language has evolved, the relative commonality of various phrases and words will differ.

  There are multiple words for similar concepts in different languages, and the most appropriate word choice depends on subtle nuances in the contextual application. In psycholinguistics, the concept “Name Agreement” is used to denote the proportion of people who use the same word to denote a picture. This is used when standardising picture sets for use in psychological tests, such as naming tests. The frequency of the word and naming agreement can then be compared/matched across languages to find pictures with the same/similar difficulty. For example, in Spanish the word “colony” is translated to “colonia,” but “colonia” means either “colony” or “neighbourhood.” Similarly, the word ‘tent’ can be translated to Spanish as either ‘caseta’ (large tent) or ‘tienda’ (which can refer to a store as well as a tent). The most appropriate word in Spanish is determined by the context of the rest of the test item, but if the word stands alone, it is not disambiguated and can present subtle test translation inequivalence.

There are instances where establishing semantic equivalence is not sufficient because the test-takers have a different way of approaching the underlying aim of the test. For example, the "Draw A Person" test (Goodenough, 1926) was developed to assess cognitive development. This test has been widely used but may be less appropriate in certain contexts such as rural Zambia where children conceptualise the human form by "making a person" using clay rather than drawing and where the Panga Munthu (Make a Person) Test is more appropriate (Serpell & Simatende, 2016). This can also apply to any type of visuoconstructional drawing test used in unschooled and/or illiterate populations that will assess drawing skills rather than visuoperceptual abilities (e.g., Nielsen & Jørgensen, 2013). The same is true for semantic fluency using the category "animals," where people who are illiterate perform more poorly than literate people (e.g., Nielsen & Waldemar, 2016); again, this seems to be due to the higher ecological relevance of drawing and the animal category for people who are literate.

Suggestions for Practice:

- *Construct function.* Test adapters consider not only whether a tested construct exists in the target culture and matches the source construct, but also how that construct may function in the target culture. For example, the English concept of "dementia" also exists as "demencia" in Spanish, but it additionally carries an implication of insanity and often carries additional functions of stigma, avoidance, marginalisation, and isolation. The same is true for Arabic (depending on the Arabic word used (marad eaqliun/eatah/khabal): mental disorder, insanity, craze, lunacy, madness, dotard) and Turkish (bunama: second childhood, become a cabbage, dotage). It is not possible to reasonably anticipate all possible future uses of a test in all cultural contexts. However, we recommend that test adapters research the major current and historical uses of a test and consider how such uses might manifest in the target culture. They should then consider whether the proposed adaptation and constructs will serve those uses well.

  Test adapters may implement a piloting process to inform the practitioner how extensive the test adaptation process needs to be, depending on the above factors. In instances where semantic equivalence does not translate to ecological relevance, the adapters should consider how much change is required while still maintaining the intended properties of the test. Where the required changes are too extensive, then adaptation efforts should be abandoned in favour of developing a more appropriate and relevant test. For example, many different versions of the Mini Mental Status Exam (MMSE) (Folstein, Folstein, & McHugh, 1975) have been proposed for different cultures and/or populations with low/no education. It is questionable if it is the same test, particularly if many of the test items have been changed (e.g., Orientation items, constructional copying, sentence repetition, object naming, calculation). Instead, the Rowland Universal Dementia Assessment Scale (Storey, et al., 2004) was developed as an alternative to the MMSE that does not require many cultural or language adaptations across several populations in High-, Low-, and Middle-Income countries (Nielsen and Jørgensen, 2020).

TD-4 (7) Provide evidence that the item formats, rating scales, scoring categories, test conventions, modes of administration, and other procedures are suitable for all intended populations.

ITC:

Explanation. *Item formats such as five-point rating scales or new item formats such as "drag and drop" or "answer all that are correct" or even "answer one and only one answer choice" can be confusing to respondents who have not seen the item formats before. Even item layouts, the use of graphics, or rapidly emerging computerised item formats can be confusing to candidates. There are many examples of these types of errors found in the United States with their initiative to move much of the standardised testing of children to the computer. Through practice exercises, the problems can be overcome for most children. These new item formats must be familiar to respondents or a source of testing bias is introduced that can distort any individual and group test results.*

A newly emerging problem might be associated with computer-administered versions of a test. If respondents are not familiar with the computer-based test platform, a tutorial is needed to ensure that these respondents gain the familiarity they need for a computer-administered test to provide meaningful scores.

Suggestions for practice. *Both qualitative and quantitative evidence have a role to play in assessing this guideline. There are several features of an adapted test that might be checked:*

> *Check that any practice exercises are sufficient to bring respondents up to the level required for them to provide honest and/or responses that reflect their level of mastery of the material.*
>
> *Ensure that respondents are familiar with any novel item formats or test administrations (such as a computer-administration) that have been incorporated into the testing process.*
>
> *Check that any test conventions (e.g., the placement of any exhibits, or the marking of answers on an answer sheet) will be clear to respondents.*
>
> *Again, the rating forms provided by Jeanrie and Bertrand (1999) and Hambleton and Zenisky (2010) are helpful. For example, Hambleton and Zenisky included questions such as "Is the item format, including physical layout, the same in the two language versions?", and "If a form of word or phrase emphasis (bold, italics, underline, etc.) was used in the source language item, was*

Neuropsychological Application

Explanation: This guideline emphasises the importance of familiarity with test items and methods of test administration. This is important in making sure that unfamiliar test items and administration methods do not influence the results of the testing. However, in the context of neuropsychological evaluation, the phrase “suitable for all intended populations” involves far more elements than just familiarity.

There are some cases where establishing semantic equivalence is not sufficient. For example, Franzen and colleagues (2019) found a systematic performance difference among culturally, linguistically, and educationally diverse individuals in the Netherlands on the Visual Association Test (Lindeboom & Schmand, 2003). They found that memory performance for black-and-white line drawings was consistently lower than picture versions of the same stimuli. Therefore, despite the participants being familiar with the test items, the administration (black-and-white line drawings versus pictures) influenced performance.

Another factor to consider, especially for neuropsychological assessments, is the approach that is taken when completing the task. In some cases, the inherent strategy employed by the intended population is different from what the test developers had in mind. Test adapters should be cognisant of this possibility and include allowances for the variations in test-taking approaches, especially during the piloting phases. For example, speed and reaction time are key measures in many neuropsychological assessments, however, some cultures place less value on speed of performance and therefore extra emphasis may be required in the test instructions. Even if speed is valued, cultures may differ in how they implement a speed-accuracy trade-off in test performance.

Neuropsychology is very rich in different test methodologies, such as repetition, recall, recognition, multiple choice, Likert scales, ranking, visual-spatial copying, constructions, mental rotation, cloze procedures, naming, fluency tasks, writing, oral spelling, reading irregular words, analogies, similarities, definitions, answering questions, maths problems, visual search, auditory vigilance, tactile integration, odour identification, fine motor movements, etc. Each of these activities varies widely across languages and cultures in the degree to which they are taught in schools, practised outside of schools, and are feasible in different languages and writing systems. Test developers need to consider not only whether a task or procedure is feasible but also whether it is familiar, how it is regarded, and whether it is likely to measure the intended construct.

It is also vital to examine whether the scoring criteria of a test need to be modified, not only for tests that need to be adapted but also when the original test requires no adaptation for use in a new linguistic and cultural context. For example, while examining the appropriateness of the Clock Drawing Test in a Bengali-speaking population in India, it was observed that the low-educated healthy participants wrote the numbers within the clock face partly in English and partly in Bengali (Crombie et al., 2023). In this case, the Rouleau scoring system (Rouleau et al., 1992) used for scoring the test performance was modified by Crombie et al. to give credit for numbers written partly in English and partly in Bengali scripts.

Another example of where scoring criteria should be considered is verbal (letter and category) fluency tests when used in a bilingual or multilingual context. A key scoring issue is whether credit is given for words from different languages (e.g., whether credit is given for correct responses in one's first language (L1) and also borrowed or loan words in the second language (L2) which have become part of the regular L1 vocabulary). It is important to develop a scoring method that maximises the validity (e.g. sensitivity and specificity) of each form of verbal fluency test (see also TD-1). Even when a test scoring system does not require adaptation, it should not be assumed that the raw scores on a test should be interpreted in the same way in a different linguistic/cultural context. The raw scores on the test should be converted into standard scores based on the score distribution of the intended linguistic/cultural context.

Suggestions for Practice:

- *Review the impact of procedures on construct measurement.* The following questions can be asked of the adapted test:
  - Does the adapted test have functional suitability? Will the changes to the test items and administration tools still serve the same purpose (i.e., will it assess the same operational definition of the construct)? Test adapters should be mindful of the test purpose and ensure that the new version of the test serves the same function.
  - Will the response categories assess similar constructs? For example, do the response categories on a Likert scale have an inherent value about the construct being tested, so that a high score on the construct still represents the same values in the intended population?
  - Is the test simple, approachable, and easy to follow? Test difficulty can be a confounding factor when interpreting test performance and it must be considered for the intended population. Qualitative, or process approaches, can be included in the piloting phase to investigate test performance more holistically. Examples of qualitative approaches include error analyses, focus groups with test-takers and administrators, and feedback surveys.

- *Practice items.* The use of practice items for unfamiliar items is often used to avoid the effects of unfamiliar tests. In the case where practice items are provided, the goal is to give enough practice items for the test-taker to understand the test instructions without giving rise to practice effects. Instructions to the tester should provide clear directions on how to determine whether a test-taker truly understood the directions for the test. This is particularly important for some cognitive domains such as executive functions, which require novel tasks to assess functioning.

- *Consider whether the scoring criteria need to be amended to maximise the validity of the test in the new linguistic/cultural context.* Although a test may, in general, assess the same construct in different linguistic/cultural contexts, it is recommended that the scoring criteria are examined to determine if modifications may be needed to maximise the validity of a test. Data collected from a validity study could be examined to determine which scoring criteria maximise the test's validity, and ideally, this should be confirmed in replication studies.

We recommend an iterative piloting and feedback strategy during the test adaptation process to ensure that the above issues are taken into consideration during the adaptation process. Furthermore, the piloting should include both brain-impaired and healthy populations (see TD-5).

TD-5 (8) Collect pilot data on the adapted test to enable item analysis, reliability assessment and small-scale validity studies so that any necessary revisions to the adapted test can be made.

ITC:

Explanation. *Prior to initiating any large-scale test scores, reliability and validity studies and/or norming studies that may be time-consuming and expensive, it is important to have confirming evidence about the psychometric quality of the adapted test. There are many psychometric analyses that can be carried out to provide initial evidence of score reliability and validity. For example, at the test development stage, an item analysis using at least a modest sample size (e.g., 100) can provide much-needed data about the functioning of particular test items. Items, which are very easy or hard by comparison to other items, or showing low or even negative discriminating powers, can be reviewed for possible item flaws. With multiple-choice items, it would be appropriate to investigate the effectiveness of item distractors. Problems can be spotted and revisions made. Also, with the same data compiled for item analysis, coefficient alpha or coefficient omega (McDonald, 1999) provide the test developer with valuable information that could be used to support decisions about the appropriate length of the source and target language versions of the test.*

Suggestions for practice. There are a number of basic analyses that can be carried out:

> *Carry out a classical item analysis study to obtain information about item level means and item discrimination indices, and, with multiple-choice items or similar selection items, carry out a distractor analysis, too.*
>
> *Carry out a reliability analysis (e.g., KR-20 with dichotomously scored items, or coefficient alpha or coefficient omega with polytomously scored items).*
>
> *As necessary, carry out a study or two to gain insight about the validity of the adapted test. For example, suppose the adapted test is to be administered via a computer. It may be desirable to carry out a study to evaluate the mode of test administration (i.e., paper and-pencil versus computer-administered). Suppose that the instructions require respondents to answer all questions. It may be necessary to do some research to determine the best instructions for accomplishing this goal. Researchers have found that it is surprisingly difficult to get some respondents to answer every question, if guessing is being encouraged when respondents do not*

Neuropsychological Application

Explanation: This ITC Guideline addresses data-based piloting for item analysis, reliability assessment, and small-scale validity studies**.** Before embarking on large-scale data collection, particularly normative data collection, it is important to run smaller, qualitative pilot studies to address comprehension and both item and method suitability with both test-takers and test administrators. Early qualitative piloting can also give indications of probable validity and reliability. Multiple pilot and revision phases may be necessary for a successful adaptation process. After early qualitative piloting, further quantitative piloting is appropriate. Results from the piloting phases may provide evidence on whether it is appropriate to abandon the test adaptation (i.e., when there is evidence that the test is not reliable or not working in the way anticipated).

Suggestions for practice: The ITC Guideline refers to item analysis with ‘modest’ sample sizes of 100. Whilst this may be feasible for group-administered tests or questionnaires that can be completed quickly, it is unlikely to be feasible for many individually-administered tests that require skilled administrators, particularly in low-resource settings. For qualitative pilot studies designed to check issues such as whether administration and scoring instructions are clear and understandable, or whether the test is acceptable to test-takers, smaller sample sizes are likely to be sufficient. A 'modest' quantitative pilot sample size of 100 would be aspirational for many sparsely-resourced neuropsychological test development projects.

- *Checking method and item suitability with qualitative piloting.* Early piloting that focuses on method and item suitability should include qualitative feedback from both test-takers and test administrators.Early qualitative feedback can address not only item suitability but also method suitability. For example, oral spelling and oral backward spelling are common neuropsychological tasks in English. Oral spelling is highly practised in English language education, to the point of national competitions (spelling bees). This is not the case for languages with transparent orthographies (highly regular spelling) or non-phonetic writing systems (Chinese). So oral spelling may be very unfamiliar or impossible in those languages and may not measure the intended language and memory functions.

  Piloting procedures also provide an opportunity to check for item difficulty, to see whether this is consistent with what is expected (see C-2). For tasks with multiple items that are ordered in terms of difficulty, it may be necessary to change the order of items, something that is particularly relevant if there is a discontinue rule. For example, the abacus is a very difficult item on the English Boston Naming Test (Kaplan, Goodglass, & Weintrab, 1983) but a very easy item when administered to Mandarin speakers. Furthermore, all responses should be recorded verbatim (written, audio, and/or video as feasible and necessary) for post hoc analysis of whether there are systematic factors leading to errors. The piloting stage should include an extensive collection of qualitative information, not only errors made but also the nature of the test-taking and test-administering experience. This can help to identify misunderstandings, poorly functioning instructions and items, and the overall perception and reception of the test.

  For an example of piloting techniques, see the procedures outlined in the project, Developing the Canadian Indigenous Cognitive Assessment for Use with Indigenous Older Anishinaabe Adults (Jacklin et al., 2020). Similarly, Franzen and colleagues (2022) have demonstrated late-piloting techniques for determining the feasibility of new measures in applied clinical contexts.

- *Piloting populations.* The sample for the pilot should be representative of the target standardisation/normative population (i.e., if the test is going to be used with the general population it should not be piloted only on college students). Ideally, the test should be piloted on the full range of the target population. This should take into account all important factors likely to characterise the target population and that may be responsible for variations in response to the test such as the full range of levels of education, the full age range, and, when relevant, the full range of language variations and dialects, people who are bilingual and likely to be tested in a second language, the full geographic range (especially rural versus urban), representatives of different cultural groups, and those with sensory and motor impairments. At this early stage of piloting, it may be prudent to oversample from the extremes of the distribution of the target population to have the best chance of encountering and fixing problems.

Pilot test administrators should also be chosen from the full range of anticipated administrators, particularly concerning professional training (e.g., psychologists, nurses, physicians, teachers, paraprofessionals), dialects, and language fluency. If it is anticipated that administrators will be able to learn to administer the test only from the manual, then some piloting should include administrators who receive no more instruction than the manual. If a test is designed to be administered using an interpreter, then some piloting should be done with representative interpreters.

- *Quantitative data piloting.* After initial qualitative piloting, changes may be required to test instructions or items. Whilst very minor changes may not require additional piloting, if changes are substantial then further piloting is advisable before moving to larger-scale formal data piloting, validation studies, or normative data collection.

  As noted in the original ITC Guidance above, the further stage of quantitative data piloting can help to identify problematic items and procedures. These can then be adapted, replaced, or abandoned. It is to be expected that any test is likely to have a proportion of translated/adapted items that do not function well. Any large project involving a neuropsychological battery is likely to have a few procedures that do not translate or function well. It may be prudent to begin projects with a larger pool of items and procedures (subtests) than the final target, with the expectation that some will be pruned away through the development process.

Confirmation Guidelines

C-1 (9) Select sample with characteristics that are relevant for the intended use of the test and of sufficient size and relevance for the empirical analyses.

