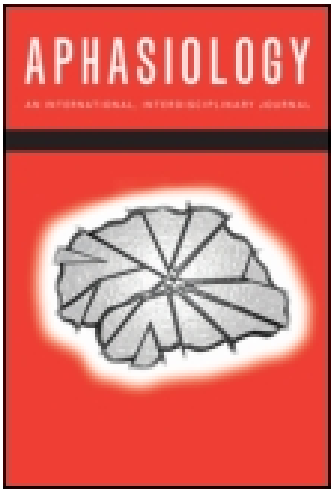

Aphasiology
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/paph20

A tutorial on aphasia test development in any language: Key substantive and psychometric considerations

Maria V. Ivanova [a] & Brooke Hallowell [b]

[a] Neurolinguistics Laboratory, Faculty of Philology , National Research University Higher School of Economics , Moscow 101000 , Russia

[b] Communication Sciences and Disorders , Ohio University , Athens 45701 , OH , USA

Published online: 25 Jun 2013.

To cite this article: Maria V. Ivanova & Brooke Hallowell (2013) A tutorial on aphasia test development in any language: Key substantive and psychometric considerations, Aphasiology, 27:8, 891-920, DOI: 10.1080/02687038.2013.805728

To link to this article: http://dx.doi.org/10.1080/02687038.2013.805728

Aphasiology, 2013
Vol. 27, No. 8, 891–920, http://dx.doi.org/10.1080/02687038.2013.805728

Routledge
Taylor & Francis Group

REVIEW

A tutorial on aphasia test development in any language: Key substantive and psychometric considerations

Maria V. Ivanova[1] and Brooke Hallowell[2]

[1]Neurolinguistics Laboratory, Faculty of Philology, National Research University Higher School of Economics, Moscow 101000, Russia
[2]Communication Sciences and Disorders, Ohio University, Athens 45701, OH, USA

Background: There are a limited number of aphasia language tests in the majority of the world's commonly spoken languages. Furthermore, few aphasia tests in languages other than English have been standardised and normed, and few have supportive psychometric data pertaining to reliability and validity. The lack of standardised assessment tools across many of the world's languages poses serious challenges to clinical practice and research in aphasia.
Aims: The current review addresses this lack of assessment tools by providing conceptual and statistical guidance for the development of aphasia assessment tools and establishment of their psychometric properties.
Main Contribution: A list of aphasia tests in the 20 most widely spoken languages is included. The pitfalls of translating an existing test into a new language versus creating a new test are outlined. Factors to be considered in determining test content are discussed. Further, a description of test items corresponding to different language functions is provided, with special emphasis on implementing important controls in test design. Next, a broad review of principal psychometric properties relevant to aphasia tests is presented, with specific statistical guidance for establishing psychometric properties of standardised assessment tools.
Conclusions: This article may be used to help guide future work on developing, standardising and validating aphasia language tests. The considerations discussed are also applicable to the development of standardised tests of other cognitive functions.

Keywords: Psychometrics; Test development; Standardised testing; Aphasia; Language assessment; Validity; Reliability.

Address correspondence to: Maria V. Ivanova, Neurolinguistics Laboratory, Faculty of Philology, National Research University Higher School of Economics, Ul. Myasnickaya, d. 20, Moscow 101000, Russia. E-mail: mvimaria@gmail.com

This work was supported in part by grant number DC00153-01A1 from the National Institute on Deafness and Other Communication Disorders, an Ohio University Graduate Fellowship in the School of Hearing, Speech and Language Sciences, a College of Health and Human Services Student Research and Scholarly Activity Award, and the Basic Research Program at the National Research University Higher School of Economics.

BACKGROUND

Standardised assessment of aphasia serves to determine important aspects of an individual's language functioning, including whether he or she has aphasia, the severity and characteristics of language impairment, prognosis for language recovery, communicative strengths and weaknesses, target areas for treatment planning and the degree of improvement or regression over time. In aphasia research contexts, standardised assessment is essential for providing reliable and valid quantification of language abilities and enabling comparisons across individuals and groups.

Clinicians and researchers working with English-speaking people with aphasia are privileged to have access to numerous standardised aphasia tests, ranging from comprehensive batteries and general measures of "functional" abilities to tests of specific language domains (see Spreen and Risser (2003) and Patterson and Chapey (2008) for overview of aphasia assessment tools in English). Unfortunately, few standardised aphasia tests exist in the majority of the world's commonly spoken languages other than English. This is largely because specialised clinical care and research foci related to aphasia are in early development states in much of the world. In Table 1, the 20 most widely spoken languages are listed in order according to the estimated number of native speakers of those languages (Lewis, 2009). Aphasia tests developed or translated for use in each language are shown.

As demonstrated in Table 1, amongst those available, most have not been standardised and normed and lack reliability and validity psychometric data. Many aphasia tests in languages other than English are translations of well-known and widely used assessment instruments in English, such as the Boston Diagnostic Aphasia Examination (BDAE; Goodglass, Kaplan, & Barresi, 2001a), the Boston Naming Test (BNT; Goodglass & Kaplan, 2001), the Multilingual Aphasia Examination (Benton, Hamsher, & Sivan, 1994) and the Western Aphasia Battery (WAB; Kertesz, 1982). Some of these tests are mere literal translations of the English-language originals (e.g., Lauterbach, 2006), while others entail thoughtful adaptations that take into account important cultural and linguistic factors associated with the target language (e.g., Benton & Hamsher, 1994; Kertesz, Pascual-Leone, & Pascual-Leone, 1990; Kim & Na, 2001). Notable examples of tests originally developed, standardised and extensively normed in a non-English language are the Standard Language Test of Aphasia (SLTA) in Japanese (SLTA Committee, 1977), the Aachen Aphasia Test (AAT) in German (Huber, Poeck, Weniger, & Willmes, 1983) and the Verb and Sentence Test (VAST) in Dutch (Bastiaanse, Maas, & Rispens, 2000).

As shown in Table 1, novel tests or adapted and translated versions of existing assessment instruments often remain unpublished (e.g., Bhatnagar, n.d.; Tseng, 1993) or have limited circulation (e.g., Kacker, Pandit, & Dua, 1991; Lauterbach, 2006). Sometimes a test is translated for a specific study or for a student's academic project and then is not used again (e.g., Sreedevi, 1991; Tsang, 2000). Several authors refer to specific translated versions of tests but merely cite the English version (e.g., Naeser & Chan, 1980). Some report the use of tests translated into additional languages, but provide no explicit source for these versions, which do not appear to have been normed on native speakers of the target languages (Chenggapa, 2009). Frequently, psychometric data on a test are published in a non-English language, preventing the larger audience of aphasiologists from evaluating the validity and reliability of the assessment instrument (e.g., Tsvetkova, Axytina, & Pulaeva, 1981; Watamari et al.,

TABLE 1
List of the 20 most commonly spoken languages and corresponding available aphasia tests

	Languages	*Aphasia language tests*	*Control norms**	*Norms on people with aphasia***
1.	Chinese (includes Mandarin and Cantonese)	– Bilingual Aphasia Test[1] (•Paradis & Shen, 1987)	No	No
		– Boston Diagnostic Aphasia Examination[2] (•Naeser & Chan, 1980; Tseng, 1993)	?	?
		– Boston Naming Test[3] (Tsang, 2000)	?	?
		– Aphasia Battery in Chinese (Gao, 1996)	?	Yes
		– Chinese Rehabilitation Research Center Standard Aphasia Examination (Zhang, Ji, & Li, 2005)	?	Yes
		– Western Aphasia Battery[4] (Cantonese version, also includes adapted items from Boston Diagnostic Aphasia Examination; Yiu, 1992)	?	Yes
2.	Spanish	– Aphasia Language Performance Scale (•Keenan & Brassell, 1975)	?	?
		– Bilingual Aphasia Test (•Paradis & Ardila, 1989; Paradis & Elias, 1987)	No	No
		– Boston Diagnostic Aphasia Examination (•Garcia-Albea, Sanchez-Bernardos, & del Viso-Pabon, 1986; Pineda et al., 2002; Rosselli, Ardila, Florez, & Castro, 1990)	Yes	?
		– Boston Naming Test (•Allegri et al., 1997; Kaplan, Goodglass, & Weintraub, 1986; Kohnert et al., 1998; Ponton et al., 1992, 1996; Taussig, Henderson, & Mack, 1992)	Yes	?
		– Communicative Abilities in Daily Living (Martin, Manning, Munoz, & Montero, 1990)	Yes	Yes
		– Multilingual Aphasia Examination (•Benton & Hamsher, 1994; Rey, Feldman, Rivas-Vazquez, Levin, & Benton, 1999, 2001)	Yes	Yes (TBI)

(*Continued*)

TABLE 1
(Continued)

	Languages	*Aphasia language tests*	*Control norms**	*Norms on people with aphasia***
		– Psycholinguistic Assessments of Language Processing in Aphasia[5] (•Valle & Cuetos, 1995)	?	?
		– Western Aphasia Battery (•Kertesz et al., 1990)	?	?
3.	English	For extensive lists of standardised aphasia language tests available in English, see Spreen and Risser (2003) and Patterson and Chapey (2008).	–	–
4.	Arabic	– Bilingual Aphasia Test (•Paradis & Abidi, 1987; Paradis & El Halees, 1989)	No	No
5.	Hindi	– All India Institute of Medical Sciences Diagnostic Test of Aphasia (Bhatnagar, n.d.; 1984; Bhatnagar et al., 2002)	?	?
		– Bilingual Aphasia Test (•Paradis & Vaid, 1987)	No	No
		– Boston Diagnostic Aphasia Examination (Kacker et al., 1991)	?	Yes
		– Communicative Abilities in Daily Living (Mahendra, 2004)	Yes	Yes
6.	Bengali	No aphasia tests were found	–	–
7.	Portuguese	– Aachen Aphasia Test[6] (Lauterbach, 2006)	Yes	No
		– Bilingual Aphasia Test (•Paradis & Hub Faria, 1989; Paradis, Simões, & Dillinger, 1987)	No	No
8.	Russian	– Bilingual Aphasia Test (•Ivanova & Hallowell, 2009; Paradis & Zeiber, 1987)	No	Yes
		– Multiple-Choice Test of Auditory Comprehension in Russian (Hallowell & Ivanova, 2009)	Yes	Yes
		– Quantitative Language Assessment in Patients with Aphasia (Kolichestvennaya Ocenka Rechi y Bol'nux s Aphasieu; •Tsvetkova et al., 1981)	Yes	Yes

9.	Japanese	– Bilingual Aphasia Test (•Paradis & Hagiwara, 1987)	No	No
		– Communication Activities of Daily Living (Watamari et al., 1987, 1990)	Yes	Yes
		– Standard Language Test of Aphasia (•Hasegawa et al., 1984; Higashikawa, Hadano, & Hata, 2006; SLTA Committee, 1977)	Yes	Yes
		– Test of Differential Diagnosis of Aphasia (Sasanuma, Itoh, Watamori, Fukusako, & Monoi, 1992)	?	?
		– Western Aphasia Battery (•Sugishita, 1986)	?	?
10.	German	– Aachen Aphasia Test (•Huber et al., 1983)	Yes	Yes
		– Bilingual Aphasia Test (•Paradis & Lindner, 1987)	No	No
		– Lexicon and Morphology Test (Lexikon Modellorientiert; •De Bleser, Cholewa, Stadie, & Tabatabaie, 2004, 1997; Stadie, De Bleser, Cholewa, & Tabatabaie, 1994)	?	?
11.	Javanese	No aphasia tests were found	–	–
12.	Lahnda (includes Panjabi and Seraiki)	– Aphasia Screening Test (Mumby, 1988, 1990)	No	Yes
13.	Telugu	No aphasia tests were found	–	–
14.	Vietnamese	– Bilingual Aphasia Test (•Paradis & Truong, 1987)	No	No
15.	Marathi	No aphasia tests were found	–	–
16.	French	– Bilingual Aphasia Test (•Paradis & Goldblum, 1987)	No	No
		– Boston Diagnostic Aphasia Examination (•Mazaux & Orgogozo, 1982)	Yes	?
		– Boston Naming Test (Roberts & Doucet, 2011; Thuillard-Colombo & Assal, 1992)	Yes	No
		– Examination of Acquired Dyslexias (Examen des Dyslexies Acquises; •Lemay, 1990, 1988)	Yes	?

