Unmasking the Psychometric Challenges of the Boston Naming Test in the North American Multicultural Context

Dévoiler les faiblesses psychométriques du *Boston Naming Test* (*Test de dénomination de Boston*) lors de son application dans le contexte multiculturel nord-américain

Manon Spigarelli
David Gotti
Audrey Lalancette

KEYWORDS
ANOMIA
BOSTON NAMING TEST
NEUROPSYCHOLOGICAL ASSESSMENT
MULTICULTURAL POPULATIONS
METROLOGICAL PROPERTIES

Manon Spigarelli, David Gotti, Audrey Lalancette

École des sciences de la réadaptation, Université Laval, Québec, QC, CANADA

Centre interdisciplinaire de recherche en réadaptation et intégration sociale, Centre intégré universitaire de santé et de services sociaux de la Capitale-Nationale du Québec, Québec, QC, CANADA

Editor:
David McFarland

Abstract

Anomia, characterized by the inability to retrieve appropriate words for naming objects, people, and actions, is a prevalent phenomenon in neuropsychology and speech-language pathology. The assessment of naming deficits is crucial in various pathological conditions, including poststroke aphasia, primary progressive aphasia, and Alzheimer's disease. The Boston Naming Test is a widely used assessment tool for common-name anomia in North America. However, concerns about its psychometric qualities, cross-cultural validity, and differentiation among ethnic groups have raised questions about its diagnostic utility. This systematic review evaluates the psychometric properties of the 60-item version of the Boston Naming Test, specific to its use in the North American multicultural context. Although the Boston Naming Test has shown promise in distinguishing groups with different clinical conditions, issues regarding its cross-cultural validity and its ability to differentiate between ethnic groups persist. This review calls for further studies to explore these concerns and delve deeper into the psychometric properties of the Boston Naming Test. Despite the Boston Naming Test being known as a valuable diagnostic tool, its interpretation should be handled with caution and awareness of its limitations in diverse clinical, cultural, and population contexts. Considering the identified challenges, the development and standardization of a new object-naming test that addresses these issues appears to be a promising path forward.

Abrégé

L'anomie, qui se caractérise par une incapacité à récupérer les bons mots pour nommer des objets, des personnes et des actions, est un phénomène courant en neuropsychologie et en orthophonie. L'évaluation des déficits de dénomination est donc une composante essentielle des évaluations neuropsychologique et orthophonique réalisées avec des individus atteints de diverses conditions pathologiques, incluant l'aphasie secondaire à un accident vasculaire cérébral, l'aphasie primaire progressive et la maladie d'Alzheimer. Le *Boston Naming Test* (*Test de dénomination de Boston*) est un outil grandement utilisé en Amérique du Nord pour évaluer l'anomie touchant les mots courants. Cependant, des préoccupations concernant ses qualités psychométriques, sa validité interculturelle et sa capacité à différencier les performances des individus de différentes origines ethniques ont soulevé des questions quant à son utilité diagnostique. La présente revue systématique évalue les propriétés psychométriques de la version de 60 items du *Boston Naming Test*, en lien avec son utilisation spécifique dans le contexte multiculturel nord-américain. Bien que le *Boston Naming Test* se soit montré prometteur pour distinguer les performances d'individus présentant différentes conditions cliniques, des questions concernant sa validité interculturelle et sa capacité à différencier les performances d'individus de diverses origines ethniques persistent. La présente revue systématique argue en faveur d'études supplémentaires examinant ces questions et investiguant plus en profondeur les propriétés psychométriques du *Boston Naming Test*. De plus, même si le *Boston Naming Test* est considéré comme un outil de diagnostic utile, l'interprétation de ses résultats doit se faire avec prudence et en tenant compte des limites liées à son utilisation dans les divers contextes cliniques, culturels et démographiques. Au regard des préoccupations soulevées, l'élaboration et la normalisation d'un nouvel outil d'évaluation de la dénomination d'objets paraissent être une avenue prometteuse.