ITC:

Explanation*. The data collection design refers to the way that the data are collected to establish norms (if needed) and equivalence among the language versions of a test, and to conduct validity and reliability studies, and DIF studies. A first requirement with respect to the data collection is that samples should be sufficiently large to allow for the availability of stable statistical information. Though this requirement holds for any type of research, it is particularly relevant in the context of a test adaptation validation study because the statistical techniques needed to establish test and item equivalence (e.g., confirmatory factor analysis, IRT approaches to the identification of potentially biassed test items) can most meaningfully be applied with samples large enough to reliably estimate model parameters (the recommended sample size will depend on model complexity and data quality).*

Also, the sample for a full-scale validity study should be representative of the intended population for the test. We draw attention to the important paper by van de Vijver and Tanzer (1997), and the methodological contributions found in van de Vijver and Leung (1997), Hambleton, Merenda, and Spielberger (2005), Byrne (2008), and Byrne and van de Vijver (2014), to guide the selection of appropriate statistical designs and analyses. Sireci (1997) provided a discussion of the problems and issues in linking multi-language tests to a common scale.

Sometimes, in practice, the intended population for the target language version of a test may score much lower or higher, and/or be more or less homogeneous than the source language group. This creates major problems for certain methods of analyses, such as reliability and validity studies. One solution is to choose a subsample of the source language group to match the target language group sample. With matched samples, any differences in the results for the matched samples that may be due to differences in the shapes of the distributions in the two groups can be eliminated (see Sireci & Wells, 2010). For example, comparisons of test structure typically involve covariances, and these will vary as a function of the score distributions. By using matched samples, whatever role the distribution of scores might play in the results is matched in the two samples, and so the role of score distributions on the results can be ruled out as an explanation for any differences in the results.

Perhaps one more example might help to explain the problem of different score distributions in the source and target language groups. Suppose the test score reliability is .80 in the source language group, but only .60 in the target language group. The difference might appear worrisome and raise questions about the suitability of the target language version of the test. However, it is often overlooked that reliability is a joint characteristic of the test and the population (McDonald, 1999) – because it depends on both the true score variance (population characteristic) and error variance (test characteristic). Therefore, the same error variance can lead to a higher reliability simply due to the larger true score variance in the source language group. McDonald (1999) shows that the Standard Error of measurement (which is the square root of error variance) is in fact a more appropriate quantity to compare between samples, not reliability. Another alternative using reliability coefficients would be to draw a matched sample of candidates from the source language group and recalculate the test score reliability.

ITC (continued):

Modern approaches to testing measurement invariance using multiple-group Confirmatory Factor Analysis (CFA) allow for samples with different distributions of the latent traits to be assessed. In such models, while measurement parameters such as item factor loadings and intercepts are assumed equal across groups, the means, variances and covariances of the latent traits are allowed to vary across groups. This allows for the use of full samples, and accommodates the more realistic scenario of different distributions of measured traits across different populations.

***Suggestions for practice**. In nearly all research, there are two suggestions that are made when describing the sample(s):*

> *Collect as large a sample as reasonable given that studies to identify potentially biassed test items require a minimum 200 persons per version of the test (Mazor, Clauser & Hambleton, 1992; Subok, 2017). To undertake item response theory analyses and model fit investigations a sample of at least 500 respondents is required (Hulin, Lissak & Drasgow, 1982; Hambleton, Swaminathan & Rogers, 1991), while studies to investigate the factorial structure of a test require fairly large sample sizes, perhaps 300 or more respondents (Wolf, Harrington, Clark & Miller, 2013). Clearly, analyses with smaller samples are possible, too – but the first rule is to generate large participating samples whenever possible.*
>
> *Choose representative samples of respondents whenever possible. Generalisations of findings from non-representative samples of respondents are limited. To eliminate differences in the results due to methodological factors such as variations in score distributions, drawing a sample from the source language group to match the target language group is often a good idea. Comparisons of standard errors of measurement may be more appropriate.*

Neuropsychology Application

Explanation: A successful adaptation process should include a strategic plan to recruit participants for validation studies that are representative of the population with whom the test is intended to be used. It is important to carefully consider key elements that define the intended population that might impact neuropsychological test performance in addition to language and ensure that these characteristics are present in the sample of participants included in the validation sample. Neuropsychological tests are used in many contexts beyond healthcare settings (e.g., schools, capacity determinations, parenting, criminal, and vocational settings) and it is also important to consider any characteristics likely to be present in these groups that should be present in validation study participants.

There is often great cultural heterogeneity within members of a given language group, particularly for languages that are spoken in various regions worldwide. Several sociocultural factors have been shown to impact neuropsychological test performance in certain populations, yet it is unknown whether the impact of these factors cuts across most or all populations. Further, research on the identification of which aspects of a population are most important to consider is limited. Bilingualism/multilingualism, socioeconomic status, sex, ethnicity, nationality, quantity or quality of education, culture/acculturation, and familiarity with tests have been some of the factors more consistently examined.

Given that most neuropsychological tests require the direct interaction of the examinee with an examiner in the collection of data, it is important to consider factors related to the examiner, and the examiner-examinee relationship when conducting studies aimed at empirically analyzing neuropsychological tests. Factors such as stereotype threat and examiner bias can impact examinees' neuropsychological test performance (Thames et al. 2013).

Suggestion for Practice:

- Carefully characterise sociocultural aspects of the population with whom the adapted test is intended to be used that might impact neuropsychological test performance and try to ensure that these characteristics are present in the validation and normative samples. It is useful to empirically analyse the impact of these characteristics on neuropsychological test performance as part of a validation study as not all may be relevant in the new linguistic/cultural context.

- When developing studies aimed at empirically validating neuropsychological tests, consider the characteristics of the examiners in the collection of data, and empirically test their impact on neuropsychological test performance.

- It is recognised that the collection of data from very large samples may not be feasible when tests must be individually administered by skilled examiners, particularly in resource-limited settings. This may limit the nature of the statistical analyses that are possible. Analyses that are undertaken should be limited to those that are appropriate for the sample size that can be obtained and limitations of using small samples must be clearly documented.

C-2 (10) Provide relevant statistical evidence about the construct equivalence, method equivalence, and item equivalence for all intended populations.

ITC:

***Explanation**. Establishing the construct equivalence of the source and target language versions of a test is important, but it is not the only important empirical analysis to carry out. Also, approaches for construct equivalence (PC-2) and method equivalence (PC-3) were addressed briefly earlier in the guidelines.*

Researchers need to address the equivalence at the item level as well. Item equivalence is studied under the title, "differential item functioning (DIF) analysis." In general, DIF exists if two test- takers, from two different (cultural- linguistic) populations, have the same level of the measured trait but have a different response probability on a test item. Overall differences in test performance across the groups could possibly occur, but this does not present a problem by itself. Whereas, when the members of the populations are matched on the construct measured by the test (typically a total test score, or total test score minus the score for the item being studied), and performance differences exist on the item across the groups, DIF is present in the item. This type of analysis is performed for each item in the test. Later, an attempt is made to understand the reasons for the DIF in the items, and, based on this judgemental review, some items may be identified as flawed, and altered or removed completely from the test.

Two important potential sources of DIF to evaluate are translation problems and cultural differences. More specifically, DIF may be due to (1) translation non-equivalence that occurs from source to target language versions of the test such as familiarity with the vocabulary used, change in item difficulty, change in equivalence of the meaning, etc., and (2) cultural contextual differences (Scheuneman & Grima, 1997; van de Vijver & Tanzer, 1997; Ercikan, 1998, 2002; Allalouf, Hambleton, & Sireci, 1999; Sireci & Berberoğlu, 2000; Ercikan, et al., 2004; Li, Cohen, & Ibera, 2004; Park, Pearson & Reckase, 2005; and Ercikan, Simon, & Oliveri, 2013).

During translation, there is the possibility of using less common vocabulary in the target language. The meanings could be the same in the translated versions, but, in one culture, the word could be more common compared to the other. It is also possible to change the difficulty level of the item as a result of translation due to sentence length, sentence complexity, and use of easy or difficult vocabulary as well. Meaning may also change in the target language with deletion of some parts of the sentences, inaccurate translations, having more than one meaning in the vocabulary used in target language, non-equivalent impressions of the meanings of some words across the cultures, etc. Above all, cultural differences might cause the items to function differently across the languages. For example, words like "hamburger" or "cash register" may not be understood or have a different meaning in two cultures.

There are at least four groups of analyses to check if items are functioning differently across the language and/or cultural groups. These are (a) IRT-based procedures (see, e.g., Ellis, 1989; Thissen, Steinberg, & Wainer, 1988; 1993; Ellis & Kimmel, 1992), (b) Mantel-Haenszel (MH) procedure and extensions (see, e.g., Dorans & Holland, 1993; Hambleton, Clauser, Mazor, & Jones, 1993; Holland & Wainer, 1993; Sireci & Allalouf, 2003), (c) logistic regression (LR) procedures (Swaminathan & Rogers, 1990; Rogers & Swaminathan, 1993) and (d) restricted factor analysis (RFA) procedures (Oort

ITC (continued):

In the IRT-based approaches, test-takers across two languages are matched based on the latent trait scores. In MH and LR methodologies, the observed or estimated test score is used as the matching criterion prior to comparing item performance of respondents in the two groups. Although the sum score is most popular matching criterion in these procedures, other estimated scores, for instance from factor analysis can also be used. These scores are also iteratively "purified" by deleting the questionable items. The matching criterion should be valid and reliable enough to evaluate the DIF properly. In RFA, each item is regressed on the grouping variable (potential violator) as well as the latent trait. Each item loading is set free and the fit to the model is evaluated with reference to the null model where no item is loaded on the grouping variable (no DIF model). If the model provides significantly better fit, this flags the item as DIF.

When a test is dimensionally complex, finding an appropriate matching criterion is an issue (Clauser, Nungester, Mazor & Ripkey, 1996). Using multivariate matching criteria, such as different factor scores obtained as a result of factor analysis, might change the item level DIF interpretations as well. Accordingly, this guideline suggests that, if the test is multidimensional, the researchers might use various criteria to flag the items as DIF, and evaluate the items which are consistently flagged as DIF with respect to various matching criteria. Multivariate matching can reduce the number of items exhibiting DIF across the language and cultural groups.

These methodologies could require different sample sizes. MH, LR, and RFA are models that may reliably and validly work for relatively small samples compared to IRT-based techniques, which require larger samples for valid parameter estimations. Another consideration is the type of item response data. MH, LR, and RFA can be applied to binary - scored data. Other approaches, such as the generalised MH, are needed with polytomous response data.

This guideline requires researchers to locate possible sources of method bias in the adapted test. Sources of method bias include (1) the different levels of test motivation of participants, (2) differential experience on the part of respondents with psychological tests, (3) more speediness of the test in one language group than the other, (4) differential familiarity with the response format across language groups, and (5) heterogeneity of response style, etc. Biases in responses have been, for example, a major concern in interpreting PISA results, and have received some research attention.

Finally, yet importantly, this guideline will require researchers to address construct equivalence. There are at least four statistical approaches for assessing construct equivalence across source and target language versions of a test: Exploratory factor analysis (EFA), confirmatory factor analysis (CFA), multidimensional scaling (MDS), and comparison of nomological networks (Sireci, Patsula, & Hambleton, 2005).

According to van de Vijver and Poortinga (1991), factor analysis (both EFA and CFA) are the most frequently used statistical technique to assess whether a construct in one culture is found in the same form and frequency in another culture. This statement from 1991 remains true today, though the statistical modelling approaches have advanced considerably (see, for example, Hambleton & Lee, 2013, Byrne, 2008). Since, with EFA, it is difficult to compare separate factor structures, and there are no commonly agreed-upon rules for deciding when the structures can be considered equivalent, statistical approaches such as CFA (see, for example, Byrne, 2001, 2003, 2006, 2008) and weighted multidimensional scaling (WMDS) are more desirable as they can simultaneously accommodate multiple groups (Sireci, Harter, Yang, & Bhola, 2003).

(continued)

ITC (continued):

There have been many studies in which CFA was used to evaluate whether the factor structure of an original version of a test was consistent across its adapted versions (e.g., Byrne & van de Vijver, 2014). CFA is attractive for evaluating structural equivalence across adapted tests because it can handle multiple groups simultaneously, statistical tests of model fit are available, and descriptive indices of model fit are provided (Sireci, Patsula, & Hambleton, 2005). The capability to handle multiple groups is especially important as it is becoming common to adapt tests into many languages (e.g., some intelligence measures are now translated/adapted into over one hundred languages, and, in TIMSS and OECD/PISA, tests are adapted into over 30 languages). As the strict requirement of zero cross-loadings in CFA, however, often does not fit well the data on complex multidimensional instruments, Exploratory Structural Equation Modelling (ESEM) is becoming more and more popular, especially with personality data or more complex and inter-related variables (Asparouhov & Muthén, 2009).

WMDS is another attractive approach for evaluating construct equivalence across different language versions of an assessment. Like EFA, WMDS analysis does not require specifying test structure a priori, and, like CFA, it allows for the analysis of multiple groups (e.g., Sireci, et al., 2003).

Van de Vijver and Tanzer (1997) have suggested that cross-cultural researchers should examine the reliability of each cultural version of the test of interest and search for both convergent and discriminant validity evidence in each cultural group. These studies may often be more practical than studies of test structure that require very substantial sample sizes.

It must be recognized, however, that comparison of test-taker performance across two language versions of a test is not always the goal of translating/adapting a test. Perhaps, for example, the goal is simply to be able to assess test-takers in a different language group on a construct. In this instance, careful examination of the validity of the test in the second language group is essential, but research to find evidence of the equivalence of the two forms is not so critical. The importance of this guideline will depend on the intended purpose or purposes of the test in the second language (i.e., target language group). Tests like those used in PISA or TIMSS require evidence of high content overlap because the results are used to compare the achievement of students in many countries. The use of a depression inventory translated from English into Chinese for researchers to study depression or for counsellors to assess depression of their clients would not require high overlap in content. Instead, validity to support the depression inventory in China would be needed.

This guideline can also be addressed with statistical methods after the test has been adapted. For example, if cultural groups are thought to differ on important variables irrelevant to the construct measured, comprehensive designs and statistical analyses can be used to control for these 'nuisance' variables. Analysis of covariance, randomised-block designs, and other statistical techniques (regression analysis, partial correlation, etc.) can be used to control the effects of unwanted sources of variation among the groups.

(continued)

ITC (continued):

Suggestions for practice. *This is a very important guideline and there are many analyses that might be carried out. For equivalence analyses, we offer the following suggestions for practice:*

> *If sample sizes are sufficient, carry out a comparative study of the construct equivalence of the source and target language versions of the test. There are lots of software packages to facilitate these analyses (see Byrne, 2006).*
>
> *Carry out exploratory (preferably rotating to a target structure – so-called "target rotation") or confirmatory factor analysis, and/or weighted multidimensional scaling analysis, to determine the level of agreement in the structure of the test of interest across language and/or cultural groups. The requirement of large sample sizes (10 persons per variable) makes these studies difficult to carry out in many cross-cultural studies. An excellent model for a study of this type is Byrne and van de Vijver (2014).*
>
> *Look for evidence of convergent and discriminant validity (essentially, look for correlational evidence among a set of constructs and check the stability of these correlations across language and/or cultural groups) (see van de Vijver & Tanzer, 1997).*

For DIF analyses, some suggestions are identified below. For more sophisticated approaches, researchers are encouraged to read the professional literature on DIF:

> *Carry out a DIF analysis using one of the standard procedures (if items are binary scored, the Mantel-Haenszel procedure may be the most straightforward; if items are polytomously scored, the generalised Mantel-Haenszel procedure is an option). Other more cumbersome solutions include IRT-based approaches. If sample sizes are more modest, a "delta plot" can reveal potentially flawed items. Conditional comparisons are another possibility (for a comparison of results with small sample methods, see, for example, Muñiz, Hambleton, & Xing, 2001).*

Neuropsychological Application

Explanation. There are several aspects of this guideline and its explanation that are particularly relevant to the translation and adaptation of neuropsychological tests.

- *Functionality in the new language and/or cultural group*. The guideline notes that it is important to recognise that the comparison of test-takers performance across two-language versions of a test is not typically the purpose of translating/adapting a test. The key question is whether the test is valid for its intended purpose in the new language and/or cultural group. For example, if a memory test is used as part of a diagnostic assessment for dementia, the most important question is whether the test has good diagnostic accuracy in the new language and/or cultural group, and not whether participants in the original language and/or cultural group perform differently to those in the new language and/or cultural group. If cross-cultural differences in a particular cognitive construct are being studied, then construct, method, and item equivalence are critical. In clinical neuropsychology settings, the purpose of translating/adapting a test is usually to assess cognitive strengths and weaknesses (impairment) as part of a diagnostic assessment or planning of rehabilitation in the new language and/or cultural group. In this case, the usefulness of the translated/adapted test should be measured by how well it serves its intended purpose in the new language and/or cultural group.