(*Continued*)

TABLE 1
(Continued)

	Languages	*Aphasia language tests*	*Control norms**	*Norms on people with aphasia***
		– Lille Test of Communication (Test Lillois de Communication; •Delacour, Wyrzykowski, Lefeuvre, & Rousseaux, 2000; Rousseaux, Delacour, Wyrzykowski, & Lefeuvre, 2003)	?	?
		– Montreal–Toulouse Aphasia Battery (•Beland & Lecours, 1990; Beland, Lecours, Giroux, & Bois, 1993; Nespoulos et al., 1992)	Yes	?
		– Picture Naming Test (Test de Dénomination Orale d'Images; •Deloche & Hannequin, 1997; Metz-Lutz et al., 1991)	Yes	?
		– Test for the Examination of Aphasia (Test pour l'Examen de l'Aphasie; •Ducarne, 1989)	?	?
17.	Korean	– Bilingual Aphasia Test (•Paradis & Suh, 1991)	No	No
		– Boston Naming Test (•Kim & Na, 1997, 1999)	Yes	No
		– Korean Aphasia Test Battery Form 1 (based on the Japanese Test of Differential Diagnosis of Aphasia; Park, Sasanuma, Sunwoo, Rah, & Shin, 1992)	Yes	Yes
		– Screening Test for Aphasia and Neurologic Communication Disorders (•Kim, 2009)	Yes	Yes
		– Western Aphasia Battery (•Kim & Na, 2001, 2004)	Yes	Yes
18.	Tamil	– Bilingual Aphasia Test (•Paradis & Devanathan, 1989)	No	No
		– Revised Token Test (•Chenggapa, 2009; Sreedevi, 1991)	?	Yes
19.	Italian	– Aachen Aphasia Test (•Luzzatti, Willmes, & De Bleser, 1996)	?	?
		– Bilingual Aphasia Test (•Paradis, Canzanella, & Baruzzi, 1987)	No	No
		– Boston Naming Test (D'Agostino, 1985)	Yes	No
		– Communicative Abilities in Daily Living (Pizzamiglio et al., 1984)	?	?

	Languages	*Aphasia language tests*	*Control norms**	*Norms on people with aphasia***
		– Clinical Test of Lexical Retrieval and Production (Test Clinici di Ricerca e Produzione Lessicale; Novelli et al., 1986)	Yes	?
		– Italian Battery for Assessing Aphasic Deficits (Batteria per l'Analisi dei Deficit Afasici; •Miceli, Laudanna, Burani, & Capasso, 1994)	?	?
20.	Urdu	– Bilingual Aphasia Test (•Paradis & Janjua, 1987)	No	No

Notes. The content for this table was developed through an extensive review of the literature, Internet searches through multiple search engines and correspondence with colleagues studying aphasia in many countries. Given that many assessment tools are unpublished and that others may only be found through searches in languages the authors do not know, this table is not necessarily exhaustive. Only quantitative tests, for which a specific source could be located, even if it was a reference to an unpublished manuscript or a conference presentation, are listed in this table. Versions of tests that are mentioned in studies/books but without a corresponding source or with a corresponding English version are not included. Also, questionnaires and rating scales are not included.

• before a citation denotes an official separate publication of the test itself; other citations refer to sources that report on the test's construction, development and standardisation. Whenever possible, the latest version of the test is provided as a reference.

*yes = test has been standardised on adult individuals without aphasia, and norms and psychometric data are available; no = not standardised; adult individuals with aphasia, and aphasia norms and psychometric data are available; ? = not possible to determine that based on the available information.

**yes = test has been standardised on adult individuals with aphasia, and aphasia norms and psychometric data are available; no = not standardised; ? = not possible to determine that based on the available information.

[1]The multiple-language versions of the Bilingual Aphasia Test were initially designed to evaluate bilingual patients; however, any single version of the test can be used on its own to assess linguistic functioning in a single language in which other assessment instruments are not available (Paradis, 1987). The BAT is available in over 60 languages (Paradis, n.d.). The BAT is currently out of print. One can obtain copies of the test by writing directly to Dr. Paradis or download the test materials from https://www.mcgill.ca/linguistics/research/bat/

[2]The Boston Diagnostic Aphasia Examination is also available in Finnish (Laine et al., 1993; Laine, Niemi, Koivuselkä-Sallinen, & Tuomainen, 1993), Norwegian (Reinvang & Graves, 1975) and Thai (Gandour, Dardarananda, Buckingham, & Viriyavejakul, 1986).

[3]The Boston Naming Test is also available in Dutch (Marien, Mampaey, Vervaet, Saerens, & De Deyn, 1998), Finnish (Laine et al., 1993; Laine, Koivuselkä-Sallinen, Hännine, & Niemi, 1993), Jamaican (Unverzagt, Morgan, & Thesiger, 1999) and Swedish (Tallberg, 2004).

[4]The Western Aphasia Battery is also available in Hebrew (Soroker, 1997) and Thai (Dardarananda, Potisuk, Grandour, & Holasuit, 1999).

[5]The Psycholinguistic Assessments of Language Processing in Aphasia is also available in Dutch (Bastiaanse, Bosje, & Visch-Brink, 1995).

[6]The Aachen Aphasia Test is also available in Dutch (Graetz, De Bleser, & Willmes, 1992; Willmes, Graetz, De Bleser, Schulte, & Keyser, 1991), English (Miller, De Bleser, & Willmes, 1997; Miller, Willmes, & De Bleser, 2000) and Thai (Pracharitpukdee, Phanthumchinda, Huber, & Willmes, 2000).

1987). In the published papers on assessment instruments, there is great variability in what constitutes a normative sample (of control participants without neurological disorders and people with aphasia) and how comprehensively the psychometric properties of the test have been established. Of course, publication of a test does not guarantee that it has been properly normed and that solid test construction principles have been incorporated (Roberts, 2001). In sum, there are little or no normative data for many non-English-language aphasia tests and most are not available to a wide audience, thus limiting their research and clinical use.

AIMS

The lack of standardised aphasia assessment tools across many of the world's languages poses a serious challenge to clinical practice and research. This tutorial is offered to help clinicians and researchers strategically address this challenge. Readers are given a framework for considering the pros and cons of translating and adapting existing tests for use in other languages or developing new tests in target languages. Then, a review of critical psychometric properties and corresponding statistical analytic methods for quantitative aphasia tests is given. The aim is to provide constructive guidance for clinical researchers aiming to develop, validate and standardise new or translated and adapted aphasia tests. Much of the content is also applicable to the development of standardised tests of other cognitive functions, such as memory, attention and executive functions, and tests of other types of communication disorders. Given its aim, the tutorial also provides a practical review of standardised testing and psychometrics and a summary of key psychometric issue to address, many of which are neglected in existing assessments. Authors of new tests are encouraged to consult additional resources on test development and statistics, while making sure to consider all aspects mentioned here. Highly recommended are the works by Allen and Yen (2002); Anastasi and Urbina (1997); Fishman and Galguera (2003); Franzen (2003); Mitrushina, Boone, Razani, and D'Elia (2005); Schiavetti and Metz (2002), and Strauss, Sherman, and Spreen (2006).

CONSTRUCTION OF AN APHASIA LANGUAGE TEST IN A NEW LANGUAGE

There are two general approaches to developing a new test: an existing standardised test in another language can be translated into the target language or a new original test may be developed. Below we examine each approach and outline its advantages and challenges. Then we describe factors in test design and administration procedures that should be addressed irrespective of the approach adopted.

Considerations in translating an existing test

The translation of an existing test may at first appear easier and more efficient than developing a new test. However, this approach has many caveats. 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). Therefore, verification of a new version through back-translation (when the new language version is translated back into the original language and compared with the initial version of the test; Brislin, 1970) is not entirely possible or appropriate because of

differing aspects of the original and the target languages, including rules of phonology, morphology and syntax (Bates, Wulfeck, & MacWhinney, 1991; Paradis, 1987).

It is important to control for potential confounding factors when translating a test from a different language. Psycholinguistic controls that may have been implemented in the design of a test developed in one language do not necessarily apply in the new version in a different target language. These 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 and cultural relevance (Edwards & Bastiaanse, 2007; Ivanova & Hallowell, 2009; Lorenzen & Murray, 2008; Murray & Clark, 2006; Roberts, 2008; Roberts & Doucet, 2011). Also, where appropriate, attention to other factors, such as verb tense, aspect, mood, noun case and gender, and type of script should be controlled. In some cases, a test is being developed so that it may be administered across multiple languages; this is important for comparing individual multilingual speakers' abilities across varied languages. Consistently implementing controls for myriad linguistic factors across all target languages is challenging and requires strong teamwork among native speakers of each of the target languages who are also experts on language and aphasia.

It is important to recognise that even when careful adaptations are made and parallel linguistic controls are implemented, culturally linked attitudes of individuals about types of tasks (e.g., question-answer, multiple-choice) and means of indexing responses (e.g., correct/incorrect scoring) may influence performance. Further, the very act of clinical testing may not be as common in the cultures of speakers of target language as it is in cultures associated with the original language (Baker, 2000; Lezak, Howieson, & Loring, 2004). Likewise, the pragmatic roles of tester and test-taker, often a younger person who "knows" the correct answers testing an older person who may feel a loss of face in having his or her deficits emphasised, are especially problematic in some cultural contexts. Such cases require especially thoughtful consideration of the validity, interpretation and generalisation of results.