Anomia, extensively studied in speech-language pathology, is defined as an inability to find the right and appropriate word to name objects, people, and actions (Alexander & Hillis, 2008; Mesulam, 2016). Although anomia tends to increase naturally with age (Albert, 1988), it remains a major symptom in many pathological conditions. Its assessment is crucial to quantify impairment extent and plan interventions for individuals with poststroke, primary progressive aphasia, or Alzheimer's disease, among other conditions (Alexander & Hillis, 2008; Mesulam, 2016). A sensitive and valid test of naming ability is therefore of critical importance in overall neuropsychological and language assessments (Harry & Crowe, 2014).

In North America, the Boston Naming Test (BNT; Kaplan et al., 2001) is one of the most widely used object-naming tests in speech-language pathology for assessing common-name anomia (Rabin et al., 2005, 2016). Its construct is based on the cognitive model of oral word production proposed by Caramazza and Hillis (1991). Although alternative object-naming tests exist, such as the Western Aphasia Battery (Kertesz, 2007), they are typically less comprehensive and integrated into language assessment batteries. Originally comprising 60 items and published in the United States (Kaplan et al., 1983), the most recent version of the BNT (2nd edition; Kaplan et al., 2001) has gained widespread popularity in many other English-speaking countries such as Canada, England, and Australia. Recognizing the need for flexibility in administration and integration into broader test batteries, such as the Boston Diagnostic Aphasia Examination (Goodglass et al., 2001), shortened versions of the BNT, featuring 30 or 15 items, have also been developed (Fastenau et al., 1998; Graves et al., 2004; Lansing et al., 1999; Mack et al., 1992; Saxton et al., 2000; Teng et al., 1989).

Although the BNT is extensively used in North America, concerns have been raised regarding its psychometric qualities (Harry & Crowe, 2014). Notably, some argue that the lack of precision in the guidelines for administration and scoring could contribute to significant interrater variability. For example, unclear interpretations of the discontinuation criterion when administering the BNT may disproportionately impact the assessment of naming abilities in individuals with impairments (Ferman et al., 1998). In addition, it has been pointed out that the BNT presents with a nonnormal score distribution. Consequently, relying on *Z*-scores may result in inaccurate test interpretations, such as overdiagnosis, particularly in cases where performance is only slightly reduced (Bortnik et al., 2013). Also, the presence of a ceiling effect diminishes the effectiveness of the BNT in detecting impairment in some cases (Brooks et al., 2009). Finally, a recurrent criticism is the variability of results among different groups of individuals sharing a similar clinical profile. For example, in individuals without impairment, mean scores exhibit significant variability based on personal factors, such as age, education, and gender (Zec et al., 2007). The impact of this variability on test interpretation is not adequately addressed in the BNT's limited standardization (Harry & Crowe, 2014).

In addition to these factors, the familiarity of a concept and the frequency with which individuals are engaged with it (Pompéia et al., 2001) influence the ability to name an image (Chedid et al., 2019; Ghasisin et al., 2015). Familiarity is known to be a significant predictor of naming reaction times, with concepts that are more familiar being named more quickly (Hirsh & Funnell, 1995). Indeed, familiar objects are semantically richer than less familiar ones, aiding in their identification. A test like the BNT should therefore consider the influence of an individual's cultural environment on the familiarity of certain items when assessing naming ability. For instance, recognizing an igloo image might be quicker for someone in Canada, where igloos are culturally and geographically relevant, than for someone in Australia. Moreover, numerous scholars emphasize the responsibility of clinicians to enhance the validity of assessments by tailoring tests, employing suitable norms, and taking into account the unique contextual factors of each patient during the interpretation of scores (Daugherty et al., 2017; Fernández & Fulbright, 2015; Harry & Crowe, 2014; Jørgensen et al., 2017; Werry et al., 2019).

Those concerns have raised questions about the BNT's applicability among multicultural populations, and ethnic group differences in neuropsychological test performance have been widely documented (Boone et al., 2007; Campbell et al., 2002; Gladsjo et al., 1999; Norman et al., 2011). The inherent variance in language and cognition within these populations underscores the importance of standardized adaptations, exemplified by the adaptations made for the BNT in Spanish (Allegri et al., 1997), French (Thuillard Colombo & Assal, 1992), Chinese (Cheung et al., 2004), and Korean (Kang et al., 2000).