- *Construct equivalence*. It is also important to evaluate whether the test is measuring the same construct in the new language and/or cultural group. This is particularly relevant to measures with multiple factors, including questionnaires and measures of complex constructs such as intelligence (see PC-2). For example, the Wechsler Adult Intelligence Scale (WAIS) (Wechsler, 2008) is the most widely used measure of intellectual ability in neuropsychological settings, and analysis of the performance of the standardisation samples of its most recent editions (WAIS-III and WAIS-IV) has indicated that its 10 subtests contribute to a four-factor structure, comprising: Perceptual Reasoning (PR); Verbal Comprehension (VC); Working Memory (WM); and Processing Speed (PS). For translations/adaptations of the WAIS-IV, it is necessary to examine whether this same factor structure is present to ensure that the test battery is measuring the same constructs in the different languages and/or cultural groups and to ensure it is appropriate to calculate and report the four index scores. For example, Cockroft et al. (2015) compared a multilingual, low socio-economic group of black South African students (SA) with a predominantly white, British, monolingual, higher socio-economic group (UK) and found that sub-test scores loaded differently for the UK and SA groups. Interestingly, exploratory factor analysis indicated that a four-factor structure was the best fit for the SA data (though Arithmetic loaded on the verbal comprehension factor more than the working memory factor), but a three-factor structure was a better fit for the UK data (though interestingly the same four-factor structure that is present in the US standardisation sample was also present in the sample used for the UK WAIS-IV standardisation sample). Another recent example is the work of Staios and colleagues (in preparation), who conducted a confirmatory factor analysis of normative data from an elderly sample of Greek Australians on the Greek adaptation of the WAIS-IV and showed a good fit for the same four-factor solution as the original version.

- *Test validity*. This can also be examined by looking at convergent and divergent validity. That is, does the test correlate with other tests considered to measure the same construct and not correlate with measures considered to measure other constructs in both the original language and/or cultural group and the new language and/or cultural group? This can be challenging to do if there are few tests available in the new language and/or cultural group.

- *Item equivalence*. A broad construct measured by the original test may be equivalent in the new language and/or cultural group but some or all of the items in the test may not be equivalent in both original and target languages. This is particularly relevant in neuropsychological tests that have items organised by increasing difficulty, with discontinue rules operating. In this case, Differential Item Functioning analysis can be carried out to examine whether items function differently in different samples.

  For some tests, it can be quickly identified that whilst the construct being examined is equivalent in the two languages and/or cultural groups, some or all of the items of the original test will not be useful in the new language and/or cultural group. An example of this is object naming. Object naming is often examined as part of a neuropsychological assessment, and it is reasonable to assume that the ability to name objects is a universal cognitive function and that difficulties in word finding may arise because of neurological injury or disease in all language and/or cultural groups. Most available naming tests, including the naming tests included in many cognitive screening tests, involve naming black and white line drawings of objects. The use of black and white line drawings may introduce some method bias as some cultures may be less familiar with line drawing representations of objects. A major issue in the translation and adaptation of naming tests is the potential lack of familiarity with the included objects that may be common in one language and/or cultural group but unfamiliar or virtually unknown in another. This may increase the difficulty of the naming test and compromise the validity of the test in the new language and/or cultural group. Therefore, it is important to assess item equivalence between the original and new language and/or cultural group for all items, assessing both adopted words and object exemplars and if necessary, replacing items. If a pool of items is available, one approach to selecting items for inclusion in the test is to use item discrimination analysis, using a mixed sample of healthy and clinical participants. Although some attempts have been made to develop 'universal' naming tests (e.g., Ardila, 2007), most naming tests have not been designed for cross-linguistic/cross-cultural purposes but include objects that are common in the language and/or cultural group in which the test was developed.

When selecting new items for a test it is important to consider all item features that were relevant to the selection of the original items, whilst ensuring that the original features of the items are relevant in the new context. For example, the Addenbrookes Cognitive Examination III (Bak & Mioshi, 2007) includes a test of reading irregularly spelled words (sew, pint, soot, dough, and height). Difficulty with reading irregular words is referred to as surface dyslexia. It would be pointless to simply translate these items to another language if they did not have irregular spellings in the new language. Furthermore, some languages have few, or no, irregular spellings of words and it is necessary then to consider how the language deficit that produces surface dyslexia manifests itself in the new context and select test items accordingly.

Suggestions for practice. Exploratory factor analysis (EFA), confirmatory factor analysis (CFA), multidimensional scaling (MDS), and other advanced statistical techniques may not be familiar to all neuropsychologists and other test users. Documentation of test features in manuals and other publications should include clear explanations of the implications of these analyses. This is particularly important when results differ from those of the source test so that test users do not simply assume that the test and its interpretations and uses are equivalent in the new context. Test developers should be particularly careful to distinguish if a test can be considered equivalent for purposes of comparisons across versions and populations. They should also be particularly careful to specify the uses and validities that have and have not been demonstrated in the new language and/or cultural group.

C-3 (11) Provide evidence supporting the norms, reliability, and validity of the adapted version of the test in the intended
populations.

ITC:

Explanation*. The norms, validity evidence, and reliability evidence of a test in its source language version do not automatically apply to other possible adaptations of the test into different cultures and languages. Therefore, empirical validity and reliability evidence of any new versions developed must also be presented. All kinds of empirical evidence supporting the inferences made from the test should be included in the test manual. Special attention should be paid to the five sources of validity evidence based on: test content, response processes, internal structure, relations to other variables, and consequences of testing (AERA, APA, NCME, 2014). Exploratory and confirmatory factor analysis, structural equation modelling, and multitrait- multimethod analyses are some of the statistical techniques that can be used to obtain and analyse data addressing validity evidence based on internal structure.*

Suggestions for practice*. The suggestions are the same as would be required for any test that is being considered for use:*

If the norms developed for the original version of the test are suggested to be used with the adapted version, evidence should be provided that this use is statistically appropriate and fair. If no evidence can be provided for such use of the original norms, specific norms should be developed for the adapted version according to the standards for norm development.

Compile a sufficient amount of reliability evidence to justify the use of the target language version of the test. The evidence might normally include an estimate of internal consistency (e.g., KR-20, or coefficients alpha or omega).

Compile as much validity evidence as is needed to determine whether the target language version of the test should be used. The type of evidence compiled would depend on the intended use of the

Neuropsychological Application

Explanation. Normative data is a key component of the usefulness of neuropsychological tests. We cannot assume that because a test is reliable and valid in the context for which it was developed a translated and adapted version will be equally reliable and valid in the context in which it is going to be used.

- *Reliability.* The reliability of the adapted test should be examined in the context in which it is intended to be used. There is a range of different forms of reliability and those that are relevant and feasible to test should be examined. This may include measures of internal consistency, test-retest, inter-rater, and parallel-version reliability.

- *Practice and Retest Effects.* Concerns and proofs of test reliability do not greatly differ between neuropsychological and other tests. However, in establishing test-retest reliability, test developers may encounter significant practice effects, which are distinct from reliability. Practice effects refer to improvement in test performance upon repeat testing with the same or a similar test or alternate version. Practice effects have been demonstrated across different neuropsychological measures, particularly for tests of memory, executive functions (Basso et al., 2002; Lemay et al., 2004), and information processing speed (Levine et al., 2004; Register-Mihalik et al., 2012). In the clinical context, practice effects may complicate data interpretation during serial assessment with the same test (Calamia, et al., 2012). Alternate versions of a test may help to mitigate any potential influence of practice effects during serial assessment. However, practice effects may still be apparent even when using alternate versions of the test in clinical practice (Calamia, et al., 2012). Attention, naming, and visuoperceptual / recognition tasks are less susceptible to such effects (Zgaljardic & Benedict, 2001).

In the context of adapting neuropsychological tests, practice effects may impact data interpretation, and contribute to bias in the psychometric properties of the adapted version. Similar to the clinical context, practice effects during test adaptation may be partly due to familiarity with the type of task, familiarity with specific test items (Zgaljardic & Benedict, 2001), and memory of the specific test stimuli (Benedict, & Zgaljardic, 1998). Healthy participants show greater practice effects than those with acquired, degenerative, and psychiatric conditions because they are usually able to retain the procedural knowledge of testing, including test-taking strategies over extended periods (Calamia, et al., 2012). With alternate versions as well, the strategies learned may affect performance and lead to practice effects. Depending on the nature of the test and the test stimuli, the time interval over which practice effects are found may vary (Calamia, et al., 2012; Rapport, et al., 1997). Therefore, the inter-assessment time intervals for test-retest validation of adapted versions should be carefully chosen to be relevant to anticipated test use so that significant practice effects can be properly understood and interpreted. In some circumstances, it may be advisable to test multiple inter-assessment time intervals to estimate the size and time course of practice effects. Practice effects should also be documented for retesting with alternate versions of a test.

Target population characteristics are another factor that contributes to practice effects during repeated neuropsychological testing. In general, those with higher IQs, higher education, and also younger participants tend to show greater practice effects (Calamia, et al., 2012). One may suspect that practice effects are unlikely to occur in test-naïve individuals. On the contrary, practice effects may become apparent if the novelty of ‘testing’ is diminished in this population. This may not only be due to familiarity with the test material, but also due to comfort with testing (“test sophistication”), familiarity and rapport with the examiner, and familiarity with the test setting. For example, some of us have noted that test-naïve test-takers often improve and perform faster on the later, supposedly more difficult conditions of attention tests such as the Coloured Trails Test (D’Elia, et al., 1996) and the Five Digit Test (Sedó, 2007). Most research on practice effects has been carried out on populations who have some familiarity with testing. Because of the considerations just mentioned, it remains an open question whether practice effects will be greater, lesser, or the same in test-naïve subjects relative to available research with test-familiar subjects. This may also vary with the specific populations and the nature of the test materials. Consequently, test-retest reliability and practice effects need to be examined for translations and adaptations of tests and with the intended populations.

Practice effects that are due to remembering the content of a memory test or the solution to a problem-solving task are likely to be relatively short, lasting weeks or months. These are likely to be seen only when retesting with the same test. Practice effects due to developing familiarity with testing or strategies for problem-solving may be more long-lasting. These types of effects may also be seen when retesting with alternate versions of a test. It may be possible to separate content effects from familiarity and strategy effects with alternate versions of tests. For example, the developers of the Brain Injury Rehabilitation Trust (BIRT) Memory and Information Processing Battery (BMIPB) investigated test-retest reliability both with the same version and with alternate versions (Disabilities Trust, 2022). Improvements seen between alternate versions may be due to familiarity and comfort effects. Improvements that go beyond that when seen in repeat testing with the same version may reflect memory for the content of the version.

One strategy for demonstrating the validity of a translated test is to administer both language versions to bilinguals. Even here, practice effects can be seen and need to be taken into account (Ibrahim, et al., 2015).

Suggestions for practice:

- Ways to address practice effects.
    - Check whether empirical data are available indicating any potential influence of practice effects on the original test in question or on its adapted versions.
    - Check whether empirical data are available regarding the optimal test-retest and inter-assessment interval that will help to minimise any potential practice effects. If appropriate, it would be best to use the same time interval used for the original test.
    - The purpose for which the test is being used (for example, to monitor change over time in degenerative conditions; pre- and post-surgery for epilepsy or tumours; baseline testing for athletes and soldiers at risk for concussion) needs to be taken into consideration for determining the time interval.
    - Determine whether the sample in whom the test will be conducted, or the alternate versions of a test will be administered is likely to be subject to practice effects.
    - Determine if any particular test items contribute to practice effects.
    - Use appropriate statistical measures to determine whether test-retest and performance on alternate versions have been confounded by practice effects.
    - Determine the appropriate statistical measures to estimate practice effects.
    - Use a counterbalanced design to reduce practice effects for tests with alternate versions.
    - Report the magnitude of practice effects and the optimal time for retesting (either with the same version or with the alternate versions). Such information proves useful to determine the minimal interval time for retesting in clinical situations where serial assessment is necessary either with the same or alternate versions.
    - If there is reason to suspect that there may be variability in practice effects within a population because of effects of education, age, native language, acculturation, presence of a neuropsychological disorder, or other variables, design the test-retest reliability studies to be able to check for such variation.
    - If using cross-language test-retest to validate a test translation/adaptation, counterbalance the order of administration and take practice effects into account in the data analysis (Ibrahim, et al., 2015).

    If exploratory research finds that all or some parts of the test-takers in the target population need substantially more practice items to comprehend the task, then test developers should reconsider whether the test is truly measuring the same intended construct as in the original language and population.

- *Validity.* Validity is a property of the use of a test, not of the test itself. The primary use of most neuropsychological tests is distinguishing brain impairment from normal brain functioning. This information contributes to diagnosis. A related function is distinguishing functioning in different neuropsychological domains such as attention, memory, language, perception, self-control, emotions, social skills, and so on. Neuropsychological tests are also often used to infer adaptive functioning. Ideally, validity will be demonstrated for each distinctive use. So primary validity will demonstrate the ability to distinguish healthy from unhealthy brains. Further validity will demonstrate the ability to distinguish one cognitive impairment such as memory from another such as attention. Each major practical application will also be validated, such as the ability to predict school or work performance. For example, a word list memory test might be excellent at discriminating Alzheimer's disease from normal and very good at discriminating memory impairment from visual-spatial impairment but only fair at distinguishing memory impairment from language impairment, marginal at distinguishing money management capacity, and poor at determining capacity to drive a vehicle. When this word list memory test is translated and adapted ideally each of the useful validities will be demonstrated in the new context.

- *Norms.* Test interpretation in neuropsychology typically involves a comparison of an individual’s performance to a normative data set. These norms are used to infer the expected performance of the individual. Performances below these expectations are used to infer brain dysfunction. Patterns of dysfunction relative to norm-based expectations are further used to refine diagnoses, treatment strategies, and goals and to predict and explain adaptive behaviour and capacities. We cannot assume that the normative data from one context will be accurate in another. For example, in one unpublished pilot study, a non-verbal reasoning test had to be abandoned as part of an epidemiological battery when none of the pilot study physicians in the target culture were able to achieve an “unimpaired” score relative to the source culture. The use of inappropriate norms is not just inaccurate and wasteful–it can be harmful.

  It is *possible* that the original norms will indeed be accurate in the new context. This is most likely to occur when adaptations to a test are minimal *and* the new context and population are very similar to the original context and population. However, this needs to be determined rather than just assumed. For example, the Wechsler Adult Intelligence Scale-IV (WAIS-IV; Wechsler, 2008) was developed with normative data collected in the US and is also widely used in the UK. The modifications required to adapt the test to make it suitable for use in the UK were relatively modest – most items in the tests remained unchanged but some individual items were changed to make them more familiar to the UK population (e.g., ‘automobile’ was changed to ‘car’; ‘gasoline’ was changed to ‘petrol’). Furthermore, the cultural context of the UK is broadly like that of the US, in terms of language, education, socioeconomic status, etc. Hence it is not unreasonable to assume that the normative data set for the US sample might be comparable to the UK population. However, rather than simply assuming this to be the case, a study was undertaken to examine this hypothesis. A sample of 270 UK participants was recruited to represent the UK population in terms of gender, age, race/ethnicity, geographic region, and educational level. The performance of this group was similar to the original US normative sample in terms of item difficulty, average scores, and reliability. Furthermore, confirmatory factor analysis showed that a similar factor structure emerged. Therefore, in this case, it is considered appropriate to use the adapted test and also to use the original normative data for the interpretation of individual performance. Even though a sample of 270 was collected in the UK, the normative sample from the US is much larger providing a more precise estimate of performance. This approach demonstrates that the UK version performs similarly to the original US version, though does not directly confirm that it is a valid tool for, for example, determining the impact of a neurodegenerative neurological condition on cognitive test performance (though this might be inferred from the psychometric properties that have been evaluated).

The WAIS-IV therefore exemplifies a test that has been adapted for use in a new context, but original norms could be used. However, as noted earlier, the cultural differences between the United States (US) and the United Kingdom (UK) are much smaller than would be the case between the US and a country with a different language, education, socioeconomic, contexts, etc. In many cases, it will be clear from the outset that the original norms are unlikely to reflect performance in the new context. In these situations, new normative data must be collected. This can be challenging and expensive, but if the original norms are used there is a very high risk that the interpretation of test performance will not be valid, which could lead to misdiagnosis and inappropriate treatment, or inappropriate interpretations in other settings. New normative data should therefore be collected from a representative sample of the population in which the adapted test will be used.

It not entirely possible to know in advance which normative language/cultural/ national/educational groupings will give the most accurate predictions of performance for any given individual. We have reason to suspect that, for many populations, people from different backgrounds or with different demographic characteristics (e.g., pure versus functional illiteracy, monolinguals versus multilingual, urban versus rural dwelling) may yield differing test norms that could produce significant differences in relevant inferences. When these groups are combined in normative data sets, it may give test users less confidence in the interpretation of their results.

Furthermore, while there has been some tendency to assume that a national identity is an appropriate unit for population norming, this is not necessarily accurate (e.g., Guàrdia-Olmos, et al., 2015). Cultural heterogeneity may challenge the generalizability of the norms developed within a specific cultural context (e.g., Bengali speaking population residing in Kolkata, in Eastern India) compared to another context (e.g., Malayalam speaking population residing in Bangalore, in Southern India) within the same country. For example, for norms for the category fluency test (animals) differed across Bengali (mean age = 66.8 ± 10 years; mean education = 7.7 ± 5.5) and Malayalam (mean age = 66.9 ± 5.6; mean education = 7.2 ± 6.1 years) speaking populations in India (Das et al.,2006; Mathuranath et al., 2003).