Even when a test has been linguistically and culturally adapted for use in a different target language, further standardisation and norming are essential before one may draw valid conclusions about an individual's language abilities (Bates et al., 1991; Edwards & Bastiaanse, 2007; Ivanova & Hallowell, 2009; Roberts & Doucet, 2011). Norms developed in one language are not applicable to an individual's performance on the test in a different language (Kohnert, Hernandez, & Bates, 1998; Manuel-Dupont, Ardila, Rosseli, & Puente, 1992). Individuals with comparable severity and analogous aphasia types might exhibit differing patterns of performance depending on the psycholinguistic characteristics of their native language (Bates et al., 1991). An identical score or change in scores on the same test in two different languages is not necessarily interpretable in the same manner. Therefore, for test results to be interpreted appropriately the test's reliability and validity must be substantiated anew. In sum, given the challenges of test translation and adaptation and given that normative data must be established in the translated version, development of an entirely new test in the target language is sometimes a preferable alternative.

Considerations in creating a new test

Creation of a new test begins with thoughtful consideration of what constructs are to be assessed and why, and what types of individuals are to be assessed and why. When deciding on the purpose of the test, one must delineate the decisions/inferences

that will be made based on assessment results and define the reference population (Wolfe & Smith, 2007a). Clearly articulating the test's purpose from the beginning helps ensure that further steps (item selection, task structuring, determination of scoring and scaling models, and so forth) are all in accordance with and serve its main purpose.

Selection of subtest and test item types is motivated by a certain framework for conceptualising aphasia as well as by the need to sample specific types of linguistic behaviours for particular purposes (e.g., Howard, Swinburn, & Porter, 2010; Marshall & Wright, 2007). For instance, providing a baseline for detection of change over time requires measurement of performance across several similar items in the same domain. In contrast, determining type of aphasia and areas of strength and weakness across language modalities requires diverse items from varied domains. A general aphasia test should address the abilities to comprehend and produce linguistic content (semantics), form (phonology, morphology and syntax) and the ability to use language appropriately in context (pragmatics) (Patterson & Chapey, 2008). It should also include items of varying difficulty (Goodglass, Kaplan, & Barresi, 2001b) so that it will be sensitive in indexing impairments of varying severity levels. General language functions to be covered by a comprehensive aphasia test include auditory comprehension, repetition, automatic speech, naming, spontaneous speech, reading and writing (Goodglass et al., 2001b; Kertesz, 2007b; Murray & Clark, 2006). A brief summary of clinical symptoms and types of test items corresponding to these seven domains is provided in Table 2.

Beyond the investigation of language domains, test developers should consider the importance of assessing nonverbal aspects of aphasia and whether these are to be assessed by the test they are developing versus through additional recommended assessments. There is growing evidence that cognitive nonlinguistic deficits in aphasia interact with and tend to exacerbate the language impairment (Helm-Estabrooks, 2002; Hula & McNeil, 2008; McNeil, Odell, & Tseng, 1991; Murray, 2004; Wright & Fergadiotis, 2012). When assessing language abilities, clinicians ideally should perform a screening of attention, memory and executive skills (Connor, MacKay, & White, 2000). Additionally, it is important that clinicians assess the impact of cognitive nonlinguistic variables on language processing in aphasia (Martin, Kohen, & Kalinyak-Fliszar, 2010; Murray, 1999, 2004). One of the main challenges in assessing cognitive nonlinguistic deficits in persons with aphasia is that most cognitive tasks require some level of verbal processing. Therefore, performance on these tasks might be confounded by individuals' receptive and expressive linguistic deficits and provide a distorted picture of their cognitive nonlinguistic strengths and weaknesses (Hallowell, Wertz, & Kruse, 2002; Ivanova & Hallowell, 2012; Odekar, Hallowell, Lee, & Moates, 2009).

Despite mounting evidence of concurrent nonlinguistic impairments in aphasia, assessment of cognitive functions has not become an integral part of standardised aphasia assessment instruments, and aphasia tests tend to have a strong psycholinguistic orientation. The Comprehensive Aphasia Test (Swinburn, Porter, & Howard, 2004) is designed to screen for associated cognitive deficits (visual neglect, semantic memory, word fluency, recognition memory, gesture use and arithmetic skills). However, these screening tools are intended primarily to minimise potential confounds of test administration, not to investigate concurrent cognitive deficits. The Cognitive Linguistic Quick Test (Helm-Estabrooks, 2001) is another test enabling evaluation of cognitive strengths and weaknesses, although it is not designed

TABLE 2
Language functions, corresponding test items and factors to be controlled

Language functions and associated deficits	*Test items*	*Linguistic/cognitive factors to be controlled in test design*
Auditory comprehension: – Phonemic and semantic comprehension errors – Receptive agrammatism	– Lexical decision (word vs. non-word discrimination) – Selection of a multiple-choice image corresponding to a verbal stimulus – Commands – Yes/no questions – True/false statements – Questions following a story (complex ideational material) – Story retell tasks – Spontaneous conversation	– Word frequency – Word familiarity – Noun case and gender – Age of acquisition – Imageability – Concreteness/abstractness – Word, phrase, sentence length – Phonemic complexity – Grammatical complexity (e.g., semantically constrained vs. not; canononicity; clausal types, verb tense, mood) – Plausibility of content
Reading: – Reduced reading ability (dyslexia/alexia)	– Matching cases/script/numbers – Copying letters, words, phrases, sentences – Orthographic lexical decision – Reading aloud – Word/sentence/paragraph reading with picture matching – Paragraph/text reading with comprehension questions	– Script, font – Word frequency – Word familiarity – Noun case and gender – Age of acquisition – Imageability – Concreteness/abstractness – Word, phrase, sentence length – Phonemic composition and articulatory difficulty – Grammatical complexity – Plausibility of content

(*Continued*)

TABLE 2
(Continued)

Language functions and associated deficits	*Test items*	*Linguistic/cognitive factors to be controlled in test design*
Repetition: – Inability to repeat – Inaccurate repetition – Perseveration	– Repetition of phonemes, words (nonsense words, single words, series of words), phrases, sentences	– Phonetic/phonemic composition and articulatory difficulty – Word, phrase, sentence length – Grammatical complexity – Grammatical and semantic – Plausibility
Automatic speech: – Limited automatic (rote, highly learned) speech – Perseveration	– Recitation of automatic sequences (numbers, days of the week, months) – Recitation of nursery rhymes, poems, songs – Spontaneous automatic utterances during conversation	– Articulatory difficulty – Familiarity of rote sequences
Naming: – Word retrieval difficulties – Paraphasias (literal/phonemic, semantic/global) – Perseverations – Circumlocution	– Confrontation naming – Word descriptions/definitions requiring naming response – Cloze sentences or phrases – Word fluency (i.e., produce words starting with a certain letter or within a given semantic category)	– Word frequency – Word familiarity – Age of acquisition – Imageability – Concreteness/abstractness – Phonemic composition and articulatory difficulty – Word length – Semantic category – Visual/tactile stimulation – Real objects versus images – Degree of control of physical stimulus properties of images

Spontaneous speech: – Expressive agrammatism, telegraphic speech – Dysnomia/anomia – Paraphasias – Perseverations	– Picture description – Conversation/discussion of a topic (assessment of expressive speech greatly depends on selected performance measures—rating scales)	– Degree of conversational structure and support – Topic complexity – Topic familiarity and personal significance – Relationship to conversational partner
Writing: – Reduced writing ability (dysgraphia/agraphia)	– Letter matching – Writing of words, phrases, sentences to dictation – Copying – Written picture naming – Narrative writing	– Word frequency – Word familiarity – Age of acquisition – Imageability – Concreteness/abstractness – Word, phrase, sentence length – Phonemic complexity – Regular vs. irregular words – Complexity of target text – Topic familiarity and personal significance – Type of writing instrument/keyboard

specifically for use with people who have aphasia. Additional standardised tools to assess cognitive functions for people with aphasia are needed.

The extent to which cognitive functions will be screened in an aphasia test depends upon the framework for conceptualising aphasia that is adopted by the authors (Hallowell & Chapey, 2008). For instance, Goodglass et al. (2001b) defined aphasia as a "disturbance of any or all of the skills, associations, and habits of spoken and written language" (p. 5). This purely language-oriented view of aphasia led to the selection of strictly language-focused tasks for the BDAE (Goodglass et al., 2001a). In contrast, one might propose taking a more cognition-oriented approach. For example, a test developer who regards language impairment in aphasia as intrinsically tied to the impaired short-term memory would emphasise indexing of short-term memory by manipulating memory load in the context of various language tasks (e.g., Martin et al., 2010).

Controlling confounding factors in test design

A confounding factor is any aspect of a verbal or visual stimulus or testing task that may influence test performance in any way that is not directly related to the language ability being assessed. For example, when assessing single-word comprehension, a person's prior experience with a word or concept would influence his or her ability to point to a correct image in a multiple-choice task. Thus, word familiarity would *confound* an assessment of the degree of comprehension deficit owing to aphasia. Attributing incorrect responses to aphasia when something other than aphasia has led to the incorrect response is invalid. Cognitive and linguistic controls for multiple potentially confounding factors should be implemented in the test design process. A summary of such factors is provided in Table 2. It should be kept in mind that many original tests that might be used as a basis for translation were not necessarily designed with careful attention to the need for such controls.

Just as there are specific cognitive and linguistic factors to be taken into account in test item design, there are also key factors to consider in the development of test stimuli and test procedures in general. Given that problems of motor control, speech, visual acuity, colour perception, hearing acuity, auditory processing, attention and memory are all common concomitant factors in the target population (Hallowell, 2008; Heuer & Hallowell, 2009; Lezak et al., 2004; Murray & Clark, 2006), it is important to take these into account to the greatest degree possible during test development, through stimulus design, test administration instructions or both. Otherwise, concomitant impairments are likely to confound assessment and leave room for alternative explanations of results. For instance, it should be assured that visual stimuli are presented clearly and are perceptible for individuals with decreased visual acuity. Type of font, size of print, competing visual stimuli within the test displays or arrays of objects to be presented, and colour and other physical properties of visual stimuli can all impact performance of individuals with aphasia (Arditi & Cho, 2007; Brennan, Worrall, & McKenna, 2005; Hallowell, 2008; Heuer & Hallowell, 2007, 2009). Likewise, clarity, intensity, rate and fundamental frequency of auditory stimuli should be taken into account when such stimuli are pre-recorded as part of the test; otherwise, these should be addressed explicitly in test administration instructions.

In general, when controlling for potentially confounding factors, it is important to minimise non-linguistic cognitive requirements of language tasks, such as reliance on remembering task instructions, intact visual-spatial processing abilities, and so forth.