Although the issue of cultural validity for the English version of the BNT has received limited attention, some concerns have been acknowledged through tailored adaptations for specific countries, such as New Zealand (Barker-Collo, 2001), and for bilinguals (Barker-Collo, 2001; Roberts & Doucet, 2011; Sheppard et al., 2016). For example, New Zealanders made 60% more errors on the BNT items "beaver" and "pretzel" than North Americans did.

Social and cultural factors are known to have a significant impact on test performance within the field of speech-language pathology. Factors such as educational background, racial socialization, socioeconomic status, and acculturation have been shown to vary among ethnic groups, influencing both test performance and the interpretation of clinical observations (Berry, 2007; Enobi et al., 2022; Manly et al., 2004). Acculturation, a multifaceted construct, plays a pivotal role in shaping an individual's cultural and psychological adaptation to their environment. As individuals navigate contact with multiple cultural groups or members, they undergo complex changes that affect their beliefs, attitudes, and values (Berry, 2007; Manly et al., 2004). In the context of speech-language pathology, acculturation levels are closely linked to the adoption of linguistic, cognitive, and behavioural characteristics associated with the dominant culture.

The need for cultural adaptations of tests is particularly relevant in the North American context. Both the United States and Canada are experiencing a growing level of ethnic diversity, particularly within younger generations. In the United States, the population distribution is estimated as follows: 58.9% non-Hispanic White, 19.1% Hispanic or Latino, 13.6% Black or African American, 6.3% Asian, and 3% identifying as two or more races (United States Census Bureau, n.d.). The proportion of people under the age of 15 years who identified as non-White exceeded 50% for the first time in 2018 (Frey, 2019). In Canada, proportions are slightly different, but ethnocultural diversity stays a dominant core value for the Canadian society. As of 2021, 25% of the total population identified as part of a racialized group, with South Asian, Chinese, and Black populations collectively representing 16.1% of the Canadian population (Statistics Canada, 2022).

The BNT adaptations mentioned above, tailored for specific targets like New Zealand or French-English bilinguals, may not be suitable for clinicians practising in multicultural settings in North America. In cases where those adaptations could be relevant, clinicians might face additional challenges in clinical settings where the test or its interpretation guidelines are not available. Time and resources constraints may also hinder clinicians from readily incorporating or seeking out adaptations. Finally, many North American residents do speak some level of English as a second language, leading clinicians to assume they can effectively use the English version of the test. However, extensive research has demonstrated significant differences in bilingual performance on specific languages (Kohnert et al., 1998; Roberts et al., 2002).

Considering these issues, the aim of this systematic review is to examine the psychometric properties of the English BNT (Kaplan et al., 2001) used in the North American context. The intent is to enlighten clinicians on the relevance and applicability of the BNT in multiethnic populations in Canada and the United States.

Methodology

Measurement Tool

The BNT comprises 60 black-and-white drawings of objects that participants are instructed to sequentially name. Each drawing is displayed individually, and participants have a 20-s window to provide a response. In the event of an error, two forms of cues are available to participants. If the error is attributed to poor perception or an inability to recognize the object (such as responding with "I don't know what it is"), a descriptive cue is provided. For example, if a participant responds to a picture of a mushroom with "umbrella," the examiner may say, "It looks a little like that, but what the artist meant to draw was something to eat," and the participant is given an additional 20 s to correctly name the object. A phonemic cue, involving the first consonant and vowel of the word, may be offered to participants who fail to respond or produce an incorrect response. The use of the phonemic cue does not contribute to the total score, as clarified by Lezak et al. (2004), Mitrushina et al. (2005), and Strauss et al. (2006). Instead, it primarily serves as a tool to guide the clinician in interpreting the level of impairment during clinical assessments.

The presentation order of the items is said to go from least difficult to most difficult, based on the frequency of the related word, from high-frequency nouns (e.g., bed) to low-frequency ones (e.g., lattice). Although an extensive review of technical manuals and neuropsychological literature did not reveal how item frequency was determined for the BNT, Yochim et al. (2013) found no noticeable decrease in their frequency from items 30 to 60 in their examination of the matter. This finding is noteworthy because individuals experiencing anomia often face challenges in naming low-frequency common nouns, such as "asparagus."