Lastly, acculturation, the process of change in an individual because of contact with a new culture, is likely to make a substantial difference in testing results (Tan, Burgess, & Green, 2020). Thus, norms may not generalise within a language or cultural group across levels of acculturation and immigrant generations, including international migration or migration within a country, especially rural-to-urban migration and in diverse normative samples including people who have immigrated from another cultural context.

C-4 (12) Use an appropriate equating design and data analysis procedures when linking score scales from different language versions of a test.

ITC:

Explanation. *When linking two language versions of a test to a single reporting scale, several options are possible. If a common set of items is used, the functioning of these common items across the two language groups should be evaluated, and if differential functioning is observed, their removal from the data used in establishing the link should be considered. Delta plots (Angoff & Modu, 1973) serve this purpose well, and Cook and Schmitt-Cascallar (2005) provided a good illustration of how to use delta plots to identify items that have a different meaning for the two groups of examinees. Not all item types have the same potential to link between language versions. Item difficulty and discrimination parameter estimates derived in the framework of item response theory for the common items can be plotted to help identify inappropriately performing common items (see Hambleton, Swaminathan, & Rogers, 1991).*

But linking (i.e., "equating") scores across two language versions of a test will always be problematic because strong assumptions need to be made about the data. Sometimes, a highly problematic assumption is made that the different language versions of the test are equivalent, and then scores from the two versions of the test are used interchangeably. Such an assumption may have merit with mathematics tests because translation/adaptation is typically straightforward. It may have merit, too, if the two versions of the test have been carefully constructed, and so the assumption can be made that the source language version of the test functions with the source language population in an equivalent way to which the target language version of the test functions in the target language population. This assumption may have merit if all of the other evidence available suggests that the two language versions of the test are equivalent and there are no method biases influencing the scores in the target language version of the test.

Two other solutions exist, but neither is perfect. First, the linking could be done with a subsample of the items that are deemed to be essentially equivalent in the two language versions of the test. For example, the items may be the ones that were judged as very easy to translate/adapt. In principle, the solution could work, but requires the linking items and the remainder of the test items to be measuring the same construct. A second solution involves linking through a sample of test-takers who are bilingual. With this sample taking both versions of the test, it would be possible to establish a score conversion table. The sample could not be too small, and in the design, the order of presentation of the forms of the test would be counterbalanced. The big assumption in this approach is that the candidates are truly bilingual, and so, apart from the relative difficulties of the forms, the candidates should do equally well on both forms. Any difference is used to adjust the scores in converting scores from one version of the test to the other.

Suggestions for practice. *Linking scores across adapted versions of a test is going to be problematic at the best of times because all of the equating designs have at least one major shortcoming. Probably the best strategy is to completely address all of the steps for establishing score equivalence. If the evidence addressing the three questions below is strong, even the scores from the two versions of the test can be treated interchangeably:*

> *Is there evidence that the same construct is being measured in the source and target language versions of the test? Does the construct have the same relationship with other external variables in the new culture?*

(continued)

ITC (continued):

Is there strong evidence that sources of method bias have been eliminated (e.g., no time issues, formats used in the test are equally familiar to candidates, no confusion about the instructions, no systematic misrepresentation in one group or the other, standardised instructions, absence of response styles (extreme ratings, differential motivation...)?

Is the test free of potentially biassed test items? Here, a plot of p values or, better, delta values, from items in the two versions of the test can be very helpful. Points not falling along the linear equating line should be studied to determine if the associated items are equally suitable in both languages. DIF analyses provide even stronger evidence about item equivalence across language and cultural groups.

If linking of scores is attempted, then an appropriate linking design needs to be chosen and implemented. Evidence for the validity of the design should be provided.

Neuropsychological Applications

Explanation. The methodology outlined above for attempting to establish “equivalency” across different language versions of a test is generally comparable to neuropsychological tests. However, it is important to consider that the varied and complex cultural factors that impact neuropsychological test performance make the comparison of scores of different versions of the tests problematic (Casaletto & Heaton, 2017). Furthermore, a great majority of the time the purpose of neuropsychological tests is not to allow direct comparison of test scores across language populations, but rather to serve similar functions in each of the different populations. In this latter case, the concept of whether scores are “equivalent” across language populations is less relevant. Rather, the focus is on whether the test measures a similar construct across language populations and whether the scores derived from a given population have strong psychometric properties within the population of interest, including whether it accurately identifies underlying neurological impairment within a given group.

Suggestions for Practice: Linking scores across adapted versions of a neuropsychological test is typically not the main purpose of a neuropsychological test. Rather, neuropsychological scores are closely tied to the normative sample that they were developed. Thus, caution is warranted when interpreting scores derived from different cultural/language groups.

If linking of scores is attempted in the adaptation of a test, then it would be important to consider and clearly describe why this is a focus, and how this would impact the utilisation of scores in clinical and research settings.

Administration Guidelines

A-1 (13) Prepare administration materials and instructions to minimise any culture- and language-related problems that are caused by administration procedures and response modes that can affect the validity of the inferences drawn from the scores.

ITC:

Explanation. *Implementing the administration guidelines should start from an analysis of all factors that can threaten the validity of test scores in a specific cultural and linguistic context. Experience with the administration of an instrument in a monolingual or monocultural context may already be helpful in anticipating problems that can be expected in a multilingual or multicultural context. For example, experienced test administrators often know which aspects of instruction may be difficult for respondents. These aspects may remain difficult after translation or adaptation. Applications of instruments in a new linguistic or cultural context could also find issues not previously found in monocultural applications.*

Suggestions for practice. *It is important with this guideline to anticipate potential factors that might create problems in test administration. Some of those factors that need to be studied to ensure fairness in test administration are the following:*

> *Clarity of test instructions (including translation of those instructions), the answering mechanism (e.g., the answer sheet), the allowable time (one common source of error is the failure to allow sufficient time for test-takers to finish), motivation for candidates to complete the test, knowledge about the purpose of the test, and how it will be scored.*

Neuropsychological Application

Explanation. The nature and purpose of neuropsychological testing may be a foreign concept for test-takers without prior assessment exposure, which can lead to misunderstanding and potentially inaccurate results. Such misunderstanding may be mistaken for brain impairment, test anxiety, poor test effort, or other concerns. To minimise these effects, neuropsychologists must take into consideration the socio-cultural, demographic, and functional context of the intended population in both the research aspects of the test (e.g., construction, validation, and standardisation) and their clinical applicability (e.g., providing clear and culturally appropriate test description, explanations to client/patients to obtain their consent for testing, and administration procedures). Thus, the true scope of this guideline needs to encompass not just the adaptation of instructions specific to a given test, but rather, the entire testing context. As a general recommendation, before working on the adaptation of tests for culturally-diverse populations ─that may or not differ from the test developers' background–it is always advisable that test developers and researchers self-evaluate their biases and preconceived notions about the population and consider how these may affect their approach to adaptation and validation assessment (i.e., potential test biases).

Suggestions for Practice. While this specific guideline focuses on testing materials and administration considerations, these should be carefully considered and addressed from the early stages of test development, translation, adaptation, and validation. Administration instructions should discourage on-the-spot translation and adaptation of a test for a culture/group for which the test has not been validated due to obvious risk to the accuracy and relevancy of the obtained results. Especially in cultures where test standardisation is not stringent, test inaccuracies and potential cultural biases must be minimised. It is also important to highlight the importance of piloting (see TD-5) to assure test applicability, construct validity (see C-2 and C-4), and ecological validity within the intended cultural context.

The accuracy and validity of tests are highly dependent upon the entire psychosocial context in which they are used. This context reaches well beyond the specific test instructions. It cannot be fully specified and accounted for in a test manual for all contexts of use. Nevertheless, test developers should recognize this, acknowledge it in the manual, and attempt to anticipate the diversity of psychosocial contexts of use as much as is feasible.

- *Orientation to test-takers.* The first suggested consideration is what type and how much information will be provided to the test-taker about the evaluation (e.g., rationale, materials, methods, procedures, and expectation; in some settings called "informed consent"). Parts of this recommendation apply to the standardised instructions (explained in more detail below) and to the application by end-users (e.g., clinicians) to provide a person-centred explanation to their clients/patients. This conversation must be culturally sensitive and lead to 1) a clear understanding of what is expected from the test-taker and 2) a reduction of any potential confounding factors associated with language, power differentials, performance anxiety, etc.

  This explanation may require substantial modifications based on the intended cultures' typical experience with assessment measures, evaluative methods, and the concept of neuropsychology in general. Any present attempt to list all possible cultural and global considerations would be limited; therefore, our recommendation is to consider the intended culture's understanding of the assessment procedures with a macro- to micro-approach. For example, a neuropsychologist may ask several questions in the process: What is the population's understanding of neuropsychological functioning? Are intelligence, cognition, and psychological functioning often evaluated with objective measures within this cultural group? Are individuals from this culture typically evaluated in the academic or clinical setting? What is the intended population's approach to test-taking? How do they best receive instructions and feedback? Information obtained from these analyses may offer important information about how to adapt the informed consent (e.g., how much or how simple the explanations would need to be) and should also help inform the evaluation of adaptations to test instructions, practice items, prompting, and feedback.

  In making cultural adaptations to test explanations, it is important to try to preserve the construct validity of the test. Tests designed to measure executive functions are particularly sensitive to such effects. Many such tests are intended to present moderately-novel tasks, not so unfamiliar as to entirely befuddle the test-taker with understanding the test expectations, but not so familiar as to be able to be approached in a routine and familiar manner and so fail to measure the intended executive functions. Task familiarity may vary across cultures and so the amount of instruction needed may vary to achieve the intended moderate novelty and preserve the construct validity of the measure.

- *Clarification of purpose, roles, and possible misconceptions.* Another recommended consideration is regarding the clarification of a neuropsychologist's role or an assessment's role in the whole care system. In many countries, psychiatrists, paediatricians, or therapists take on multiple responsibilities including testing. Therefore, some test-takers may have greater difficulty understanding and adapting to the specific procedures, goals, and nature of neuropsychological evaluation and the client-provider relationship. It might be worthwhile to provide an extra explanation or distinction of the different roles and distinct types of services to make sure test-takers understand what to expect and what is expected from them. The evaluator may also consider spending time destigmatizing or clarifying misconceptions about mental health and the evaluation of intelligence, cognition, and neurobehavioral functioning. Such clarifications may be beyond the domain of what can be reasonably described and expected in a test manual, but they may nevertheless be essential to accurate testing.

- *Patient-provider relationship.* We cannot emphasise enough how important it is the process of developing a good working relationship (also known as "rapport") and an understanding of how individuals are motivated and comfortable at a social level and identifying cultural factors that may interfere with or affect test performance. For example, in the Hispanic/Latino, because of the value of "familismo" (strong values of dedication, respect, and loyalty to immediate and extended family), the inclusion of family members in the evaluative process will serve to improve the level of trust and comfort of the test-taker with the evaluative processes. This is also an important value in many Asian cultures.

  "Personalism" is another value often characteristic of Hispanic/Latino cultures, but also in some European (e.g., Italy) and African cultures, and it refers to the emphasis on courtesy, respect, politeness, and open communication in social interactions. For individuals who value "personalism," a distant and sterile attitude from the clinician may affect the establishment of rapport and trust and hinder cooperation, response to feedback, and engagement in testing. For some other cultures, who may experience historical mistrust towards healthcare providers, such as Black or African American individuals, spending extra time in the development of rapport may also be necessary. Opportunities for increased rapport with these populations may include more extended assessment interviews, questions that gather information beyond the clinical essentials, cultural humility, and curiosity, and framing of the assessment relationship as collaborative and mutually beneficial. Now, for some cultures that value individualism (e.g., White Americans) or populations where a more respectful stance towards authority figures is emphasised (e.g., China), such approaches may be intrusive, distasteful, or even intimidating. These same cultural extremes can be found in the need for personal space, direct or indirect eye contact, and comfort with environmental elements (e.g., some military individuals/Veterans may be uncomfortable sitting with their back towards the door). The aspect of cultural humility and respect can go a long way to help transgender or gender nonconforming individuals to feel respected and valued within professional relationships. When the provider is from a different cultural background than the client/patient, the acknowledgment of cultural differences and possible effects on the evaluation should also be considered and when considered helpful, discussed with the client. Again, specific procedures for establishing rapport may be beyond the domain of what can be reasonably described and expected in a test manual, but they may nevertheless be essential to accurate testing.

- *Task Explanation.* To attain the most accurate data, the clear presentation and description of the assessment tasks and the explanation of what is needed from the test-taker is the third most important element to assure the best possible results when translating or adapting a test and must be considered early on the adaptation methodology.

- Instructions may need variations to accommodate the culturally accepted healthcare model of the target populations (e.g., directive versus collaborative). For example, whereas in North America an evaluator may obtain the best possible effort from a test-taker by offering directive instructions, such as, "I would like you to draw a clock," some cultures may require more specific instructions or an explanation of the rationale behind the task and how their performance is being evaluated.
- Instructions must be adapted beyond translation, also considering socially accepted commands, requests, etc. In some cultures, offering direct commands may imply a level of power, thereby enhancing the power dynamics in the evaluation environment. In this context, a younger evaluator providing directive requests to an elder patient may be considered disrespectful and potentially threatening for both rapport and effort. On the contrary, some individuals may be inclined to trust their providers as the experts and traditionally do not seek much explanation about procedural considerations. At the level of adequate social context of test instructions, also consider whether variability in the administrator's sociability (e.g., inflection, warmth, comfort, etc.) may affect task engagement, effort, and overall performance on a task. For example, in a culture where mutual respect and collaboration are valued, a direct demand or instruction may be received with resistance and reduced effort, which was also addressed in our applications for establishing pre-assessment rapport. Likewise, many languages express relationships in the second person pronoun and/or verb form that is chosen in ways that English no longer does (but used to with "thee" and "thou"). For example, in Spanish it is appropriate to address a child as "tu," but this would usually be inappropriate for an adult, for whom "usted" should be used in most settings, and possibly "vos" in some settings. Such choices should be appropriately addressed in test instructions, for example, by giving a choice of pronouns in the instructions. It is recommended that the rationale for adapting instructions is clearly documented so evaluators administering the tests understand the rationale and the importance of following specified instructions (see Doc-1 and Doc-2).
- Another important consideration is how much 'clarity' or 'intentionally blurring' (or being less than clear about the purpose of the test) will be provided in the explanation of the test, which may include offering additional information about the goal, objective, and expectations of the task. In a speeded test, for example, instructions may instruct test-takers to "Work as quickly as you can without making mistakes," but these instructions may be contradictory in cultures where working without mistakes is an antonym to working as quickly as one can; again, addressing the speed/accuracy paradigm (see C-2). In that example, one may consider it necessary to clarify that performance is scored by the time used. However, the trade-off of cultural adaptation must prevent inadvertently changing the construct of the test.
- In addition, consider that expressions of the level of effort encouraged in the task instructions (e.g., "make sure you do a good job", "do not make any mistakes") may be worded in a manner that is too strong and anxiety-provoking for some cultures (e.g., a culture highly emphasizing perfectionism and lack of room for failure/mistakes). It may also trigger some shame in individuals who grow up in a culture in which criticising and shaming are socially popular approaches to parenting.
- Some populations, such as those with reduced or affected hearing acuity, may need a visual print of the test's instructions or additional adaptations.

- *Practice items.* After oral instructions are provided, most tests include practice items to ensure understanding and for the test-taker to become comfortable with the task and ask questions if there are any doubts. The number of examples/trials needed may vary from one culture to another mediated by the commonality of and familiarity with the task (e.g., visuospatial puzzles can be a new concept for some cultures), and the socially accepted level of comfort expected for an individual to attempt a task independently (e.g., some cultures emphasise practice until perfection under the guidance of a guide/teacher/leader). Providing extra practice may improve test performance for some tasks in some cultures, but inadequate practice for the culture may result in misunderstanding of the expectations and invalidation of the construct. While changing the amount of practice from one culture to another may alter the standardisation in one sense, this may be needed to preserve the construct and functionality of the test.

  The original test may not include any practice items. When adapting a neuropsychological test for a new population, it needs to be determined whether practice items need to be introduced. In case they are required, the number of practice items that should be included also needs to be determined. Here again, several factors impact the decision, including the type/nature of the test, whether it is a screener, and population characteristics.

  Changes to practice item content and procedures need to be done during the cognitive interviewing phase and should be further examined during the pilot phase. These types of adaptations and their justification should be clearly documented on the test instructions/technical supplements (see Doc-1 and Doc-2). These choices will need to consider the relative priorities of comparisons across cultures versus validity within the culture (see Application Background).

- *Cueing and corrections.* During task administration, test adaptations should ensure the culturally appropriate forms of prompting or feedback to acquire additional information are either part of the standardised test structure (e.g., correction for an error, correct/incorrect feedback, semantic and phonemic cueing that is added to a score) or through qualitative information (e.g., to see whether they can provide additional answers or to redirect attention to the task).