It is crucial for authors to acknowledge overtly non-linguistic demands within linguistic tasks that cannot be eliminated, such as auditory and visual acuity, focused attention and so forth. This will help cue future consumers of the test to screen their examinees for these prerequisite abilities before test administration and, in the case of identified deficits, will help to provide alternative explanations for performance errors.

Standardising test administration

The term "standardise" has two meanings in the assessment context. In a broad sense, the term refers to the collection of psychometric data on a test. In a narrower sense, it refers to a uniform administration protocol from patient to patient and from one examiner to another. This includes consistent use of specific and concrete instructions on how the various tasks that compose the test should be presented to an individual, how repetition and prompting may be used, how responses are to be scored and how any qualitative judgments are to be made. It also includes instructions on required screenings and ratings of premorbid proficiency in each language of a multilingual speaker; recommendations for taking into account non-linguistic cognitive, perceptual and motor deficits in test administration and interpretation; reminders to control the test environment by eliminating background noise, ensuring appropriate lighting and controlling extraneous visual clutter; and suggestions for considering such factors as depression, fatigue, anxiety, concomitant illness and pain in determining optimal testing times. Thorough instructions are important for ensuring consistent administration, which minimises inter- and intra-examiner differences.

Some standardised tests include ceiling/floor rules in the description of administration procedures. These rules make test administration shorter by helping clinicians assess individuals with aphasia on relevant items. A ceiling rule indicates the number of items that may be answered incorrectly within a given task or subtest before the clinician discontinues testing on that particular task or subtest. For instance, a ceiling rule might state that if an individual incorrectly names five objects in a row, then testing should continue with a different task of naming actions and the rest of the items in the object naming tasks should be marked as incorrect. A floor rule guides the decision of which items to start testing, so that time is not spent on items that are too easy. If ceiling/floor rules are not provided, the examiner should not arbitrarily skip items that intuitively seem too easy or too difficult for the individual being tested. This would violate the standardised structure of the test and render an individual's results incomparable to the normative group.

Once test items are well developed (translated or created anew in a target language) based on solid practical and theoretical motivations and once a standardised protocol is established, it is then important to administer the test to large samples of individuals with and without aphasia so that the psychometric properties of the test may be studied. Key areas to address are norm referencing, validity, reliability and item analysis.

PSYCHOMETRIC PROPERTIES OF APHASIA LANGUAGE TESTS

Norm referencing

It is important that any test for use with people with language impairments be thoroughly normed on a large sample of people with no history of neurological disorder

so that appropriate cut-off scores for "normal" performance may be determined. It is desirable for normative groups to include at least 100 individuals (Franzen, 2003). An assumption that people without language impairment should perform errorlessly on such a test is not acceptable as there is variability in "normal" linguistic performance, and individuals without neurogenic language impairments do not always obtain perfect scores on existing aphasia batteries (Ross & Wertz, 2004). Furthermore, performance on language tests is often influenced by age, education and socio-economic status (Lezak et al., 2004; Mitrushina et al., 2005). Thus, the normative sample should be comparable in terms of each of those three factors to the clinical group for whom it is intended. For example, it is inappropriate to norm an aphasia test exclusively on college students. Also, normative results can be used to evaluate item validity. If a sufficient number of participants without aphasia respond incorrectly to an item and the score on that item stands out from the distribution of scores on most items, then it is advisable to re-evaluate the item for potential confounds (e.g., poor wording, ambiguous stimuli, culturally inappropriate material) and possibly revise or eliminate it.

An additional set of normative data should be based on a sample representative of the population for which the test is intended. For example, an aphasia test should be normed on people with various types of aphasia and various levels of severity of language impairment, as well as people without neurological disorders. Relevant features (e.g., age, education level, socio-economic status, time post onset, site of lesion and concomitant medical conditions) of the standardised sample should be described in detail. Normative clinical data allow the examiner to interpret an individual's raw test score by considering where his or her performance lies in the distribution of scores. This also helps to associate that individual's performance with a certain clinical profile. For this purpose, raw scores of the clinical normative samples are ideally transformed into standardised scores, the most common of which are percentiles and Z-scores. Diagnostic percentiles allow clinicians to determine the percentage of people from a particular group who perform better or worse than any particular individual. For example, if an individual's score corresponds to the 65th percentile, then 65% of all people in the normative sample (the group used for comparison) scored lower on the test.

A Z-score shows how many standard deviations above or below the mean of the normative group an individual is performing. The following formula is used to calculate a Z-score:

$$Z = \frac{x - \mu}{\sigma}$$

where x is the individual's raw score, μ is the population mean and σ is the population standard deviation. For example, if the mean of a normative population is 0.6 with a standard deviation of 0.2 and the raw score for a given individual is 0.4, then the Z score for that individual is $(0.4 - 0.6)/0.2 = -1$. If the Z score is negative, then it indicates that the individual scored this many standard deviations *below* the mean of the normative group.

Again, for standard scores to be reliable, it is typically desirable for the clinical normative sample of a test to include 100 or more individuals. In some cases, it is also advisable to collect normative data on individuals who have brain injury but no language impairment (see below discussion on criterion validity; Kertesz, 2007b).

A possible alternative to norm referencing is criterion or domain referencing (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999; McCauley, 1996), which is used to compare an individual's performance to a specific functional performance standard or body of knowledge. While norm referencing permits interpretation of how an individual performs compared to other individuals within a similar diagnostic category, criterion referencing enables inferences regarding what skills individuals with aphasia have lost and to what extent. In clinical practice, criterion-referenced tasks are often implemented for determining treatment candidacy and tracking progress during intervention. While criterion referencing is not typically implemented in standardised aphasia batteries, clinicians and researchers may implement their own criterion references for specific applications of test results (e.g., for determining inclusion in a certain study or for meeting treatment goals). For example, for determining that a treatment goal has been met, pre-determined criterion scores for verbal fluency might be set through performance on the picture description task from the WAB-Revised (Kertesz, 2007a) or for naming on the BNT (Goodglass & Kaplan, 2001). In the current overview, we limit our discussion to norm-referenced tests.

Reliability

The term reliability refers to the consistency, stability and accuracy of a measuring instrument. Although reliability is related to the quality and the soundness of the test, it is also influenced by the domain and the population being measured. For example, people with aphasia show greater variability in behavioural performance compared to individuals without neurological disorders (Strauss et al., 2006); hence, there are always limitations to how reliable any aphasia test may be. Three types of reliability are particularly relevant for aphasia tests: internal consistency, test–retest stability and inter-rater reliability.

Internal consistency (or *internal reliability*) reflects the constancy of results across items within a subtest or the overall test. It is usually estimated through a procedure called split-half reliability, with a larger correlation indicating higher reliability. For dichotomous items (e.g., right/wrong), a suitable correlation to use is Kuder-Richardson (KR-20). If the items are continuous in nature, then Cronbach's alpha is more appropriate. Each of these correlation coefficients can be interpreted as the mean of all possible split-half coefficients and therefore is preferred to simply correlating results from two halves of the test. As a general rule, a reliability coefficient over .7 is considered an indication of sufficient internal consistency (Allen & Yen, 2002; Strauss et al., 2006).

For a subtest/test to have sound internal reliability, it should have an adequate number of items. Including too few items can make performance easily prone to influence of fluctuations in extraneous factors. Such factors may be internal to the test taker (e.g., distraction, fatigue or motivation to participate, which are often inconsistent for a given individual) or related to fluctuations in the probability that a given person could respond well across linguistic items even without a language deficit (e.g., due to familiarity with constructs being assessed). All such factors make measurement unreliable. Conversely, including too many items, although it may lead to higher internal reliability, is generally impractical because of increasing fatigue and limits on available testing time. It is recommended that a pilot version of a test include more items than one intends to include in the actual test; based on analyses of results from

the larger set, the number can be reduced while still ensuring high internal reliability. Additionally, the number of items will depend on the aims of the test—whether it is to sample a broad range of behaviours or index a specific domain in detail. Item reliability and discriminability parameters can also be helpful in the quest for optimal subtest length (see section on Item Analysis).

Split-half reliability is important for estimating the standard error of measurement (SEM), another important psychometric property. SEM is an estimate of the error inherent in interpreting an individual's test score. In other words, as the obtained score does not necessarily reflect the individual's true score on the test, but rather an estimate of it, the standard error of measurement reflects how accurate and precise that estimate is likely to be (Mitrushina et al., 2005). Standard error of measurement is calculated according to the following formula:

$$\mathrm{SEM} = \mathrm{SD}\sqrt{1 - r_{xx}}$$

where SD is the standard deviation of all the scores in the sample and r_{xx} is the reliability coefficient for the test. The higher the reliability of a test, the smaller the standard error of measurement is.

Test–retest reliability is a measure of the stability of measurement across time. It requires administering the same test to the same group of participants twice, within a time period during which the construct being measured is unlikely to change. It is calculated by correlating the two sets of scores. The higher the correlation coefficient, the stronger the reliability of the instrument is. When testing a person with the same instrument twice, especially if the time between sessions is brief, learning effects should be taken into consideration. On the other hand, the longer the wait between the test and the retest, the greater the likelihood that the skills and abilities under examination might change, thereby confounding test–retest reliability. This is especially relevant for individuals with aphasia, whose language abilities are prone to variation because of a number of factors. For these reasons, two similar versions of the same test can be used to estimate test–retest reliability; in this case, item variance should be taken into account as well. Unfortunately, few aphasia tests have multiple forms available for this purpose.

Inter-examiner reliability is a measure of how much the examiner influences participants' performance on a test. It is established by correlating results obtained by the same individuals when the test is administered and scored by two different examiners. As mentioned earlier, detailed presentation and scoring guidelines enhance inter-examiner reliability.

Validity

Validity refers to the degree to which a test actually measures what it purports to measure such that conclusions drawn from performance on that test are appropriate. Tests in and of themselves are not valid or invalid; it is the inferences made from them that can be evaluated for validity (Franzen, 2003). Therefore, the issue of validity should be approached from a perspective of the questions one hopes to answer using the results of the test. Several types of validity are of particular relevance to aphasia batteries. These include face, content, criterion and construct validity.

Face validity is based on a subjective judgment by individuals considering test results that there is a match between the purpose of the test and its actual content. For example, one may be interested in face validity according to test administrators, test-takers or readers of articles describing tests administered in research protocols.

Content validity refers to the adequacy of sampling from the domain of the construct to be measured. Ideally, items of a test should be representative of the constructs under study. For instance, a subtest on reading comprehension should include items that target various aspects of reading comprehension, rather than evaluate breadth of general knowledge, verbal reasoning or ability to produce coherent speech. The importance of attending to this seemingly obvious tenet was highlighted by Nicholas, MacLennan, and Brookshire (1986), who showed that people with and without aphasia who had not even read passages taken from published aphasia tests were still able to answer correctly more than half of the corresponding "comprehension" questions at well above chance levels. Establishing content validity entails detailed conceptual item and task analysis. The content of each construct being tested should agree with content areas as defined by other researchers in the field. While establishment of content validity is based primary on rational arguments, statistically it can be partially supported through investigation of item discriminability, described below under Item Analysis.

Criterion validity reflects how well the score on the test can predict a certain outcome (criterion). Criterion validity can be quantified as the correlation coefficient between scores on the test under study and an external measure. Depending on the type of criteria chosen, several subtypes of criterion validity may be distinguished. *Concurrent validity* reflects the relationship between one or more specific criteria and scores from the test under development. The criterion may entail a score on a previously validated test or another objective measure (e.g., presence of language impairment). *Predictive validity* indicates how well the test scores can predict future events (e.g., a person's language ability several months following test administration). In aphasia language testing, predictive validity is demonstrated less frequently than concurrent validity.

As mentioned above, one of the means of establishing concurrent validity is to investigate the correlation between a new test and another test of known validity (termed by some as convergent validity; see Mitrushina et al., 2005). A strong relationship between two tests implies that one is measuring the same construct as the other. However, the initial motivation for developing a test may be that there is no instrument available for assessment of a specific ability in a certain population or that the existing tools in and of themselves have questionable validity and reliability; in such cases, a high degree of concurrent validity is not necessarily the test author's goal. Another frequently used method for demonstrating concurrent validity is using presence of a language impairment as the criterion. In this instance, the degree to which the test distinguishes performance of individuals with aphasia from that of individuals without neurological impairment indicates the test's concurrent validity. Statistically this is verified by a significant difference in language scores between a sample of people with aphasia and a control group without neurological deficits. However, it is important to recognise that differences in performance between the two groups may be attributable to a number of factors (e.g., concomitant cognitive and perceptual problems) that may be relevant but not central to the language deficit the examiner wishes to study. To increase confidence that the differences in performance are associated with distinctions relevant to the nature of the disorder being assessed, other groups should

be used to establish criterion validity of a test as well. Ideally, a language test should distinguish people with aphasia not only from controls without neurological impairment, but also from individuals with neurological disorders but without language impairment (Kertesz, 2007b; Spreen & Risser, 2003).

As a part of the examination of criterion validity for a categorical criterion, other indices can be considered. When examining an instrument's ability to distinguish normal from impaired performance, even though the mean scores of the two groups might be significantly different, there is still likely to be some overlap between them. Thus, when making a clinical decision based on test results, it is important to consider the degree of overlap (the proportion of individuals with aphasia scoring at or above the minimum expected score for people without aphasia) and/or the index of determination (the degree to which being diagnosed with aphasia compared to having no aphasia predicts performance on a test) (Ross & Wertz, 2003). Psychometric properties of a test such as sensitivity (the percent of individuals with aphasia who perform below a cut-off score for normal performance) and specificity (the proportion of individuals without aphasia who obtain results above the cut-off for normal language abilities) are also valuable when evaluating the likelihood that a test score leads to a correct interpretation (Ross & Wertz, 2004). The cut-off for performance of people with aphasia should be established by simultaneously considering the sensitivity and specificity of a test (Pepe, 2003; Strauss et al., 2006). This may be done simply by examining sensitivity and specificity information in graphical or tabular format and selecting a score that provides an optimal balance between false positives (suggesting that there is a deficit that is not there) and false negatives (failing to detect a deficit that is actually present), thus enhancing criterion validity of the instrument.

Construct validity represents the test's ability to index the underlying theoretical construct it is intended to measure. It is a determination of whether the created measure behaves the way it should according to the theory in which it is grounded. Construct validity is supported by an accumulation of theoretical rationale and empirical evidence. Theoretical considerations include consideration of how accurately the specific operationalisation (items and tasks of the test) matches the underlying construct. Empirical demonstration of *convergent* validity (indexed via high correlations with other tests measuring the same construct) and *discriminant* validity (indexed via low correlations with tests measuring other constructs) can be used as supporting evidence for construct validity (Mitrushina et al., 2005). Additionally, factor-analytic statistical techniques are used to show whether the subtests in a given battery each contribute to one or more major factors that represent language/cognitive functions consonant with the conceptual framework behind the test. At the early stages of instrument development, all subtest scores may be entered into exploratory factor analysis. This analysis can be used to examine which factors account for the most variance in distribution of subtest scores and whether scores from different language functions influence (or "load onto") different factors or just a single severity-related factor. Then follow-up hypothesis-driven confirmatory factor analysis performed on a new sample of data may provide a more in-depth evaluation of the goodness of fit of the specific model of language processing to the obtained pattern of results. Other options for examining construct validity might be studies demonstrating differences between clearly differing clinical groups or between pre- and post-test data collected during an aphasia intervention study.

Recently, there has been a theoretical shift from distinguishing diverse types of validity to viewing validity as a unified construct supported by multiple levels of

convergent evidence (Messick, 1995; Strauss et al., 2006; Wolfe & Smith, 2007a). That is, "various sources of evidence may illuminate different aspects of validity, but they do not represent distinct types of validity. Validity is a unitary concept. It is the degree to which all accumulated evidence supports the intended interpretation of test scores for the proposed purposes." (American Educational Research Association et al., 1999, p. 11). In the test development process, all aspects of validity, whether considered part of a common construct or not, should be taken into account, as they each make a unique and a significant contribution to the adequacy of interpretations and actions based on scores of the test instrument under development. Furthermore, test validation is an ongoing process; data collected with the instrument may contribute to further refinement of its psychometric properties, provide guidance for test revisions or help delineate the test's limitations.

Item analysis

It is also crucial to investigate an assessment instrument at the level of its individual items. It is important to consider three interrelated but different aspects of test items: item difficulty, item reliability (consistency) and item discriminability.

Item difficulty for dichotomously scored (e.g., right/wrong) items is regarded as the percent of participants who respond correctly to the item. Ideally, the distribution of item difficulty should be close to normal, and the range should be from near 0% to about 100% (or to the cut-off for normal performance). Such a distribution will ensure that the test includes easy and difficult items (assuming that other sources of variation are eliminated through analyses of item consistency, as discussed below). Items of varying degrees of difficulty are critical for making subtle distinctions among participants (Goodglass et al., 2001b). Item difficulty should be evaluated based on a representative sample of people with aphasia. A representative sample in this case depends on the instrument being validated. If a test is designed to evaluate general language performance, then item difficulty should be examined based on performance of people with various types and severity of linguistic deficits; if it is designed for assessment of severe aphasia, then results of a group of individuals with severe aphasia should be examined.

Item consistency, or *item reliability*, is evaluated by recomputing the instrument's (or subtest's) reliability *n* times, deleting a different item from the instrument/subtest each time; this is called alpha-if-item-deleted (Fishman & Galguera, 2003). When deleting a particular item causes the overall reliability of the instrument/subtest to rise, then the item is considered to be a poor item, as the instrument becomes more consistent without it. On the other hand, if deletion of an item causes the overall instrument/subtest reliability to fall, then the item is desirable, since it contributes to the instrument's reliability. A second approach to estimating a test's reliability at the item level is based on determining each item's inter-item consistency, which is a measure of how this item relates to other items on the test. An item's average correlation with every other item in the instrument/subtest is calculated. Items with extremely high intercorrelations (over .8) should be examined for redundancy. If the item is also highly correlated with the overall score and content analysis reveals its similarity to other items, then it may be reasonable to discard the item. Items that correlate poorly with the rest of the test should be further considered in terms of potentially confounding factors, such as cultural relevance and familiarity or physical stimulus characteristics.

Item discriminability is an item validity estimation procedure. It is computed through corrected item total correlation, which is accomplished by correlating scores on a particular item to the total score minus the score on that item. A bell-shaped curve with a mean of .60 to .80 is desirable for this statistic (Fishman & Galguera, 2003). A moderate to high correlation with the overall score reflects that the item measures the same underlying construct as that reflected by the instrument overall. Special attention should be paid to items that demonstrate a high correlation (around .9) with the overall score, as this might be an indication of the redundancy of that particular item. Item reliability and discriminability parameters guide researchers in selecting the optimal number of items for a given subtest. Both item reliability and item discriminability for an aphasia language test should be evaluated based on data from people with aphasia.

Item analysis can also be done using Rasch or more general item response theory models, which are becoming increasingly popular in test development (Bond & Fox, 2007; Hula, Doyle, McNeil, & Mikolic, 2006). Rasch analysis models the relationship between person ability, item difficulty and response probability. It permits examination of goodness of fit of individual test items in relation to the construct being measured and provides a justification for regarding the generated test scores as interval-level measures. It helps ascertain that all test items within a subtest measure a single difficulty dimension; this allows detection of items that poorly discriminate the construct being measured relative to the summary discrimination of all items. Only recently Rasch analysis has started to be used as a statistical foundation for development of new aphasia tests (Hula et al., 2010) and to investigate aspects of validity in existing assessment instruments (Hula, Donovan, Kendall, & Gonzalez, 2009; Hula et al., 2006; Willmes, 1997). An in-depth account of instrument development based on Rasch modelling is beyond the scope of this tutorial; for further information on the topic, see Baylor et al. (2011), Bond and Fox, Embretson and Reise (2000), and Wolfe and Smith (2007a, 2007b).

A summary of steps in developing a standardised quantitative aphasia test along with corresponding psychometric indices is presented in Table 3.

CONCLUSIONS

Thorough psychometric evaluation of a test is effortful and time consuming, requiring data to be collected on large samples of participants and then meticulously analysed. However, this should not daunt prospective researchers. First, this work can be accomplished in several steps, with information gained at every stage contributing to a valuable empirical evidence base for professionals using the assessment tool. Second, the process of data collection can become more manageable through a multi-centre collaboration, where several teams in different settings work together on gathering data pertaining to a certain assessment instrument.

Once relevant psychometric properties of the test have been established, it is important to publish the test. Otherwise, professionals unacquainted with the authors have little or no access to the test materials and its psychometric data. Also, other professionals cannot evaluate content, reliability or validity of unpublished tests used in research or clinical evaluations. This undermines the quality of the research and the generalisations that can be made based on research results. Additionally, it is important to publish reports on test development and standardisation in international journals that are accessible to a wide audience. While such publications might not

TABLE 3
Requirements for a standardised aphasia test

Requirement	*Related psychometric indices*
1. Standardised administration and scoring	Inter-examiner reliability Test–retest reliability
2. Items covering intended domains of language functioning	Face validity Content validity Item discriminability Construct validity
3. Discrimination between relevant characteristics of impairment	Content validity Construct validity
4. Items of varying difficulty	Item difficulty index
5. Sufficient number of items for stability of measurements	Internal consistency Test–retest reliability Item reliability Construct validity
6. Minimal effect of demographic (age, education) and other cognitive variables (attention, memory, intelligence) on performance	Lack of correlation between these demographic/cognitive factors and test performance Representative normative sample
7. Differentiation between individuals with aphasia, individuals without cognitive or language impairments, and individuals without aphasia but with other cognitive deficits due to brain damage	Criterion validity Sensitivity Specificity Construct validity
8. Measurement of similarity to results on known aphasia tests	Concurrent validity

be particularly interesting in and of themselves, they will become invaluable as references for researchers and clinicians who use those tests to quantify language deficits. Future readers can then evaluate the content, reliability and validity of assessment instruments in a language that they do not speak. Dissemination of test properties in peer-reviewed publications is essential and ultimately renders findings obtained in different languages more comparable. Improvement in our clinical procedures across languages and the augmentation of the scientific merit of our studies through psychometrically valid assessment tools are well worth the efforts invested in developing standardised tests.