  Tasks that require a correction of an error (e.g., Trail Making Test (Reitan,1955)): “You made a mistake, go back to…”) or direct “wrong/right” feedback (e.g., Wisconsin Card Sorting Test (Heaton, 1981); Test of Memory Malingering (Tombaugh, 1996)) sometimes causes increased levels of anxiety, particularly for shame-based/performance-based cultures. In some tests, the administrators provide “wrong/right” feedback for the first part of the test to ensure the test-taker’s understanding of the rules and then they stop giving feedback in the later part (e.g., Comprehensive Test of Phonological Processing-2 (Wagner, et al., 2013)). It may also induce anxiety in some cultures when the administrators stop providing feedback without any explanation. A way to minimise this impact is through a combination of a clear explanation of the expectations and examples of the type of feedback that will be used and normalisation of errors on the task (e.g., “During this task, I may have to interrupt you, and this is to be expected if you make an error, but it is normal. If that happens, I will say, ...”) and the use of culturally accepted forms of feedback (e.g., instead of right/wrong, saying, “No, that wasn’t it.”).

For tasks that depend on specific forms of cueing, such as the case of naming tasks (requiring phonemic and semantic cueing), using technologies such as standardised recordings would be recommended especially in a multicultural setting. An administrator's accent could be a confounding factor for differentiation diagnoses of language disorders. But accent should not be an exclusive criterion for choosing administrators or an excuse to minimise the validity of the test.

Prompting for additional information, such as asking questions like “what else?” on a memory recall test, may induce additional stress in a culture where providing no more answers would be a sign of disrespect or disinterest. Similarly, offering prompts for an answer on a timed task (e.g., “Do you have an answer?” or “take your best guess”) may place undue stress on individuals who are taught to examine problems carefully and not provide an answer until sure.

- *Time/discontinuation procedures.* When adapting tests, consider the intended culture’s sense of the strictness of the time limit (e.g., in some cultures, when people say “soon” may mean hours) and how the time urgency or time-limited nature of a test is explained to these individuals. People from some cultures may feel that they are being criticised if they are cut off from a task due to a time limit. In such cases, ways to balance assessment time limits and cultural considerations may include a clear explanation of discontinuation procedures, annotating where in the task the test-taker was at the time limit but allowing them to finish all of it, or a test design that allows the test-taker to finish the entire task. All these options should also consider frustration or fatigue elements.

- *Formatting of test answer sheets.* Close consideration should be placed on the design of answer sheets and other test materials. For example, test bubble sheet answers may pose increased difficulty for test-takers with dyslexia. The use of electronic devices (e.g., tablets) may limit access to some populations that are not familiar with or have no access to electronics. For example, Porrselvi and Shankar (2018) found that Tamil speakers in Chennai, India, although well-educated and accustomed to electronic tablets, did worse on drawing on tablets than on paper.

- *Test selection.* It is important to address the limitations that may be present when assessing bilingual or bicultural individuals. There are many important and detailed considerations when working for this population, amongst which the most important include language, dominant culture, and normative data for selected tests. There have been many considered ways to calculate bilingualism and language dominance. Language dominance or bilingualism quotient may be assessed with a qualitative interview that describes age and context of language acquisition, language use in variable contexts, language preference, and language switching and context (De Bruin, 2019). There are other more quantitative ways to assess bilingualism and language dominance including the Language Acculturation Meter and the verbal fluency computation to name a few (Marian & Hayakawa, 2021; Suarez et al., 2020; Blumenfeld, Bobb, & Marian, 2016). Cultural considerations suggested for suitability (see TD-4 and C-2), construct validity (see C-2 and C-4), and herein discussed administration recommendations will be relevant in these cases. There may not be a normative dataset that best represents a specific client/patient, but neuropsychologists must do their best to compare available and relevant data sets and to explain how and why results were congruent or incongruent with one another and the benefits and limitations of any used dataset.

- *Piloting the manual.* Test manuals should be piloted with their intended users in the intended contexts of use. For example, if it is expected that certain types of professionals should be able to use the manual alone to learn to administer the test accurately with particular populations, then the manual should be given to those types of professionals to try out with those populations. There should then be feedback mechanisms to determine if their use was appropriate and accurate. If the manual is only intended to reinforce direct instruction in the use of the test, then feedback on the manual should reflect independent use after instruction. Again, this piloting will ideally sample the full range of test administrators and test-takers concerning professional roles, languages, ethnicities, ages, disabilities, etc.

A-2 (14) Specify testing conditions that should be followed closely in all populations of interest.

ITC:

Explanation. *The goal of this guideline is to encourage test developers to establish testing instructions and related procedures (e.g., testing conditions, time limits, etc.) that can be followed closely in all populations of interest. This guideline is primarily meant to encourage test administrators to stick to standardised instructions. At the same time, accommodations might be specified to address special subgroups of individuals within each population who may need testing accommodations such as additional time, larger print, extra quiet test administration conditions, and so on. In the testing field today, these are known as "test accommodations." The goal of these accommodations is not to inflate test-taker scores, but rather to create a testing environment for these candidates so that they can show what they may feel, or know and can do.*

Variations from the standardised testing conditions should be noted, so that, later in the process, these variations and their impact on generalisations and interpretations can be considered.

Suggestions for practice. *This guideline may in part overlap with A-1 (13), but it is restated here to highlight the importance of candidates taking the test under as similar conditions as possible. This is essential if the scores from the two language versions are going to be used interchangeably. Here are some suggestions:*

> *Testing instructions and related procedures should be adapted and re-written in a standardised way, which is suitable to the new language and culture.*
>
> *If testing instructions and related procedures are changed to the new cultures, administrators should be trained on the new procedures; they should be informed with respect to these procedures and not to the original ones.*

Neuropsychological Applications

Explanation. Neuropsychologists often experience the need to balance standardisation procedures and adaptation of tests for clinical use. Possible cultural adaptations to face-to-face administration procedures were discussed in our previous A-1 (13) applications. Here, we will discuss special considerations for adapting tests for different settings such as inpatient versus outpatient and teleneuropsychology. We will provide suggestions for test adaptations for individuals with disabilities (e.g., hearing or visually impaired), and culturally specific recommendations, when appropriate.

Suggestions for Practice.

- When adapting tests for a setting that differs from the setting in which the test was originally validated (e.g., outpatient, inpatient, school, community environments, or in-home services), it is worth considering not only test-specific factors but also environment-specific elements that may confound the results or affect the construct being measured. Researchers seeking to adapt a test for a setting that differs from the validation setting, even if it has been validated for the intended culture/population, must also review the applications to all ITC guidelines because many will be pertinent. Individuals completing the adaptation will be recommended to document all changes in test structure, materials, and administration specifics in detail for the benefit of test users (see Doc-1 and Doc-2).

- When adapting tests for the inpatient setting, evaluators may account for the patient's posture (e.g., seated in bed versus seated on a chair with a table), limitations of space and time, potential distractions, and frequent interruptions. In this setting, where brief and easily administered tests are preferred, adaptations may require changes in the length of the test, practice items, task timing or between-test delays, instruction, stimuli booklets, or materials amongst others.

- When tests developed for outpatient settings are used in inpatient settings without a formal adaptation process, test administrators should consider setting- and patient-related variables in the interpretation of results and clinical integration. There are potential cultural and psychological setting-related factors that may affect privacy and the test-taker's level of psychological distress, perception of the evaluation, and effort and test engagement. For example, in Puerto Rico, being hospitalised is often a socially alarming event to some patients and their families, and when this is the case, these individuals may experience an elevated level of stress during that time frame. Even during short stays, family members (which include extended family) visit daily and, when available, caregivers spend many hours in the room. This may lead to increased distractibility for the patient during testing and reduced privacy. Similarly, most hospital rooms are outfitted with two patient beds with medical curtain dividers, which will also be an important consideration for test security, privacy during testing, potential interruptions, and the patient's potential response to feedback or corrections or their willingness to ask for clarifications (e.g., shame). Also, some medical conditions may impact testing in the inpatient setting, including disorders affecting reality testing (e.g., schizophrenia and other psychotic disorders), delirium or other CNS-related disorders, medication side effects, and altered sleep patterns, which may negatively affect the construct equivalence and reliability of the test.

- Tests adapted for teleneuropsychological use should account for potential equipment failures, display or audio glitches or delays, or internet speed and connectivity issues that may be relevant during both a formal adaptation process and the use of a non-adapted version for clinical use through telehealth in distinct scenarios. Special considerations are needed when tests are performed virtually in the test-taker's home where distractibility, interruptions, and external prompting are more likely to occur compared to an office or other more structured setting. It is also important to consider possible threats to test security and privacy and these should be addressed during the informed consent process. For example, the test administrator may request the test-taker be evaluated private room where nobody else can be to limit exposure of exposed to test materials, including individuals acting as facilitators or interpreters, unless they are providers/clinicians that have been appropriately trained (Additional guidance for third party observers from a US perspective can be found in Glen et al., 2021). There may need to be an agreement with the patient that they are not to take photos or make copies, screenshots, or recordings of any test materials.

Equipment or system failures may interrupt the assessment or add time to specific tests with fixed time delays, such as memory tests. Visually presented stimuli may also vary in size, tone, contrast, and clarity depending on the screen/monitor (e.g., size of the screen, display settings, keyboard or touch screen, phone, PC, or iPad), potentially affecting the test-taker's responses. As such, smaller devices may not be optimal, and confirmation of display quality must be obtained. There is also a possible delay or time desynchronization, typically of a couple of seconds, between the test-taker and the evaluator's side, which may significantly impact performance on some tests, such as the case with speeded tests. In a test where a response is required in a set limit of time (e.g., 30 seconds) there could be a few seconds' delay from the presentation of the stimuli (either visually or verbally) to the time the test-taker receives the image or audio, so, the timing would need to be adjusted to account for that delay. This delay is often noticeable in casual conversation and is typically stable during a given video call but variable between test-takers or video calls, as an effect of internet speed. A computation of this delay can be attempted by asking simple, quick tasks, such as "close your eyes," and timing the response time. In a case where the delay is five seconds (5"), the evaluator may adjust for this delay by counting to five seconds before starting the timer or by adding five seconds at the end of the time limit.

Adaptations of tests for teleneuropsychological use should also consider the intended population's familiarity and comfort level with the technology, and some may need to allow for technical assistance. These considerations are also important during face-to-face evaluations. For example, test adaptations targeting younger generations may consider the effect of paper-pen format vs. electronic devices (e.g., computer, tablet) to answer questions. Individuals with hearing or vision difficulties may also experience increased technological disadvantages in teleneuropsychological testing secondary to changes in the qualities of sound or visual stimuli. On tasks requiring adequate comprehension of auditory stimuli, test adaptation may consider offering an option for visual stimuli to be simultaneously presented to test-takers with auditory difficulties.

- Regardless of the environment, test adaptations should include considerations needed for test-takers with specific needs (e.g., hearing, visual, or motor impairment) or sociodemographic characteristics such as age and lower levels of formal education, to name a few. For example, researchers pursuing an adaptation of a test for individuals with limited vision should consider whether the tasks can be adapted to a verbally mediated form without affecting construct validity (e.g., oral and motor trail making test, oral and motor Symbol Digit Modalities Test (Smith, 1973)). These adaptations should also include considerations of how many practice items are allowed and how feedback will be offered (see A1). Another example of how test materials may need to be adapted includes the use of bubble answer sheets, which can be difficult for individuals with visual difficulties or dyslexia.

Score Scales and Interpretation Guidelines

SSI-1 (15) Interpret any group score differences with reference to all relevant available information.

ITC:

Explanation. *Even if a test has been adapted through technically sound procedures, and validity of the test scores has already been established to some extent, it should be kept in mind that the meaning of inter-group differences can be interpreted in many ways because of cultural or other differences across the participating countries and/or cultures. Sireci (2005) reviewed the approach for evaluating the equivalence of two different language versions of a test by administering the separate language versions of the test to a group of test-takers who are proficient in both languages (bilingual) and who come from the same cultural or language group. He outlined some research design options for equivalence studies using bilingual respondents, listed the possible confounding variables needing to be controlled, and offered some valuable suggestions for interpreting findings.*

Suggestions for practice. *One suggestion for improving practice follows:*

> *Depending on the research question (or context for which group comparisons are made), a number of possible interpretations may be considered, before finally settling on one. For example, it is important to rule out differential motivation to perform well on the test prior to inferring that one group performed better on the test than another. There may be context effects, too that significantly impacted test performance. For example, one group of persons may simply be part of a less effective education system, and this would have a significant impact on test performance.*

Neuropsychological Applications

Explanation. Accounting for culture is necessary for accurate interpretation of neuropsychological test performance. Instead of relating group characteristics to a single factor such as race/ethnicity, immigration, or bilingual status, such differences should be interpreted considering relevant available information that may contribute to the observed group characteristics, including cultural and contextual effects.

For example, Melikyan and collaborators (2021) found that a group of American adults obtained higher test raw scores on English versions than a demographically matched Russian group on Russian translations of several well-known neuropsychological tests. Inter-group differences were interpreted as being mediated by cultural differences in attitudes to timed activities, experience with timed tests and multiple-choice formats, attention to detail, and length of digit-words that put differential demands on short-term memory in Russian and English.

Suggestions for Practice. When interpreting group scores and norms on neuropsychological tests, relevant characteristics of countries and/or cultures that should be considered include, but are not limited to:

- Characteristics of educational systems between countries and/or cultures (e.g., whether or not the educational system places importance on the teaching of syntax rules), geographic regions (e.g., urban/rural), genders, generations, and educational facilities (e.g., private vs. public schools).
- Levels and types of literacy and writing systems.
- Familiarity with test materials and strategies.
- Attitudes, motivations, expectations, and strategies concerning testing. Attitudes towards timed tests and test procedures.
- Impact of immigration patterns and policies between countries (e.g., refugees; selection for skills, language, ethnicity, ages, etc., self-selection) on acculturation patterns and representativeness of home-country norms
- The level of acculturation in immigrant groups.
- Influence of stereotype threat in marginalised groups (Thames, et al., 2013).
- Other social determinants of health or non-medical factors that impact disease, treatment, and outcomes.

SSI-2 (16) Only compare scores across populations when the level of invariance has been established on the scale on which scores are reported.

ITC:

Explanation. *When comparative studies across language and cultural groups are the central focus of the translation and adaptation initiative, the multi-language versions of a test need to be placed on a common reporting scale, and this is carried out through a process called "linking" or "equating." This requires substantial sample sizes, and evidence that construct, method, and item bias are not present in the adapted version of the test.*

Van de Vijver and Poortinga (2005) delineated several levels of test equivalence across language and cultural groups and their work is especially helpful in understanding this concept; in fact, the original concept was introduced by these authors. For example, they pointed out that measurement unit equivalence requires that reporting scales in each group have the same metric, thus ensuring differences between people within the groups have the same meaning. (For example, differences between males and females in a Chinese sample can be compared to a French sample). However, valid direct score comparisons can only be done when scores show the highest level of equivalence, called scalar equivalence or full score equivalence, which requires scales in each group to have the same measurement unit and the same origin across groups.

Numerous methods (both in the framework of classical test theory and item response theory) have been put forward for linking or equating scores from two groups (or language versions of a test). Interested readers can refer to Angoff (1984) and Kolen and Brennan (2004) to gain a deeper understanding of this topic. Cook and Schmitt-Cascallar (2005) suggest a basis for understanding statistical methods that are currently available for equating and scaling educational and psychological tests. The authors describe and critique specific scale linking procedures used in test adaptation studies, and illustrate selected linking procedures and issues by describing and critiquing three studies that have been carried out over the past twenty years to link scores from the Scholastic Assessment Test to the Prueba de Aptitude Académica.

Suggestions for practice. The key point here is that the test scores should not be over-interpreted:

> *Interpret the results based on the level of validity evidence that is available. For example, do not make comparative statements about the levels of respondent performance in the two language groups unless measurement invariance has been established for test scores being compared.*

Neuropsychological Applications

Explanation. A valid interpretation of neuropsychological tests requires the availability of normative data based on national and/or cultural characteristics to determine performances that are within normal limits versus below expectations based on cultural, language, or educational characteristics. For most neuropsychological tests, a direct comparison of raw scores from people with different demographic characteristics, including cultural and language characteristics, may not be valid. A good example is WAIS. Despite having a common reporting scale across cultural and language versions (i.e., Full-Scale IQ score of 100 ± 15 points, Subscale scores of 10 ± 3 points), the raw scores used to calculate scaled scores vary greatly depending on the country/cultural population-based normative data applied. Even comparison of the same language versions across countries may be problematic as the same raw scores may lead to different scaled scores, WAIS profiles, and diagnostic classifications depending on the adopted normative data. For instance, substantial differences are present between the American and Canadian WAIS-IV norms (Harrison et al., 2014), and WAIS-IV norms in other countries such as Colombia, Mexico, and Spain (Duggan et al., 2019).

Suggestions for Practice. Neuropsychological tests rarely have the equivalence of content, the substantial sample sizes in more than one population, and the research required for scalar equivalence or full score equivalence; therefore, direct score comparisons across populations should be avoided. Thus, direct data and research are generally not available and should generally not be used to infer whether people from one cultural group have better cognitive abilities than people in another cultural group.