Manuscript received 14 January 2013
Manuscript accepted 10 May 2013
First published online 19 June 2013

REFERENCES

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

Allen, M. J., & Yen, W. M. (2002). *Introduction to measurement theory*. Prospect Heights, IL: Waveland Press.

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

Anastasi, A., & Urbina, S. (1997). *Psychological testing*. Upper Saddle River, NJ: Prentice Hall.

Arditi, A., & Cho, J. (2007). Letter case and text legibility in normal and low vision. *Vision Research*, *47*, 2499–2505.

Baker, R. (2000). The assessment of functional communication in culturally and linguistically diverse populations. In L. E. Worrall & C. M. Frattali (Eds.), *Neurogenic communication disorders: A functional approach* (pp. 81–100). New York: Thieme.

Bastiaanse, R., Bosje, M., & Visch-Brink, E. (1995). *PALPA: Nederlandse Versie*. Hove: Lawrence Erlbaum.

Bastiaanse, R., Maas, E., & Rispens, J. (2000). *Werkworoden- en Zinnetest (WEZT)*. Lisse: Swets Test.

Bates, E., Wulfeck, B., & MacWhinney, B. (1991). Crosslinguistic research in aphasia: An overview. *Brain and Language*, *41*, 123–148.

Baylor, C., Hula, W., Donovan, N., Doyle, P., Kendall, D., & Yorkston, K. (2011). An introduction to item response theory for speech language pathologists. *American Journal of Speech-Language Pathology*, *20*, 243–259.

Beland, R., & Lecours, A. R. (1990). The MT-86 β aphasia battery: A subset of normative data in relation to age and level of school education. *Aphasiology*, *4*, 439–462.

Beland, R., Lecours, A. R., Giroux, F., & Bois, M. (1993). The MT-86 β aphasia battery: A subset of normative data in relation to age and level of school education (Part II). *Aphasiology*, *7*, 359–382.

Benton, A. L., & Hamsher, K. D. (1994). *Examen de Afasia Multilingue (MAE-S)*. San Antonio, TX: Psychological Corporation.

Benton, A. L., Hamsher, Kd. eS., & Sivan, A. B. (1994). *Multilingual aphasia examination* (3rd ed.). San Antonio, TX: Psychological Corporation.

Bhatnagar, S. C. (1984). Aphasia in the Indian context: An indigenously developed aphasia test battery in Hindi. In *Continuing medical education proceedings of the neurological society of India* (pp. 183–219). Banaras: Neurological Society of India.

Bhatnagar, S. C. (n.d.). *AIIMS diagnostic test of aphasia*. Unpublished test and manuscript.

Bhatnagar, S. C., Jain, S. K., Bihari, M., Bansal, N. K., Pauranik, A., Jain, D. C., . . . Packma, M.V. (2002). Aphasia type and aging in Hindi-speaking stroke patients. *Brain and Language*, *83*, 353–361.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

Brennan, A. D., Worrall, L. E., & McKenna, K. T. (2005). The relationship between specific features of aphasia-friendly written material and comprehension of written material for people with aphasia: An exploratory study. *Aphasiology*, *19*, 693–711.

Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology*, *1*, 185–216.

Chenggapa, S. (2009). *Bi/Multilingualism and issues in management of communication disorders with emphasis on Indian perspectives*. Retrieved from http://www.languageinindia.com

Connor, L. T., MacKay, A. J., & White, D. A. (2000). Working memory: A foundation for executive abilities and higher-order cognitive skills. *Seminars in Speech and Language*, *21*, 109–119.

D'Agostino, L. (1985). *Taratura su soggetti normali di prove di denominazione per l'afasia*. Tesi di Laurea della Facolta' di Medicina e Chirurgia, Istituto di Clinica Neurologica, Universita' degli studi di Modena.

Dardarananda, R., Potisuk, S., Grandour, J., & Holasuit, S. (1999). *Thai adaptation of the Western Aphasia Battery (WAB)*. Thailand: Chiangmai Medical Bulletin.

De Bleser, R., Cholewa, J., Stadie, N., & Tabatabaie, S. (1997). LeMo, an expert system for single case assessment of word processing impairments in aphasic patients. *Neurosychological Rehabilitation*, *7*, 339–366.

De Bleser, R., Cholewa, J., Stadie, N., & Tabatabaie, S. (2004). *LeMo—Lexikon modellorientiert. Einzelfalldiagnostik bei Aphasie, Dyslexie und Dysgraphie*. München: Urban & Fischer.

Delacour, A., Wyrzykowski, N., Lefeuvre, M., & Rousseaux, M. (2000). Élaboration d'une nouvelle évaluation de la communication, le TLC. *Glossa*, *72*, 20–29.

Deloche, G., & Hannequin, D. (1997). *Test de Dénomination Orale D'images (DO 80)*. Paris: Centre de Psychologie Appliquée.

Ducarne, B. (1989). *Test pour l'examen de l'aphasie* (Forme révisée). Paris: Éditions du Centre de psychologie appliquée.

Edwards, S., & Bastiaanse, R. (2007). Assessment of aphasia in a multi-lingual world. In M. J. Ball & J. S. Damico (Eds.), *Clinical aphasiology: Future directions* (pp. 245–258). New York, NY: Psychology Press.

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

Fishman, J. A., & Galguera, T. (2003). *Introduction to test construction in social and behavioral sciences*. Oxford: Rowman & Littlefield.

Franzen, M. D. (2003). *Reliability and validity in neuropsychological assessment*. New York, NY: Kluwer Academic/Plenum.

Gandour, J., Dardarananda, R., Buckingham Jr., H., & Viriyavejakul, A. (1986). A Thai adaptation of the BDAE. *Crossroads: An Interdisciplinary Journal of Southeast Asian Studies*, *2*, 1–39.

Gao, S. (1996). Clinical diagnosis of aphasia in Chinese. *Stroke and Nervous Diseases*, *3*, 57–59.

Garcia-Albea, J. E., Sanchez-Bernardos, M. L., & del Viso-Pabon, S. (1986). Test de Boston parallel diagnostico de la afasia: Adaptacion Espagnola. In H. Goodglass and E. Kaplan (Eds.), *La evolacion de la afasio y de transtornos relacionados* (2nd ed., translated by Carlos Wenicke; pp. 129–196). Madrid: Editorial Medica Panamericana.

Goodglass, H., & Kaplan, E. (2001). *Boston naming test*. Philadelphia, PA: Lippincott Williams and Wilkins.

Goodglass, H., Kaplan, E., & Barresi, B. (2001a). *Boston diagnostic aphasia examination* (3rd ed.). Philadelphia, PA: Lippincott Williams & Wilkins.

Goodglass, H., Kaplan, E., & Barresi, B. (2001b). *The assessment of aphasia and related disorders* (3rd ed.). Philadelphia, PA: Lippincott Williams & Wilkins.

Graetz, P., De Bleser, R., & Willmes, K. (1992). *Akense afasie test*. Lisse: Swetz and Zeitlinger.

Hallowell, B. (2008). Strategic design of protocols to evaluate vision in research on aphasia and related disorders. *Aphasiology*, *22*, 600–617.

Hallowell, B., & Chapey, R. (2008). Introduction to language intervention strategies in adult aphasia. In R. Chapey (Ed.), *Language intervention strategies in aphasia and related neurogenic communication disorders* (5th ed.) (pp. 3–19). Philadelphia, PA: Lippincott Williams & Wilkins.

Hallowell, B., & Ivanova, M. V. (2009). Development and standardization of a multiple-choice test of auditory comprehension for aphasia in Russian. *Journal of Medical Speech-Language Pathology*, *2*, 83–98.

Hallowell, B., Wertz, R. T., & Kruse, H. (2002). Using eye movement responses to index auditory comprehension: An adaptation of the revised token test. *Aphasiology*, *16*, 587–594.

Hasegawa, T., Kishi, H., Shigeno, K., Tanemura, J., Kusunoki, T., Kifune, Y., & Yoshihisa, K. (1984). A study on aphasia rating scale: A method for overall assessment of SLTA results [Japanese]. *Higher Brain Function Research (Shitsugosho-Kenkyu)*, *4*, 638–646.

Helm-Estabrooks, N. (2001). *Cognitive linguistic quick test*. San Antonio, TX: Psychological Corporation.

Helm-Estabrooks, N. (2002). Cognition and aphasia: A discussion and a study. *Journal of Communication Disorders*, *35*, 171–186.