This also means that it becomes even more problematic to compare a single individual to a population that they are not part of. Therefore:

- Interpret the results based on the level of validity evidence that is available. For example, be cautious in the interpretation of neuropsychological test scores when applying normative data that are not fully representative of the respondent's cultural or language background. Deciding on which available normative data (if any) to use may be particularly challenging in the case of multicultural and/or multilingual respondents, including (older) migrants as neither the normative data from the country of origin nor the new country may be representative of this population (Evans et al., 2022; Plitas, et al., 2009; Staios, et al., in press).

- These concerns do not necessarily apply to criterion-based testing. For example, it may be acceptable to use scores and norms that are based on a population that is not representative of the individual if the issue concerns his or her capacity for specific adaptive behaviours such as a test of driving safety or of worker qualifications, or other competencies. But even then, caution is warranted. For example, using an absolute score on the Trail Making Test Part B to predict driving abilities would not be acceptable, but taking the individual out on the road behind the wheel to test their driving ability would be reasonable.

Documentation Guidelines

Doc-1 (17) Provide technical documentation of any changes, including an account of the evidence obtained to support equivalence, when a test is adapted for use in another population.

ITC:

Explanation. *The importance of this guideline has been realised and emphasised by many researchers (see, for example, Grisay, 2003). TIMSS and PISA have been very successful in observing this guideline by carefully documenting the changes throughout the adaptation work. With this information, there can be focus on the suitability of the changes that were made.*

The technical documentation should also contain sufficient detail of the methodology for future researchers to replicate the procedures used on the same or other populations. It should contain sufficient information from the evidence of construct equivalence and scaling equivalence (if carried out) to support the use of the instrument in the new population. Where inter-population comparisons are to be made, the documentation should report the evidence used to determine the equating of scores between populations.

Sometimes, the question arises about the intended audience for the technical documentation. The documentation should be written for the technical expert and for persons who will be required to evaluate the utility of the test for use in the new or other populations. (A brief supplementary document could be added for the benefit of a non-expert.)

Suggestions for practice. *Adapted tests should have a technical manual that documents all the qualitative and quantitative evidence associated with the adaptation process. It is especially helpful to document any changes that were made to accommodate the test in a second language and culture. Basically, technical persons and journal editors will want documentation on the process which was completed to produce and validate the target language version of the test. Of course, too, they will want to see the results from all of the analyses. Here are the types of questions that need to be addressed:*

> *What evidence is available to support the utility of the construct and the adapted test in the new population?*
>
> *What item data were collected and from what samples?*
>
> *What other data were obtained to assess content, criterion-related, and construct validity?*

Neuropsychological Applications

A technical document and test manual should be developed, including comprehensive details of the adaptation process and the evidence that the adapted test is reliable in the new context in which it is intended to be used and valid for its intended uses. There should be a description of the normative data collection process, if and how interpreters were used to collect data, the normative sample characteristics, and the metrics used to reflect test performance (e.g., scaled scores, T scores, or regression-based norms).

Suggestions for practice: The technical document could be written in the form of a manual to accompany the test or a technical paper in the form of a journal article. The document should include the following:

- Describe the intended purpose of the adapted test in the new context.
- Provide information on how it was established that the construct being measured in the original context is relevant in the intended new context.
- Explain how it was decided whether items in the original test should be translated or adapted in the new test and the process by which translation/adaptation was undertaken.
- Provide evidence that the test formats, including the response formats (e.g., rating scales; use of pen/pencil), are familiar to the population for whom the new test is intended.
- Give details of how the test development process addressed all the factors that are not directly relevant to the variables that the test is intended to measure but are considered likely to influence test performance, such as education/literacy, and familiarity with test-taking.
- Provide evidence relating to the initial piloting of the test in the new context, including evidence that the translated/adapted test instructions and items have the intended meaning in the new context and any changes that were made as a result of the piloting process.
- If interpreters are used to collect data, specify how this process was undertaken (e.g., level of qualification, training in test administration, tests administered with the aid of a neuropsychologist).
- Document the evidence that the test is reliable in the new context and provide detailed information about all types of reliability assessed (e.g., interrater reliability, test-retest reliability).
- Provide evidence that the test is valid for its intended purpose in the new context (e.g., construct validity, confirmatory factor analysis).
- Provide details of the normative sample and the process by which data were collected from the sample, such as details of how participants were selected (e.g., representative of the census), recruited, and tested (including the contexts in which participants were tested).
- Provide as many detailed demographic characteristics of the normative sample as possible to help individuals identify whether the normative set applies to their clients/patients (e.g., age, quality and years of education, sex, ethnicity).
- Describe the intended population and any limitations in the applicability of the test for specific populations. For example, if the test poses a limitation or may underrepresent the ability of individuals with motor difficulties, how could the evaluator provide accommodations to minimise these effects? Are there any interpretation considerations?
- Explain what metrics were used to represent test performance and how the applicable metrics were determined.
- Provide details/results of analysis of the normative data, including analysis of factors considered likely to influence test performance.
- Provide the rationale for the approach to the provision of normative data (e.g., stratification of norms according to age/education/sex; regression-based norms)
- Provide the normative data in a form that facilitates scoring and test interpretation (e.g., t-scores, scaled scores)

This list of suggestions for the content of a technical document is not necessarily comprehensive and other aspects of the test adaptation specific to the individual test being reported may need to be included.

Doc-2 (18) Provide documentation for test users that will support good practice in the use of an adapted test with people in the context of the new population.

ITC:

Explanation. *The documentation should be written for people who will be using the test in practical assessment settings. It should be consistent with the good practice defined by the International Test Commission Guidelines on Test Use (see www.InTestCom.org).*

Suggestions for practice*. The test developer should provide specific information on the ways in which socio-cultural and ecological contexts of the populations might affect performance on the test. The user manual should:*

Describe the construct(s) measured by the test and summarise the information; describe the adaptation process.

Summarise the evidence supporting the adaptation, including evidence for cultural suitability of the item content, test instructions, response format, etc.

Define the suitability of using the test with various subgroups within the population and any other restrictions on use.

Explain any issues that need to be considered in relation to good practice in test administration.

Explain if and how inter-population comparisons can be made.

Provide the information necessary for scoring and norming (e.g., relevant norm look-up tables) or describe how users can access scoring procedures (e.g., where these are computer-based).

Provide guidelines for the interpretation of results, including information on the implications of

Neuropsychological Applications

Explanation. A user manual is required for all neuropsychological tests. It should explain the rationale for the test in its original and new language and/or cultural group, provide evidence that the new test is reliable and valid, provide administration and scoring instructions, and give guidance on interpretation.

Suggestions for practice. The user manual should:

- Describe the background rationale for the original/source test and its intended use in the new language and/or cultural group.
- Briefly explain how the test was adapted from the original (refer to the technical manual for more detail).
- Provide evidence that the test is reliable and valid in the new language and/or cultural group.
- Provide detailed administration instructions. Ensure that any context-specific instructions that may be different from those in the original language and/or cultural group are explained carefully. This may include how the test is introduced; how instructions should be given; management of the test environment; management of the presentation format for people with physical impairment or via atypical presentation formats (e.g., teleneuropsychology); and how responses should be recorded (see A-1 and A-2).

- Explain how the test should be scored, including how to obtain the relevant test-metric scores from the normative data set. If this involves using a computerised scoring program, give details explaining how this should be accessed.
- Provide guidance on the interpretation of scores specific to the new language and/or cultural group in which the test is being used. This may include, but is not limited to, categorical interpretation of test scores (e.g., average; low-average; below average, etc.); interpretation of differences between sub-tests; interpretation of test performances that commonly reflect impairment in different cognitive processes; consistency of test performance or profile with particular patient populations; etc.
- Indicate to what degree the adapted version is appropriate for direct comparison of populations with the original test (psychometric equivalence) or only similar purposes (family of tests) and the basis for such a judgement.

FINAL WORDS

ITC:

We have done our best to deliver a set of guidelines to help test developers and test users in their work. However, for the guidelines and other efforts to change poor practices to have an effect, there must be good dissemination mechanisms in place. A recent systematic review by Rios and Sireci (2014) demonstrated that the majority of test adaptation projects in the published literature did not, in fact, followed the ITC Guidelines that have been available for around 20 years now. We thereby encourage the readers to make every effort to increase awareness among colleagues of this Second Edition as a primary source of best practices to which so many professionals around the world have contributed.

At the same time, we know that just as the first edition of these guidelines is now being replaced, so in turn will these second edition guidelines. The well-known AERA, APA, and NCME test standards are now in their sixth edition (AERA, APA, & NCME, 2014). We expect the ITC Guidelines for Adapting Tests to undergo another revision too in the coming years. If you know of new studies that should be cited, or influence the third edition, or you want to offer new guidelines or revisions to the 18 guidelines presented here please let the ITC know. You can contact the current chair of the Research and Guidelines committee that produced the second edition and/or the secretary of the ITC at the email address found on www.InTestCom.org.

Neuropsychological Applications

We are grateful for this opportunity to bring the insights and developments of the ITC more fully to the attention of the neuropsychology community. We hope that this will contribute to improved health equity and neuropsychological accuracy globally. We encourage the dissemination of these applications under the terms of our Creative Commons license. We particularly encourage test translators/adaptors/developers as well as reviewers and editors to use these guidelines, to report to us on their usefulness, and to suggest improvements and updates.

In addition to the ITC Guidelines reproduced above, the original (second edition) Guidelines https://www.intestcom.org/files/guideline_test_adaptation_2ed.pdf also contain a Glossary and a checklist.

The Cultural Neuropsychology Special Interest Group of the International Neuropsychological Society has its web home at https://www.the-ins.org/sigs/.

REFERENCES

Acevedo, A., Loewenstein, D. A., Barker, W. W., Harwood, D. G., Luis, C., Bravo, M., ... & Duara, R. (2000). Category fluency test: normative data for English-and Spanish-speaking elderly. *Journal of the International Neuropsychological Society*, *6*(7), 760-769.

Al Salman, Ahmed Saeed Ali (2013) The Saudi Arabian Adaptation of the Addenbrooke's Cognitive Examination – Revised (Arabic ACER). PhD thesis. http://theses.gla.ac.uk/4706/

Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement, 36*(3), 185-198.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Angoff, W. H. (1984). *Scales, norms, and equivalent scores*. Educational Testing Service.

Angoff, W. H., & Modu, C. C. (1973). Equating the scales of the Prueba de Apititud Academica and the Scholastic Aptitude Test (Research Rep No. 3). College Entrance Examination Board.

Ardila, A. (2005). Cultural values underlying psychometric cognitive testing. *Neuropsychology Review, 15*(4), 185-195. https://doi.org/10.1007/s11065-005-9180-y

Ardila, A. (2007). Toward the development of a cross-linguistic naming test. *Archives of Clinical Neuropsychology*, *22*(3), 297–307.

Ardila, A., Bertolucci, P. H., Braga, L. W., Castro-Caldas, A., Judd, T., Kosmidis, M. H., Matute, E., Nitrini, R., Ostrosky-Solis, F., & Rosselli, M. (2010). Illiteracy: the neuropsychology of cognition without reading. *Archives of Clinical Neuropsychology*, *25*(8), 689–712. https://doi.org/10.1093/arclin/acq079

Asparouhov, T. & Muthén, B. (2009). Exploratory structural modeling. *Structural Equation Modeling, 16*(3), 397-438. https://doi.org/10.1080/10705510903008204

assessments. *International Journal of Testing*, *2*(3), 199-215. https://doi/org/10.1080/15305058.2002.9669493

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 8, pp. 47–89). Academic Press.

Bak, T.H. & Mioshi, E. (2007). A cognitive bedside assessment beyond the MMSE: the Addenbrooke's Cognitive Examination. Practical Neurology 7(4), 245-9.

Ballard, E., Charters, H., & Taumoefolau, M. (2018). A guide to designing a naming test for an under-researched bilingual population: adapting the Boston Naming Test to Tongan. *Clinical Linguistics & Phonetics*, *33*(4), 376-392. https://doi.org/10.1080/02699206.2018.1518488

Barker-Collo, S., Clarkson, A., Cribb, A., & Grogan, M. (2002). The impact of American content on California verbal learning test performance: a New Zealand illustration. *The Clinical Neuropsychologist*, *16*(3), 290-299.

Basso, M. R., Carona, F. D., Lowery, N., & Axelrod, B. N. (2002). Practice effects on the WAIS-III across 3- and 6-month intervals. *The Clinical neuropsychologist*, *16*(1), 57–63. https://doi.org/10.1076/clin.16.1.57.8329

Behr, D. (2017). Assessing the use of back translation: The shortcomings of back translation as a quality testing method. *International Journal of Social Research Methodology*, *20*(6), 573-584. https://doi.org/10.1080/13645579.2016.1252188

Bender, A. H., GarcÍa. A. M., & Barr, W. B. (2010). An interdisciplinary approach to neuropsychological test construction: Perspectives from translation studies. *Journal of the International Neuropsychological Society*, *16*(2), 227-232. https://doi.org/10.1017/S1355617709991378

Benedict, R. H., & Zgaljardic, D. J. (1998). Practice effects during repeated administrations of memory tests with and without alternate forms. *Journal of clinical and experimental neuropsychology*, *20*(3), 339–352. https://doi.org/10.1076/jcen.20.3.339.822

Blanco-Elorrieta, E., & Pylkkänen, L. (2018). Ecological validity in bilingualism research and the bilingual advantage. *Trends in Cognitive Sciences*, *22*(12), 1117-1126. https://doi.org/10.1016/j.tics.2018.10.001

Blumenfeld, H. K., Bobb, S. C., & Marian, V. (2016). The role of language proficiency, cognate status, and word frequency in the assessment of Spanish–English bilinguals' verbal fluency. *International Journal of Speech-language Pathology*, *18*(2), 190-201. https://doi.org/10.3109/17549507.2015.1081288

Brandt, J., & Benedict, R. H. B. (2001). *Hopkins Verbal Learning Test–Revised*. Odessa, FL: PAR.

Brislin, R. W. (1986). The wording and translation of research instruments. In W. J. Lonner & J. W. Berry (Eds.), *Field methods in cross-cultural psychology* (pp. 137-164). Sage Publications.

Byrd, D. A., Rivera Mindt, M. M., Clark, U. S., Clarke, Y., Thames, A. D., Gammada, E. Z., & Manly, J. J. (2021). Creating an antiracist psychology by addressing professional complicity in psychological assessment. *Psychological Assessment*, *33*(3), 279. https://doi.org/10.1037/pas0000993

Byrne, B. (2001). Structural equation modeling with AMOS, EQS, and LISREL: Comparative approaches to testing for the factorial validity of a measuring instrument. *International Journal of Testing*, *1*9(1), 55-86. https://doi.org/10.1207/S15327574IJT0101_4

Byrne, B. (2003). Measuring self-concept measurement across culture: Issues, caveats, and application. In H. W. Marsh, R. Craven, & D. M. McInerney (Eds.), *International advances in self research*. Greenwich, CT: Information Age Publishing.

Byrne, B. (2006). *Structural equation modeling with EQS: Basic concepts, applications, and programming* (2nd ed.). Lawrence Erlbaum publishers.

Byrne, B. M. (2008). Testing for multigroup equivalence of a measuring instrument: A walk through the process. *Psicothema*, *20*, 872-882

Byrne, B. M., & van de Vijver, F.J.R. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence. *International Journal of Testing*, *10*(2), 107-132. https://doi.org/10.1080/15305051003637306

Byrne, B. M., & van de Vijver, F.J.R. (2014). Factorial structure of the family values scale from a multilevel-multicultural perspective. *International Journal of Testing, 14*(2), 168-192. https://doi.org/10.1080/15305058.2013.870903

Calamia, M., Markon, K., & Tranel, D. (2012). Scoring higher the second time around: meta-analyses of practice effects in neuropsychological assessment. *The Clinical neuropsychologist*, *26*(4), 543–570. https://doi.org/10.1080/13854046.2012.680913

Casaletto, K. B., & Heaton, R. K. (2017). Neuropsychological assessment: Past and future. *Journal of the International Neuropsychological Society*, *23*(9-10), 778-790. https://doi.org/10.1017/S1355617717001060

Chan, M. E., & Elliott, J. M. (2011). Cross-linguistic differences in digit memory span. *Australian Psychologist*, *46*(1), 25-30.

Chatterjee, Sumita (2021) *Differences in object perception: a comparison of Indian and British participants on scene and silhouetted object perception tasks.* PhD thesis, University of Glasgow.

Cheung F. M. (2012). Mainstreaming culture in psychology. *The American psychologist*, *67*(8), 721–730. https://doi.org/10.1037/a0029876

Chincotta, D., & Underwood, G. (1996). Mother tongue, language of schooling and bilingual digit span. *British Journal of Psychology*, *87*(2), 193-208.

Clauser, B. E., Nungester, R. J., Mazor, K., & Ripley, D. (1996). A comparison of alternative matching strategies for DIF detection in tests that are multidimensional. *Journal of Educational Measurement, 33*(2), 202-214. https://doi.org/10.1111/j.1745-3984.1996.tb00489.x

Cockcroft, K., Alloway, T., Copello, E., & Milligan, R. (2015). A cross-cultural comparison between South African and British students on the Wechsler adult intelligence scales third edition (WAIS-III). *Frontiers in Psychology*, *6*, 297. https://doi.org/10.3389/fpsyg.2015.00297

Cohen, M. J., & Stanczak, D. E. (2000). On the reliability, validity, and cognitive structure of the Thurstone Word Fluency Test. *Archives of clinical neuropsychology*, *15*(3), 267-279.

Cook, L. L., & Schmitt-Cascallar, A. P. (2005). Establishing score comparability for tests given in different languages. In R. K. Hambleton, P. F. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 139-170).

Coste, D., Moore, E., & Zarate, G. (2009). *Plurilingual and pluricultural competence.* Council of Europe, Language Policy Division, Strasbourg. available at www.coe.int/lang

Council of Europe (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*, Council of Europe Publishing, Strasbourg, available at www.coe.int/lang-cefr.

Council of National Psychological Associations for the Advancement of Ethnic Minority Interests (CNPAAEMI) *Testing and Assessment with Persons & Communities of Color.* American Psychological Association; Washington, DC: 2016. Available online: https://www.apa.org/pi/oema [Google Scholar]

Crombie, M., Dutt, A., Dey, P., Nandi, R., & Evans, J. (2023). Examination of the validity of the 'Papadum test': An alternative to the clock drawing test for people with low levels of education.

The Clinical Neuropsychologist, *37*(5), 1025–1042. https://doi.org/10.1080/13854046.2022.2047789

Crystal, David (2003). English as a Global Language (2nd ed.). Cambridge University Press.

Das, S. K., Banerjee, T. K., Mukherjee, C. S., Bose, P., Hazra, A., Dutt, A., Das, S., Chaudhuri, A., & Raut, D. K. (2006). An urban community-based study of cognitive function among non-demented elderly population in India. *Neurology Asia*, *11*, 37–48.

Daugherty, J. C., Puente, A. E., Fasfous, A. F., Hidalgo-Ruzzante, N., & Pérez-Garcia, M. (2017). Diagnostic mistakes of culturally diverse individuals when using North American neuropsychological tests. *Applied Neuropsychology: Adult*, *24*(1), 16-22. https://doi.org/10.1080/23279095.2015.1036992

De Bruin, A. (2019). Not all bilinguals are the same: A call for more detailed assessments and descriptions of bilingual experiences. *Behavioral Sciences*, *9*(3), 33. https://doi.org/10.3390/bs9030033

D'Elia, L. F., Satz, P., Uchiyama, C. L., & White, T. (1996). Color Trails Test: Professional Manual. Odessa, FL: Psychological Assessment Resources.

Disabilities Trust (2022). https://www.thedtgroup.org/research/bmipb downloaded 19/9/22.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning: Theory and Practice* (pp. 137-166).

Duggan, E. C., Awakon, L. M., Loaiza, C. C., & Garcia-Barrera, M. A. (2019). Contributing towards a cultural neuropsychology assessment decision-making framework: Comparison of WAIS-IV norms from Colombia, Chile, Mexico, Spain, United States, and Canada. *Archives of Clinical Neuropsychology*, *34*(5), 657-681. https://doi.org/10.1093/arclin/acy074

Dutt, A, Evans J., & Fernandez, A.L. (2022). *Challenges for neuropsychology in the global context.* Understanding Cross-Cultural Neuropsychology, Science, Testing and Challenges (Current Issues in Neuropsychology), Oxford, UK: *Routledge* | Taylor & Francis Group

Ellis, B. B. (1989). Differential item functioning: Implications for test translation. *Journal of Applied Psychology*, *74*(6), 912-921. https://doi.org/10.1037/0021-9010.74.6.912

Ellis, B. B., & Kimmel, H. D. (1992). Identification of unique cultural response patterns by means of item response theory. *Journal of Applied Psychology*, *77*(2), 177-184. https://doi.org/10.1037/0021-9010.77.2.177

Ellis, N. (1992). "Linguistic Relativity Revisited: The Bilingual Word-Length Effect in Working Memory During Counting, Remembering Numbers, and Mental Calculations." In *Cognitive Processing in Bilinguals*, edited by R. I. Harris, 137–155. Amsterdam: Elsevier. https://doi.org/10.1016/S0166-4115(08)61492-2

Ercikan, K. (1998). Translation effects in international assessments. *International Journal of Educational Research*, *29*(6), 543-533. https://doi.org/10.1016/S0883-0355(98)00047-0

Ercikan, K. (2002). Disentangling sources of differential item functioning in multilanguage

Ercikan, K., Gierl, J. J., McCreith, T., Puhan, G., & Koh, K. (2004). Comparability of bilingual versions of assessments: Sources of incomparability of English and French versions of Canada's national achievement tests. *Applied Measurement in Education, 17*(3), 301-321. https://doi.org/10.1207/s15324818ame1703_4

Ercikan, K., Simon, M., & Oliveri, M. E. (2013). Score comparability of multiple language versions of assessments within jurisdictions. In M. Simon, K. Ercikan, & M. Rousseau (Eds.), *An international handbook for large-scale assessments* (pp. 110-124). New York: Grégoire, J., & Hambleton, R. K. (Eds.). (2009). Advances in test adaptation research [Special Issue]. *International Journal of Testing, 9*(2), 73-166.

Evans J., Dutt, A, & Fernandez, A.L. (2022). *The future of neuropsychology in a global context.* Understanding Cross-Cultural Neuropsychology, Science, Testing and Challenges (Current Issues in Neuropsychology), Oxford, UK: *Routledge* | Taylor & Francis Group

Fasfous, A. F., Al-Joudi, H. F., Puente, A. E., & Pérez-García, M. (2017). Neuropsychological measures in the Arab world: A systematic review. *Neuropsychology review*, *27*(2), 158-173. doi:10.1007/s11065-017-9347-3

Fernández, A. L., & Abe, J. (2018). Bias in cross-cultural neuropsychological testing: problems and possible solutions. *Culture and Brain*, *6*(1), 1-35. https://doi.org/10.1007/s40167-017-0050-2

Folstein MF, Folstein SE, McHugh PR. "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res.* 1975;12(3):189–198. doi: 10.1016/0022-3956(75)90026-6.

Franzen, S., van den Berg, E., Bossenbroek, W., Kranenburg, J., Scheffers, E. A., van Hout, M., van de Wiel, L., Goudsmit, M., van Bruchem-Visser, R. l., van Hemmen, J., Jiskoot, L.C., & Papma, J. M. (2022). Neuropsychological assessment in the multicultural memory clinic: Development and feasibility of the TULIPA battery. *The Clinical Neuropsychologist*, 1-21. https://doi.org/10.1080/13854046.2022.2043447

Franzen, S., van den Berg, E., Kalkisim, Y., van de Wiel, L., Harkes, M., van Bruchem-Visser, R. L., de Jong, F. J., Jiskoot, L. C., & Papma, J. M. (2019). Assessment of visual association memory in low-educated, non-Western immigrants with the modified visual association test. *Dementia and Geriatric Cognitive Disorders, 47*(4–6), 345–354. https://doi.org/10.1159/000501151

Ghasemian-Shirvan, E., Shirazi, S. M., Aminikhoo, M., Zareaan, M., & Ekhtiari, H. (2018). Preliminary normative data of Persian phonemic and semantic verbal fluency test. *Iranian journal of psychiatry*, *13*(4), 288.

Glen, T., Barisa, M., Ready, R., Peck, E., & Spencer, T. R. (2021). Update on third party observers in neuropsychological evaluation: An interorganizational position paper. *Archives of Clinical Neuropsychology, 36*(5), 686-692. https://doi.org/10.1093/arclin/acab016

Goodenough, F. L. (1926). *Measurement of intelligence by drawings*. New York: Harcourt, Brace, and World.

Grisay, A. (2003). Translation procedures in OECD/PISA 2000 international assessment. *Language Testing, 20*(2), 225-240. https://doi.org/10.1191/0265532203lt254oa

Guàrdia-Olmos, J., Peró-Cebollero, M., Rivera, D., & Arango-Lasprilla, J.C. (2015). Methodology for the development of normative data for ten Spanish-language neuropsychological tests in eleven Latin American countries. *Neuro Rehabilitation, 37*(4), 493-499. https://doi.org/10.3233/NRE-151277

Hambleton, R. K. (2002). The next generation of the ITC test translation and adaptation guidelines. *European Journal of Psychological Assessment, 17*(3), 164-172. https://doi.org/10.1027/1015-5759.17.3.164

Hambleton, R. K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R. K. Hambleton, P. F. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 3-38). Lawrence Erlbaum Publishers.

Hambleton, R. K., & de Jong, J. (Eds.). (2003). Advances in translating and adapting educational and psychological tests. *Language Testing, 20*(2), 127-240. https://doi.org/10.1191/0265532203lt247xx

Hambleton, R. K., & Lee, M. (2013). Methods of translating and adapting tests to increase cross-language validity. In D. Saklofske, C. Reynolds, & V. Schwean (Eds.), *The Oxford handbook of child assessment* (pp. 172-181). Oxford University Press.

Hambleton, R. K., & Patsula, L. (1999). Increasing the validity of adapted tests: Myths to be avoided and guidelines for improving test adaptation practices. *Applied Testing Technology, 1*(1), 1-16.

Hambleton, R. K., & Zenisky, A. (2010). Translating and adapting tests for cross-cultural assessment. In D. Matsumoto & F. van de Vijver (Eds.), *Cross-cultural research methods* (pp. 46-74). Cambridge University Press.

Hambleton, R. K., Clauser, B. E., Mazor, K. M., & Jones, R. W. (1993). Advances in the detection of differentially functioning test items. *European Journal of Psychological Assessment, 9*(1), 1-18.

Hambleton, R. K., Merenda, P. F., & Spielberger, C. (Eds.). (2005). *Adapting educational and psychological tests for cross-cultural assessment*. Lawrence Erlbaum Publishers.

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

Hambleton, R. K., Yu, L., & Slater, S. C. (1999). Field-test of ITC guidelines for adapting psychological tests. *European Journal of Psychological Assessment, 15* (3), 270-276.

Harris, S. (2022, September 09). *Translation vs. localization vs. transcreation: Is there a difference?* Argos Multilingual. https://www.argosmultilingual.com/blog/translation-localization-difference

Harrison, A. G., Armstrong, I. T., Harrison, L. E., Lange, R. T., & Iverson, G. L. (2014). Comparing Canadian and American normative scores on the Wechsler adult intelligence scale. *Archives of Clinical Neuropsychology, 29*(8), 737-746. https://doi.org/10.1093/arclin/acu048

Heaton RK. *A Manual for the Wisconsin Card Sorting Test.* Odessa, Fla, USA: Psychological Assessment Resources; 1981.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Lawrence Erlbaum Associates.

http://doi.org/10.1017/S1041610220000344

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement, 6*(3), 249-260.

International Test Commission. (2005). The ITC Guidelines for Translating and Adapting Tests (First edition). [www.InTestCom.org]

International Test Commission. (2017). The ITC Guidelines for Translating and Adapting Tests (Second edition). [www.InTestCom.org]

Ivanova, M. V., & Hallowell, B. (2013). A tutorial on aphasia test development in any language: Key substantive and psychometric considerations. *Aphasiology, 27*(8), 891–920. https://doi.org/10.1080/02687038.2013.805728.

Jacklin, K., Pitawanakwat, K., Blind, M., O'Connell, M. E., Walker, J., Lemieux, A. M., & Warry, W. (2020). Developing the Canadian Indigenous cognitive assessment for use with Indigenous older Anishinaabe adults in Ontario, Canada. *Innovation in Aging*, *4*(4), igaa038. https://doi.org/10.1093/geroni/igaa038

Javaras, K. N., & Ripley, B. D. (2007). An 'unfolding' latent variable model for Likert attitude data: Drawing inferences adjusted for response style. *Journal of the American Statistical Association, 102*(478), 454-463. https://doi.org/10.1198/016214506000000960

Jeanrie, C., & Bertrand, R. (1999). Translating tests with the International Test Commission Guidelines: Keeping validity in mind. *European Journal of Psychological Assessment*, *15*(3), 277-283. https://doi.org/10.1027/1015-5759.15.3.277

Johnson, T. R. (2003). On the use of heterogeneous thresholds ordinal regression models to account for individual differences in response style. *Psychometrika, 68*(4), 563-583. https://doi.org/10.1007/BF02295612

Kaplan, E., Goodglass, H., & Weintrab, S. (1983). *The Boston naming test*. Philadelphia: Lea & Febiger.

Kavé G. (2005). Phonemic fluency, semantic fluency, and difference scores: normative data for adult Hebrew speakers. *Journal of clinical and experimental neuropsychology*, *27*(6), 690–699. https://doi.org/10.1080/13803390490918499

Kempler, D., Teng, E. L., Dick, M., Taussig, I. M., & Davis, D. S. (1998). The effects of age, education, and ethnicity on verbal fluency. *Journal of the International Neuropsychological Society*, *4*(6), 531-538.

Kolen, M. J., & Brennan, R. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). Springer.

Kosmidis, M. H., Vlahou, C. H., Panagiotaki, P., & Kiosseoglou, G. (2004). The verbal fluency task in the Greek population: Normative data, and clustering and switching strategies. *Journal of the International Neuropsychological Society*, *10*(2), 164-172.

Lehtonen, M., Soveri, A., Laine, A., Järvenpää, J., de Bruin, A., & Antfolk, J. (2018). Is bilingualism associated with enhanced executive functioning in adults? A meta-analytic review. *Psychological bulletin*, *144*(4), 394–425. https://doi.org/10.1037/bul0000142

Lemay, S., Bédard, M. A., Rouleau, I., & Tremblay, P. L. (2004). Practice effect and test-retest reliability of attentional and executive tests in middle-aged to elderly subjects. *The Clinical neuropsychologist*, *18*(2), 284–302. https://doi.org/10.1080/13854040490501718

Levin, K., Willis, G. B., Forsyth, B. H., Norberg, A., Stapleton Kudela, M., Stark, D., & Thompson, F. E. (2009). Using cognitive interviews to evaluate the Spanish-language translation of a dietary questionnaire. *Survey Research Methods, 3*(1), 13-25. https://doi.org/10.18148/srm/2009.v3i1.88

Levine, A. J., Miller, E. N., Becker, J. T., Selnes, O. A., & Cohen, B. A. (2004). Normative data for determining significance of test-retest differences on eight common neuropsychological instruments. *The Clinical neuropsychologist*, *18*(3), 373–384. https://doi.org/10.1080/13854040490524 20

Lezak, M.D., Howieson, D.B., Bigler, E., Tranel, D. (2012). *Neuropsychological Assessment, 5th Edition*. Oxford University Press.

Li, Y., Cohen, A. S., & Ibarra, R. A. (2004). Characteristics of mathematics items associated with gender DIF. *International Journal of Testing*, *4*(2), 115-135. https://doi.org/10.101207/s15327574ijt0402_2

Lindeboom J, Schmand B. *Visual Association Test*. Leiden: PITS; 2003.

Luk, G. (2022). Justice and equity for whom? Reframing research on the "bilingual (dis)advantage". *Applied Psycholinguistics*, 1-15. doi:10.1017/S0142716422000339

Marian, V., & Hayakawa, S. (2021). Measuring bilingualism: The quest for a "bilingualism quotient". *Applied Psycholinguistics*, *42*(2), 527-548. https://doi.org/10.1017/S0142716420000533

Mathuranath, P. S., George, A., Cherian, P. J., Alexander, A. L., Sarma, S. G., & Sarma, P. S. (2003). Effects of age, education and gender on verbal fluency. *Journal of clinical and experimental neuropsychology*, *25*(8), 1057-1064. https://doi.org/10.1076/jcen.25.8.1057.16736

Matthews, C. G. (1992). Truth in labeling: Are we really an international society? *Journal of Clinical and Experimental Neuropsychology, 14*(3), 418–426. https://doi.org/10.1080/01688639208407617

Mazor, K. H., Clauser, B. E., & Hambleton, R. K. (1992). The effect of simple size on the functioning of the Mantel-Haenszel statistic. *Educational and Psychological Measurement, 52*(2), 443-451. https://doi.org/10.1177/0013164492052002020

Melikyan, Z. A., Puente, A. E., & Agranovich, A. V. (2021). Cross-cultural comparison of rural healthy adults: Russian and American groups. *Archives of Clinical Neuropsychology, 36*(3), 359–370. https://doi.org/10.1093/arclin/acz071

Messinis, L., Nasios, G., Mougias, A., Politis, A., Zampakis, P., Tsiamaki, E., Malefaki, S., Gourzis, P., & Papathanasopoulos, P. (2016). Age and education adjusted normative data and discriminative validity for Rey's Auditory Verbal Learning Test in the elderly Greek population. *Journal of Clinical and Experimental Neuropsychology, 38*(1), 23–39. https://doi.org/10.1080/13803395.2015.1085496

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97. https://doi.org/10.1037/h0043158

Mindt, M. R., Arentoft, A., Coulehan, K., Summers, A. C., Tureson, K., Aghvinian, M., & Byrd, D. A. (2019). Neuropsychological Evaluation of Culturally/Linguistically Diverse Older Adults. In L. D. Ravdin & H. L. Katzen (Eds.), *Handbook on the Neuropsychology of Aging and Dementia* (pp. 25–48). Springer International Publishing. https://doi.org/10.1007/978-3-319-93497-6_3

Muñiz, J., Elosua, P., & Hambleton, R. K. (2013). Directrices para la traduccion y adaptacion de los tests: segunda edicion. *Psicothema*, *25(2)*, 151-157. http://doi.org/ 10.7334/psicothema2013.24

Muñiz, J., Hambleton, R. K., & Xing, D. (2001). Small sample studies to detect flaws in item translations. *International Journal of Testing*, *1*(2), 115-135. https://doi.org/10.1207/S15327574IJT0102_2

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

Nielsen, T. R., & Jørgensen, K. (2013). Visuoconstructional abilities in cognitively healthy illiterate Turkish immigrants: A quantitative and qualitative investigation. *The Clinical Neuropsychologist, 27*(4), 681–692. http://doi.org/10.1080/ 13854046.2013.767379

Nielsen, T. R., & Jørgensen, K. (2020). Cross-cultural dementia screening using the Rowland Universal Dementia Assessment Scale: a systematic review and meta-analysis. *International psychogeriatrics*, *32*(9), 1031-1044.