Heuer, S., & Hallowell, B. (2007). An evaluation of test images for multiple-choice comprehension assessment in aphasia. *Aphasiology*, *21*, 883–900.
Heuer, S., & Hallowell, B. (2009). Visual attention in a multiple-choice task: Influences of image characteristics with and without presentation of a verbal stimulus. *Aphasiology*, *23*, 351–363.
Higashikawa, M., Hadano, K., & Hata, T. (2006). The core factor for improvement in recovery from aphasia extracted by factor analysis. *Psychologia*, *49*, 143–151.
Howard, D., Swinburn, K., & Porter, G. (2010). Putting the CAT out: What the comprehensive aphasia test has to offer. *Aphasiology*, *24*, 56–74.
Huber, W., Poeck, K., Weniger, D., & Willmes, K. (1983). *Der aachener aphasia test*. Gottingen: Verlag fur Psychologie Hogrefe.
Hula, W., Donovan, N., Kendall, D., & Gonzalez, L. R. (2009, May). *Item response theory analysis of the Western Aphasia Battery*. Paper presented at the Clinical Aphasiology Conference, Teton Village, WY.
Hula, W., Doyle, P. J., McNeil, M. R., & Mikolic, J. M. (2006). Rasch modeling of revised token test performance: Validity and sensitivity to change. *Journal of Speech, Language, and Hearing Research*, *49*, 27–46.
Hula, W., Ross, K., Hula, S. A., Wambaugh, J., Schumacher, J., & Doyle, P. (2010, May). *Validity of and agreement between self- and surrogate-reported communicative functioning in persons with aphasia*. Paper presented at the Clinical Aphasiology Conference, Isle of Palms, SC.
Hula, W. D., & McNeil, M. R. (2008). Models of attention and dual-task performance as explanatory constructs in aphasia. *Seminars in Speech and Language*, *29*, 169–187.
Ivanova, M. V., & Hallowell, B. (2009). Short form of the bilingual aphasia test in Russian: Psychometric data of persons with aphasia. *Aphasiology*, *23*, 544–556.
Ivanova, M. V., & Hallowell, B. (2012). Validity of an eye-tracking method to index working memory in people with and without aphasia. *Aphasiology*, *26*, 556–578.
Kacker, S. K., Pandit, R., & Dua, D. (1991). Reliability and validity studies of examination for aphasia test in Hindi. *Indian Journal of Disability and Rehabilitation*, *5*, 13–19.
Kaplan, E. F., Goodglass, H., & Weintraub, S. (1986). *Test de Vocabulario de Boston*. Madrid: Panamericana.
Keenan, J. S., & Brassell, E. G. (1975). *Aphasia language performance scales* (Spanish version). Murfreesboro, TN: Pinnacle Press.
Kertesz, A. (1982). *Western aphasia battery*. New York, NY: Grune & Stratton.
Kertesz, A. (2007a). *Western aphasia battery-revised*. San Antonio, TX: Harcourt Assessment.
Kertesz, A. (2007b). *Western aphasia battery-revised, examiner's manual*. San Antonio, TX: Harcourt Assessment.
Kertesz, A., Pascual-Leone, P., & Pascual-Leone, G. (1990). *Western Aphasia Battery en versión y adaptación castellana* [Western Aphasia Battery-Spanish version]. Valencia: Nau Libres.
Kim, H. (2009). *Screening test for aphasia and neurologic communication disorders*. Seoul: KOPS.
Kim, H., & Na, D. L. (1997). *Korean version—Boston naming test*. Seoul: Hak Ji Sa.
Kim, H., & Na, D. L. (1999). Normative data on a Korean version of the Boston Naming Test. *Journal of Clinical and Experimental Neuropsychology*, *21*, 127–133.
Kim, H., & Na, D. L. (2001). *Korean version—the Western aphasia battery*. Seoul: Paradise Institute for Children with Disabilities.
Kim, H., & Na, D. L. (2004). Normative data on the Korean version of the Western aphasia battery. *Journal of Clinical and Experimental Neuropsychology*, *26*, 1011–1020.
Kohnert, K., Hernandez, A. E., & Bates, E. (1998). Bilingual performance on the boston naming test: Preliminary norms in Spanish and English. *Brain and Language*, *65*, 422–440.
Laine, M., Goodglass, H., Niemi, J., Koivuselka-Sallinen, P., Tuomainen, J., & Martilla, R. (1993). Adaptation of the Boston diagnostic aphasia examination and the Boston naming test into Finnish. *Scandinavian Journal of Logopedics and Phoniatrics*, *18*, 83–92.
Laine, M., Koivuselkä-Sallinen, P., Hännine, R., & Niemi, J. (1993). *Bostonin Nimentätestin Suomenkielinen versio*. Helsenki: Psykologien Kustannus.
Laine, M., Niemi, J., Koivuselkä-Sallinen, P., & Tuomainen, J. (1993). *Bostonin Diagnostises Afasiatestistön Suomenkielinen versio*. Helsenki: Psykologien Kustannus.
Lauterbach, M. (2006). *Influence of educational level in aphasia testing: experiences from standardizing the Portuguese version of the AAT (Aachener Aphasie Test)*. Retrieved from http://www.cplol.org/CD-Rom_2006/content/List%20of%20papers/org_text/07_Lauterbach_full%20text_EN.htm
Lemay, M. A. (1988). Protocole d'évaluation des dyslexies acquises. *Rééducation Orthophoniques*, *26*, 363–376.

Lemay, M. A. (1990). *Examen des Dyslexies Acquises*. Montréal: Point-Carré.

Lewis, M. P. (Ed.). (2009). *Ethnologue: Languages of the world* (16th ed.). Dallas, TX: SIL International.

Lezak, M. D., Howieson, D. B., & Loring, D. W. (with Hannay H.J., & Fischer, J. S.) (2004). *Neuropsychological assessment* (4th ed.). New York, NY: Oxford University Press.

Lorenzen, B., & Murray, L. L. (2008). Bilingual aphasia: A theoretical and clinical review. *American Journal of Speech-Language Pathology*, *17*, 299–317.

Luzzatti, C., Willmes, W., & De Bleser, R. (1996). *Aachner Aphasia Test (AAT): Versione Italiana* (2nd ed.). Florence: Organizzazioni Speciali.

Mahendra, N. (2004). *Modifying the communicative abilities of daily living (CADL-2) for use with illiterate persons with aphasia: Preliminary results*. Retrieved from http://www.speechpathology.com/Articles/article_detail.asp?article_id=242

Manuel-Dupont, S., Ardila, A., Rosseli, M., & Puente, A. E. (1992). Bilingualism. In A. E. Puente & R. J. McCaffrey (Eds.), *Handbook of neuropsychological assessment* (pp. 193–210). New York, NY: Plenum Press.

Marien, P., Mampaey, E., Vervaet, A., Saerens, J., & De Deyn, P. P. (1998). Normative data for the Boston naming test in native Dutch-speaking Belgian elderly. *Brain and Language*, *65*, 447–467.

Marshall, R. C., & Wright, H. H. (2007). Developing a clinician-friendly aphasia test. *American Journal of Speech-Language Pathology*, *16*, 295–315.

Martin, N., Kohen, F. P., & Kalinyak-Fliszar, M. (2010, May). *A processing approach to the assessment of language and verbal short-term memory abilities in aphasia*. Paper presented at the Clinical Aphasiology Conference, Isle of Palms, SC.

Martin, P., Manning, L., Munoz, P., & Montero, I. (1990). Communicative abilities in daily living: Spanish standardization. *Evaluacion Psicologica*, *6*, 369–384.

Mazaux, J. M., & Orgogozo, J. M. (1982). *Echelle d'évaluation de l'aphasie. Adaptation Française du Boston Diagnostic Aphasia Examination*. Paris: Editions Scientifiques et Psychologiques.

McCauley, R. J. (1996). Familiar strangers: Criterion-referenced measures in communication disorders. *Language, Hearing, & Speech Services in Schools*, *27*, 122–131.

McNeil, M. R., Odell, K., & Tseng, C. H. (1991). Toward the integration of resource allocation into a general theory of aphasia. *Clinical Aphasiology*, *20*, 21–39.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741–749.

Metz-Lutz, M. N., Kremin, H., Deloche, G., Hannequin, D., Ferrand, I., Perrier, D., . . . Blavier, A. (1991). Standardisation d'un test de dénomination orale: Contrôle des effets de l'âge, du sexe et du niveau de scolarité chez les sujets adultes normaux. *Revue De Neuropsychologie*, *1*, 73–95.

Miceli, G., Laudanna, A., Burani, C., & Capasso, R. (1994). *Batteria per l'analisi dei deficit afasici*. Roma: Centro di Promozione e Sviluppo dell'Assistenza Geriatrica.

Miller, N., De Bleser, R., & Willmes, K. (1997). The English language version of the aachen aphasia test. In W. Zeiegler & K. Deger (Eds.), *Clinical phonetics and linguistics* (pp. 257–265). London: Whurr.

Miller, N., Willmes, K., & De Bleser, R. (2000). The psychometric properties of the English language version of the aachen aphasia test (EAAT). *Aphasiology*, *14*, 683–722.

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

Mumby, K. (1988). An adaptation of the aphasia screening test for use with Punjabi speakers. *British Journal of Disorders of Communication*, *23*, 209–226.

Mumby, K. (1990). Preliminary results from using the Punjabi adaptation of the aphasia screening test. *British Journal of Disorders of Communication*, *25*, 209–226.

Murray, L. L. (1999). Attention and aphasia: Theory, research and clinical implications. *Aphasiology*, *13*, 91–111.

Murray, L. L. (2004). Cognitive treatments for aphasia: Should we and can we help attention and working memory problems? *Journal of Medical Speech-Language Pathology*, *12*, xxv–xi.

Murray, L. L., & Clark, H. M. (2006). *Neurogenic disorders of language: Theory driven clinical practice*. New York, NY: Thomson Delmar Learning.

Naeser, M. A., & Chan, S. W. (1980). Case study of a Chinese aphasic with the Boston diagnostic aphasia exam. *Neuropsychologia*, *18*, 389–410.

Nespoulos, J. L., Lecours, A. R., Lafond, D., Lemay, A., Puel, M., Joannette, Y., . . . Rascol, A. (1992). *Protocole Montréal-Toulouse d'Examen Linguistique d l'Aphasie* (Version M1 beta). Isbergues: Ortho Édition.

Nicholas, L. E., MacLennan, D. L., & Brookshire, R. H. (1986). Validity of multiple-sentence reading comprehension tests for aphasic adults. *Journal of Speech and Hearing Disorders*, *51*, 82–87.

Novelli, G., Papagno, C., Capitani, E., Laiacona, M., Vallar, G., & Cappa, S. F. (1986). Tre test clinici di ricerca e produzione lessicale: Taratura su soggetti normali. *Archivio Di Psicologia, Neurologia E Psichiatria*, *4*, 477–506.

Odekar, A., Hallowell, B., Lee, C., & Moates, D. (2009). Validity of eye movement methods and indices for capturing semantic (associative) priming effects. *Journal of Speech, Language and Hearing Research*, *52*, 1–18.

Paradis, M. (with Libben, G.). (1987). *The assessment of bilingual aphasia*. Hillsdale, NJ: Lawrence Erlbaum.

Paradis, M. (n.d.) *Publications: The bilingual aphasia tests*. Retrieved from http://www.semioticon.com/virtuals/cv/paradis.htm

Paradis, M., & Abidi, R. (1987). *Fahs hallat fuqdan taqat annotk aw el kalam ande mudjeedi loghatein aw akthar* (Maghrebian Arabic version). Hillsdale, NJ: Lawrence Erlbaum.

Paradis, M., & Ardila, A. (1989). *Prueba de afasia para bilingües* (American Spanish version). Hillsdale, NJ: Lawrence Erlbaum.

Paradis, M., Canzanella, M. A., & Baruzzi, A. (1987). *Esame dell'afasia di una persona bilingue* (Italian version). Hillsdale, NJ: Lawrence Erlbaum.

Paradis, M., & Devanathan, T. (1989). *Bilingual aphasia test* (Tamil version). Hillsdale, NJ: Lawrence Erlbaum.

Paradis, M., & El Halees, Y. M. (1989). *Bilingual aphasia test* (Jordanian Arabic version). Hillsdale, NJ: Lawrence Erlbaum.

Paradis, M., & Elias, J. (1987). *Test de la afasia en los bilingües* (European Spanish version). Hillsdale, NJ: Lawrence Erlbaum.

Paradis, M., & Goldblum, M. -C. (1987). *Test de l'aphasie chez les bilingues* (French version). Hillsdale, NJ: Lawrence Erlbaum.