Nielsen, T. R., & Waldemar, G. (2016). Effects of literacy on semantic verbal fluency in an immigrant population. *Aging, Neuropsychology, and Cognition*, *23*(5), 578-590. http://doi.org/ 10.1080/13825585.2015.1132668

Nielsen, T. R., Segers, K., Vanderaspoilden, V., Bekkhus-Wetterberg, P., Minthon, L., Pissiota, A., Bjørkløf, G. H., Beinhoff, U., Tsolaki, M., Gkioka, M., & Waldemar, G. (2018). Performance of middle-aged and elderly European minority and majority populations on a Cross-Cultural Neuropsychological Test Battery (CNTB). *The Clinical Neuropsychologist*, *32*(8), 1411–1430. https://doi.org/10.1080/13854046.2018.1430256

Oort, F. J., & Berberoğlu, G. (1992). Using restricted factor analysis with binary data for item bias detection and item analysis. In T. J. Plomp, J. M. Pieters, & A. Feteris (Eds.), *European Conference on Educational Research: Book of Summaries* (pp. 708-710). Twente, the Netherlands: University of Twente, Department of Education.

Ostrosky, F., Ardila, A., Rosselli, M., Lo´pez-Arango, G., & Uriel-Mendoza, V. (1998). Neuropsychological test performance in illiterates. *Archives of Clinical Neuropsychology, 13*(7), 645–660. https://doi.org/10.1093/arclin/13.7.645

Ozolins, U., Hale, S., Cheng, X., Hyatt, A., & Schofield, P. (2020). Translation and back-translation methodology in health research–a critique. *Expert Review of Pharmacoeconomics & Outcomes Research*, *20*(1), 69-77. https://doi.org/10.1080/14737167.2020.1734453

Paradis, M. & Libben, G. (1987). *The assessment of bilingual aphasia.* Psychology Press.

Paradis, M. (2011). Principles underlying the Bilingual Aphasia Test (BAT) and its uses. *Clinical Linguistics & Phonetics*, 25: 427-443

Park, H., Pearson, P. D., & Reckase, M. D. (2005). Assessing the effect of cohort, gender, and race on DIF in an adaptive test designed for multi-age groups. *Reading Psychology*, *26*(1), 81-101. https://doi.org/10.1080/02702710590923805

Plitas, A., Tucker, A., Kritikos, A., Walters, I., & Bardenhagen, F. (2009). Comparative study of the cognitive performance of Greek Australian and Greek national elderly: Implications for neuropsychological practice, *Australian Psychologist,* 44(1), 27-39. http://doi.org/10.1080/00050060802587694

Porrselvi A. P. (2022) TAM battery: Development and pilot testing of a Tamil computer-assisted cognitive test battery for older adults, The Clinical Neuropsychologist, DOI: 10.1080/13854046.2022.2156396

Porrselvi, A.P. & Shankar, V. (2018). Limitations of Integrating Technology in Cognitive Testing of Non-verbal memory. The International Journal of Indian Psychology 6(3):167-173. DOI: 10.25215/0603.98

Qi, W. G., Sun, X., & Hong, Y. (2022). Normative Data for Adult Mandarin-Speaking Populations: A Systematic Review of Performance-Based Neuropsychological Instruments. *Journal of the International Neuropsychological Society: JINS*, *28*(5), 520–540.

Rapport, L.J., Brines, D.B., Axelrod, B.N., & Theisen, M.E. (1997). Full scale IQ as mediator of practice effects: The rich get richer. *Clinical Neuropsychologist*, *11*, 375-380.

Register-Mihalik, J. K., Kontos, D. L., Guskiewicz, K. M., Mihalik, J. P., Conder, R., & Shields, E. W. (2012). Age-related differences and reliability on computerized and paper-and-pencil neurocognitive assessment batteries. *Journal of athletic training*, *47*(3), 297–305. https://doi.org/10.4085/1062-6050-47.3.13

Rios, J., & Sireci, S. (2014). Guidelines versus practices in cross-lingual assessment: A disconcerting disconnect. *International Journal of Testing*, *14*(4), 289-312. https://doi.org/10.1080/15305058.2014.924006

Rock, D. & Price, I.R. (2019). Identifying culturally acceptable cognitive tests for use in remote northern Australia. *BMC Psychol* 7, 62. https://doi.org/10.1186/s40359-019-0335-7

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement*, *17*(2), 105-116. https://doi.org/10.1177/014662169301700201

Rosselli, M., & Ardila, A. (2003). The impact of culture and education on non-verbal neuropsychological measurements: A critical review. *Brain and cognition*, *52*(3), 326-333.

Rotter, J.B. & Rafferty, J.E. (1950). *Manual: The Rotter Incomplete Sentences Blank: College Form*. The Psychological Corporation.

Rouleau, I., Salmon, D. P., Butters, N., Kennedy, K. C., & McGuire, K. (1992). Quantitative and qualitative analyses of clock drawings in Alzheimer' and Huntington' disease. *Brain and Cognition*, *18*(1), 70–87. https://doi.org/10.1016/0278-2626(92)90112-Y

Scheuneman, J. D., & Grima, A. (1997). Characteristics of quantitative word items associated with differential performance for female and Black examinees. *Applied Measurement in Education*, *10*(4), 299-319. https://doi.org/10.1207/s15324818ame1004_1

Sedo, M. A. (2007) Five Digit Test (Test de los Cinco Digitos). Manual. Madrid, TEA Ediciones.

Serpell, R. & Simatende, B. (2016). Contextual Responsiveness: An Enduring Challenge for Educational Assessment in Africa. *J. Intell.*, *4*(1), 3; https://doi.org/10.3390/jintelligence4010003

Serpell, R. (1993). *The significance of schooling: life-journeys in an African society*. Cambridge University Press.

Shaharban, N.V., Rangaiah, B.m & Thirumeni, D. (2022) Executive control functions and theory of mind among plurilingual adults, Journal of Cognitive Psychology, DOI: 10.1080/20445911.2022.2119989

Shuttleworth-Edwards, A. & Truter, S. (2023). Cross-Cultural Cognitive Test Norms: An Advanced Collation from Africa. Inter-Ed.

Sireci, S. G. (1997). Problems and issues in linking tests across languages. *Educational Measurement: Issues and Practice, 16*(1), 12-19. https://doi.org/10.1111/j.1745-3992.1997.tb00581.x

Sireci, S. G. (2005). Using bilinguals to evaluate the comparability of different language versions of a test. In R. K. Hambleton, P. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 117-138). Lawrence Erlbaum Publishers.

Sireci, S. G., & Allalouf, A. (2003). Appraising item equivalence across multiple languages and cultures. *Language Testing, 20*(2), 148-166. https://doi.org/10.1191/0265532203lt249oa

Sireci, S. G., & Berberoğlu, G. (2000). Using bilingual respondents to evaluate translated-adapted items. *Applied Measurement in Education, 13*(3), 229-248. https://doi.org/10.1207/S15324818AME1303_1

Sireci, S. G., & Wells. C. S. (2010). Evaluating the comparability of English and Spanish video accommodations for English language learners. In P. Winter (Ed.), *Evaluating the comparability of scores from achievement test variations* (pp. 33-68). Washington, DC: Council of Chief State School Officers.

Sireci, S. G., Harter, J., Yang, Y., & Bhola, D. (2003). Evaluating the equivalence of an employee attitude survey across languages, cultures, and administration formats. *International Journal of Testing, 3*(2), 129-150. https://doi.org/10.1207/S15327574IJT0302_3

Sireci, S. G., Patsula, L., & Hambleton, R. K. (2005). Statistical methods for identifying flaws in the test adaptation process. In R. K. Hambleton, P. Merenda, & C. Spielberger, C. (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 93-116). Lawrence Erlbaum Publishers.

Skutnabb-Kangas, T. & Dunbar, R, (2010). *Indigenous children's education as linguistic genocide and a crime against humanity? A global view*. (pp. 1-28). Guovdageaidnu/Kautokeino, Gáldu, Resource Centre for the Rights of Indigenous Peoples.

Smith, A. (1973). *Symbol digit modalities test* (p. 22). Los Angeles: Western psychological services.

Solano-Flores, G., Trumbull, E., & Nelson-Barber, S. (2002). Concurrent development of dual language assessments: An alternative to translating tests for linguistic minorities. *International Journal of Testing, 2*(2), 107-129. https://doi.org/10.1207/S15327574IJT0202_2

Son, J. (2018). Back translation as a documentation tool. *The International Journal for Translation & Interpreting Research, 10*(2), 89–100.

Staios, M., Kosmides, M.H., Nielsen, T.R., Kokkinis, N., Stogoannidou, A., March, E., & Stolwyk, R.J. (in press) The Wechsler Adult Intelligence Scale-Fourth Edition, Greek Adaptation (WAIS-IV GR): Confirmatory Factor Analysis and Specific Reference Group Normative Data for Greek Australian Older Adults.

Storey, J. E., Rowland, J. T., Basic, D., Conforti, D. A., & Dickson, H. G. (2004). The Rowland Universal Dementia Assessment Scale (RUDAS): a multicultural cognitive assessment scale. *International psychogeriatrics*, *16*(1), 13–31. https://doi.org/10.1017/s1041610204000043

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

Suárez, P. A., Marquine, M. J., Díaz-Santos, M., Gollan, T., Artiola i Fortuny, L., Rivera Mindt, M., Heaton, R., & Cherner, M. (2020). Native Spanish-speaker's test performance and the effects of Spanish-English Bilingualism: results from the neuropsychological norms for the US-Mexico Border Region in Spanish (NP-NUMBRS) project. *The Clinical Neuropsychologist*, *35*(2), 453-465. https://doi.org/10.1080/13854046.2020.1861330

Subok, L. (2017). Detecting differential item functioning using the logistic regression procedure in small samples. *Applied Psychological Measurement, 41*(1), 30-43. https://doi.org/10.1177/0146621616668015

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361-370. https://doi.org/10.1111/j.1745-3984.1990.tb00754.x

Tan, Y. W., Burgess, G. H., & Green, R. J. (2021). The effects of acculturation on neuropsychological test performance: A systematic literature review. *The Clinical neuropsychologist*, *35*(3), 541–571. https://doi.org/10.1080/13854046.2020.1714740

Tanzer, N. K., & Sim, C. O. E. (1999). Adapting instruments for use in multiple languages and cultures: A review of the ITC Guidelines for Test Adaptation. *European Journal of Psychological Assessment, 15*(3), 258-269. https://doi.org/10.1027/1015-5759.15.3.258

Thames, A. D., Hinkin, C. H., Byrd, D. A., Bilder, R. M., Duff, K. J., Mindt, M. R., ... & Streiff, V. (2013). Effects of stereotype threat, perceived discrimination, and examiner race on neuropsychological performance: simple as black and white? *Journal of the International Neuropsychological Society*, *19*(5), 583-593. https://doi.org/10.1017/s1355617713000076

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 147- 169). Lawrence Erlbaum Publishers.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning: Theory and practice* (pp. 67-113). Lawrence Erlbaum Publishers.

Tombaugh, T. N. (1996). *Test of memory malingering*. North Tonawanda, NY: Multi-Health Systems.

Tombaugh, T. N., Kozak, J., & Rees, L. (1999). Normative data stratified by age and education for two measures of verbal fluency: FAS and animal naming. *Archives of clinical neuropsychology*, *14*(2), 167-177.

Tsegaye, M.T., De Bleser, R., & Iribarren, C. (2011). The effect of literacy on oral language processing: Implications for aphasia tests. *Clinical Linguistics & Phonetics*, 25: 628-639.Uldis Ozolins, Sandra Hale, Xiang Cheng, Amelia Hyatt & Penelope Schofield (2020) Translation and back-translation methodology in health research – a critique, Expert Review of Pharmacoeconomics & Outcomes Research, 20:1, 69-77, DOI: 10.1080/14737167.2020.1734453

United Nations General Assembly, (1948). *Universal Declaration of Human Rights*. available at: https://www.refworld.org/docid/3ae6b3712c.html [accessed 13 May 2022] van de Vijver, F. J. R., & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. *European Psychologist*, *1*(2), 89-99. https://doi.org/ 10.1027/1016-9040.1.2.89

van de Vijver, F. J. R., & Leung, K. (2000). Methodological issues in psychological research on culture. *Journal of Cross-Cultural Psychology, 31*(1), 33-51. https://doi.org/10.1177/0022022100031001004

van de Vijver, F. J. R., & Poortinga, Y. H. (1991). Testing across cultures. In R. K. Hambleton & J. Zaal (Eds.), *Advances in educational and psychological testing* (pp. 277-308). Dordrecht, the Netherlands: Kluwer Academic Publishers.

van de Vijver, F. J. R., & Poortinga, Y. H. (1992). Testing in culturally heterogeneous populations: When are cultural loadings undesirable? *European Journal of Psychological Assessment, 8*, 17-24.

van de Vijver, F. J. R., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross- cultural assessment. *European Journal of Psychological Assessment, 13,* 29-37. https://doi.org/10.1027/1015-5759.13.1.29

van de Vijver, F. J. R., & Poortinga, Y. H. (2005). Conceptual and methodical issues in adapting tests. In R. K. Hambleton, P. F. Merenda, & C. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 39-64). Lawrence Erlbaum Publishers.

van de Vijver, F. J. R., & Tanzer, N. K. (1997). Bias and equivalence in cross-cultural assessment. *European Review of Applied Psychology*, *47*(4), 263-279.

van de Vijver, F. J., & Leung, K. (2021). *Methods and data analysis for cross-cultural research* (Vol. 116). Cambridge University Press.

Vicente, S. G., Ramos-Usuga, D., Barbosa, F., Gaspar, N., Dores, A. R., Rivera, D., & Arango-Lasprilla, J. C. (2021). Regression-based norms for the Hopkins verbal learning test-revised and the Rey–Osterrieth complex figure in a Portuguese adult population. *Archives of Clinical Neuropsychology*, *36*(4), 587-596.

Vlahou, C. H., Kosmidis, M. H., Dardagani, A., Tsotsi, S., Giannakou, M., Giazkoulidou, A., Zervoudakis, E., & Pontikakis, N. (2013). Development of the Greek Verbal Learning Test: reliability, construct validity, and normative standards. *Archives of clinical neuropsychology: the official journal of the National Academy of Neuropsychologists*, *28*(1), 52–64. https://doi.org/10.1093/arclin/acs099

Wagner, R. K., Torgesen, J. K., Rashotte, C. A., & Pearson, N. A., "Comprehensive Test of Phonological Processing-2nd Ed. (CTOPP-2)." Austin, Texas: Pro-Ed

Wang, R., Hempton, B., Dugan, J.P., & Komives, S.R. (2008). “Cultural Differences: Why Do Asians Avoid Extreme Responses?” *Survey Practice* 1 (3). https://doi.org/10.29115/SP-2008-0011.

Warrington, E.K. & James, M. (1991). The Visual Object and Space Battery Perception. Bury St Edmunds: Thames Valley Company.

Wechsler, D. (2008). Wechsler Adult Intelligence Fourth UK Edition Administration and Scoring Manual. Pearson: London.

Wolf, E.J., Harrington, K.M., Clark, S.L., & Miller, M.W. (2013). Sample size requirements for structural equation models: An evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement, 73(6),* 913–934. https://doi.org/10.1177/0013164413495237

Yi Wen Tan, Gerald H. Burgess & Robin J. Green (2020): The effects of acculturation on neuropsychological test performance: A systematic literature review, The Clinical Neuropsychologist, DOI: 10.1080/13854046.2020.1714740

Zgaljardic, D. J., & Benedict, R. H. (2001). Evaluation of practice effects in language and spatial processing test performance. *Applied neuropsychology*, *8*(4), 218–223. https://doi.org/10.1207/S15324826AN0804_4