Paradis, M., & Hagiwara, H. (1987). *Ni gengo heiyosha ni okeru shitsugosho kensa* (Japanese version). Hillsdale, NJ: Lawrence Erlbaum.

Paradis, M., & Hub Faria, I. (1989). *Teste de afasia para bilingues* (European Portuguese version). Hillsdale, NJ: Lawrence Erlbaum.

Paradis, M., & Janjua, N. (1987). *Bilingual aphasia test* (Urdu version). Hillsdale, NJ: Lawrence Erlbaum.

Paradis, M., & Lindner, O. (1987). *Aphasie test in Deutsch für Zweisprachige* (German version). Hillsdale, NJ: Lawrence Erlbaum.

Paradis, M., & Shen, R. (1987). *Bilingual aphasia test* (Modern Standard Chinese version). Hillsdale, NJ: Lawrence Erlbaum.

Paradis, M., Simões, A., & Dillinger, M. (1987). *Teste de afasia para bilíngües* (Brazilian Portuguese version). Hillsdale, NJ: Lawrence Erlbaum.

Paradis, M., & Suh, J. -M. (1991). *Bilingual aphasia test* (Korean version). Hillsdale, NJ: Lawrence Erlbaum.

Paradis, M., & Truong, Y. (1987). *Bilingual aphasia test* (Vietnamese version). Hillsdale, NJ: Lawrence Erlbaum.

Paradis, M., & Vaid, J. (1987). *Dvibhashai ka pratikshan* (Hindi version). Hillsdale, NJ: Lawrence Erlbaum.

Paradis, M., & Zeiber, T. (1987). *Bilingual aphasia test* (Russian version). Hillsdale, NJ: Lawrence Erlbaum.

Park, H. S., Sasanuma, S., Sunwoo, I. N., Rah, U. W., & Shin, J. S. (1992). The preliminary clinical application of the Tentative Korean aphasia test battery form (1). *Korean Neuropsychology*, *10*, 350–357.

Patterson, J. P., & Chapey, R. (2008). Assessment of language disorders in adults. In R. Chapey (Ed.), *Language intervention strategies in aphasia and related neurogenic communication disorders* (5th ed.) (pp. 65–160). Philadelphia, PA: Lippincott Williams & Wilkins.

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. New York: Oxford University Press.

Pineda, D. A., Rosselli, M., Ardila, A., Mejia, S. E., Romero, M. G., & Perez, C. (2002). The Boston diagnostic aphasia examination–Spanish version: The influence of demographic variables. *Journal of the International Neuropsychological Society*, *6*, 802–814.

Pizzamiglio, L., Laicardi, C., Appicciafuoco, A., Gentili, P., Judica, A., Luglio, L., . . . Razzano, C. (1984). Capacita communicative di pazienti afasici in situationi di vita quotidiana: Addatamento italiano. *Archivio Di Psicologia, Neurologia E Psychiatria*, *45*, 187–210.

Ponton, M. O., Satz, P., Herrera, L., Ortiz, F., Urrutia, C. P., Young, R., . . . Namerow, N. (1996). Normative data stratified by age and education for the Neuropsychological Screening Battery for Hispanics (NESBHIS): Initial report. *Journal of the International Neuropsychological Society*, *2*, 96–104.

Ponton, M. O., Satz, P., Herrera, L., Young, R., Oritz, F., D'Elia, L., . . . Namerow, N. (1992). A modified Spanish version of the Boston naming test. *The Clinical Neuropsychologist*, *6*, 334.

Pracharitpukdee, N., Phanthumchinda, K., Huber, W., & Willmes, K. (2000). The Thai version of the German Aachen Aphasia Test (THAI-AAT). *Journal of the Medical Association of Thailand*, *83*, 601–610.

Reinvang, I., & Graves, R. (1975). A basic aphasia examination: Description with discussion of first results. *Scandinavian Journal of Rehabilitation Medicine*, *7*, 129–135.

Rey, G. J., Feldman, E., Hernandez, D., Levin, B. E., Rivas-Vazquez, R., Nedd, K. J., & Benton, A. L. (2001). Application of the multilingual aphasia examination-Spanish in the evaluation of Hispanic patients post closed-head trauma. *The Clinical Neuropsychologist*, *15*, 13–18.

Rey, G. J., Feldman, E., Rivas-Vazquez, R., Levin, B. E., & Benton, A. (1999). Neuropsychological test development and normative data on Hispanics. *Archives of Clinical Neuropsychology*, *14*, 593–601.

Roberts, P. M. (2001). Aphasia assessment and treatment in bilingual and multicultural populations. In R. Chapey (Ed.), *Language intervention strategies in aphasia and related neurogenic communication disorders* (4th ed.) (pp. 208–232). Philadelphia, PA: Lippincott Williams & Wilkins.

Roberts, P. M. (2008). Issues in assessment and treatment for bilingual and culturally diverse patients. In R. Chapey (Ed.), *Language intervention strategies in aphasia and related neurogenic communication disorders* (5th ed.) (pp. 245–275). Philadelphia, PA: Lippincott Williams & Wilkins.

Roberts, P. M., & Doucet, N. (2011). Performance on French-speaking Quebec adults on the Boston naming test. *Canadian Journal of Speech-Language Pathology and Audiology*, *35*, 254–267.

Ross, K. B., & Wertz, R. T. (2003). Discriminative validity of selected measures for differentiating normal from aphasic performance. *American Journal of Speech-Language Pathology*, *12*, 312–319.

Ross, K. B., & Wertz, R. T. (2004). Accuracy of formal tests for diagnosing mild aphasia: An application of evidence-based medicine. *Aphasiology*, *18*, 337–355.

Rosselli, M., Ardila, A., Florez, A., & Castro, C. (1990). Normative data on the Boston diagnostic aphasia examination in a Spanish-speaking population. *Journal of Clinical and Experimental Neuropsychology*, *12*, 313–322.

Rousseaux, M., Delacour, A., Wyrzykowski, N., & Lefeuvre, M. (2003). *TLC Test Lillois de Communication*. Isbergues: Ortho Édition.

SLTA Committee. (1977). *Standard language test of aphasia: Manual of directions* (2nd ed.). Tokyo: Homeido.

Sasanuma, S., Itoh, M., Watamori, T., Fukusako, Y., & Monoi, H. (1992). *Treatment of aphasia*. Tokyo: Igaku-Shoin.

Schiavetti, N., & Metz, D. E. (2002). *Evaluating research in communication disorders* (4th ed.). Boston, MA: Allyn & Bacon.

Soroker, N. (1997). *Hebrew Western aphasia battery*. Ra'anana: Loewenstein Hospital Rehabilitation Center.

Spreen, O., & Risser, A. H. (2003). *Assessment of aphasia*. Oxford: Oxford University Press.

Sreedevi, N. (1991). *Comprehension deficits in bilingual aphasics*. Unpublished doctoral dissertation, Mysore University, India.

Stadie, N., De Bleser, R., Cholewa, J., & Tabatabaie, S. (1994). Das neurolinguistische Expertensystem LeMo: Theoretischer Rahmen und Konstruktionsmerkmale des Testteils LEXIKON. *Neurolinguistik*, *1*, 1–25.

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

Sugishita, M. (1986). *Japanese edition of Western aphasia battery*. Tokyo: Igaku-Shoin.

Swinburn, K., Porter, G., & Howard, D. (2004). *Comprehensive aphasia test*. Routledge: Psychology Press.

Tallberg, I. M. (2004). The Boston naming test in Swedish: Normative data. *Brain and Language*, *94*, 19–31.

Taussig, I. M., Henderson, V. W., & Mack, W. (1992). Spanish translation and validation of a neuropsychological battery: Performance of Spanish- and English-speaking Alzheimer's disease patients and normal comparison subjects. *Clinical Gerontologist*, *11*, 95–108.

Thuillard-Colombo, F., & Assal, G. (1992). Adaptation française du test de dénomination de Boston: Versions abrégées. *Revue Européene De Psychologie Appliquée*, *42*, 67–71.

Tsang, H. L. (2000). *Confrontation naming abilities of the young, the elderly, and people with aphasia* (Unpublished thesis). University of Hong Kong, Hong Kong, China.

Tseng, O. J.-L. (1993). *A Chinese version of the Boston Diagnostic Aphasia Examination*. Unpublished manuscript.

Tsvetkova, L. S., Axytina, T. V., & Pulaeva, N. M. (1981). *Kolichestvennaya ocenka rechi y bol'nux s aphasieu*. Moscow: Izdatelstvo Moskovskogo Gosydarstvennogo Yniversiteta.

Unverzagt, F. W., Morgan, O. S., & Thesiger, C. H. (1999). Clinical utility of CERAD neuropsychological battery in elderly Jamaicans. *Journal of International Neuropsychological Society*, *5*, 255–259.

Valle, F., & Cuetos, F. (1995). *EPLA: Evaluacion del Procesamiento Linguistico en la Afasia*. Hillsdale, NJ: Lawrence Erlbaum.

Watamari, T., Takauechi, M. I., Fukusako, Y., Itoh, M., Suzuki, K., Endo, K., . . . Sasanuma, S. (1987). Development and standardization of Communication Abilities in Daily Living (CADL) test for Japanese aphasic patients. *Japanese Journal of Rehabilitation Medicine*, *24*, 103–112.

Watamari, T., Takauechi, M. I., Itoh, M., Fukusako, Y., Suzuki, K., Endo, K., . . . Sasanuma, S. (1990). *Test for functional abilities—CADL test*. Tokyo: Ishiyaku.

Willmes, K. (1997). Application of polytomous Rasch models to the subtest written language of the Aachen Aphasia Test (AAT). In J. Rost & R. Langeheine (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 127–137). New York: Waxmann, Munster.

Willmes, K., Graetz, P., De Bleser, R., Schulte, B., & Keyser, A. (1991). De Akense afasie test. *Logopedie En Foniatrie*, *63*, 375–386.

Wolfe, W. E., & Smith Jr, E. V. (2007a). Instrument development tools and activities for measure validation using Rasch models: Part I—Instrument development tools. *Journal of Applied Measurement*, *8*, 97–123.

Wolfe, W. E., & Smith Jr, E. V. (2007b). Instrument development tools and activities for measure validation using Rasch models: Part II—Validation activities. *Journal of Applied Measurement*, *8*, 204–234.

Wright, H. H., & Fergadiotis, G. (2012). Conceptualizing and measuring working memory and its relationship to aphasia. *Aphasiology*, *26*, 258–278.

Yiu, E. M. L. (1992). Linguistic assessment of Chinese-speaking aphasics: Development of a Cantonese aphasia battery. *Journal of Neurolinguistics*, *7*, 379–424.

Zhang, Q., Ji, S., & Li, S. (2005). Reliability and validity of the Chinese rehabilitation research center standard aphasia examination. *Chinese Journal of Rehabilitation Theory and Practice*, *11*, 703–705.