Research Strategy

The research strategy was developed according to the COSMIN initiative recommendations (COSMIN, n.d.; Mokkink, Prinsen, et al., 2018). It was carried out according to the four steps suggested for conducting the search: (a) formulation of the objective of the review, (b) formulation of the eligibility criteria, (c) searching of scientific literature, and (d) selection on the basis of the abstract, then the full text (Mokkink, Prinsen, et al., 2018). A consulting librarian affiliated with Université Laval supported the completion of these steps. Her expertise in review methods and specialization in the healthcare field improved the accuracy, comprehensiveness, and relevance of the search according to the study objective.

An initial list of keywords and descriptors was identified via five articles from a preliminary review. To capture all the psychometric properties of the tool, the search filter proposed by COSMIN (Terwee et al., 2009) was used, in addition to the terms identified by the team. The terminology and search equations are presented in **Table 1**. The search strategy was deployed across three databases (Medline, CINAHL, and Web of Science). The databases were selected according to the specificity of the BNT to the healthcare field. EndNote software (https://endnote.com) was used to collate all references from the databases. The databases were consulted on February 2, 2023.

Eligibility Criteria

The included studies met these criteria: (a) the English BNT was used with adults (healthy or pathological), (b) the BNT was studied in a Canadian or American context, and (c) the BNT's psychometric qualities were evaluated. Studies were excluded if they met any of these criteria: (a) The article was published in 2000 or before, corresponding with the publication of the most recent edition of the test (Kaplan et al., 2001); (b) The BNT was used only as a comparator for the validation of another tool; (c) the article was in a language other than French or English; (d) the article studied the psychometric qualities of a battery of tests including the BNT (i.e., not the BNT alone); or (e) the document was not a scientific journal article. Reviews and meta-analyses were not included, but their reference lists were consulted to allow the addition of articles not identified by the initial search.

Table 1

Search Strategy

Search terms	Free/controlled vocabulary	Database(s) searched
(Boston naming test) AND ((valid*) OR (hypothese test) OR (Interpreta*) OR (consistenc*) OR (unidimensional*) OR (measurement variance or measurement invariance) OR (reliab* OR unreliab*) OR (Comprehensi*) OR (Relevanc*) OR (discrimin*) OR (responsive*) OR (coefficient of variation) OR (instrumentation) OR (validation stud*) OR (comparative stud*) OR (psychometr*) OR (clinimetr*) OR (outcome assessment OR outcome measure) OR (observer variation) OR ((repeatab* OR repeated OR replicab*) N3 (measure OR measures OR findings OR result* OR test*)) OR (homogen*) OR (agreement) OR (imprecision OR precision) OR (precise value) OR (test-retest OR test retest OR (test N2 retest)) OR (stability) OR (inter*) OR (intra*) OR (generali*) OR (concordance) OR (correlation*) OR (factor analys*) OR (factor structure) OR (scaling N3 analys*) OR (error*) OR (variab*) OR (measur* N3 uncertainty) OR (sensitiv*) OR (limit* N3 detect*) OR ((minimal* OR clinical*) N3 (important OR significant OR detectable) N3 (change OR difference)) OR (small* N3 (real OR detectable) N3 (change OR difference)) OR (meaningful change) OR (ceiling effect) OR (floor effect) OR (item response model) OR (differential item functioning))	Free vocabulary	Medline CINAHL Web of Sciences
(MH "Measurement Issues and Assessments+") OR (MH "Instrument Validation") OR (MH "Instrument Scaling") OR (MH "Instrument Construction+") OR (MH "Instrument Adaptation") OR (MH "Statistical Significance+") OR (MH "Data Analysis+") OR (MH "Comparative Studies") OR (MH "Outcome Assessment") OR (MH "Inferential Statistics") OR (MH "Variable+") OR (MH "Sensitivity and Specificity") OR (MH "Bias (Research)+") OR (MH "Reproducibility of Results")	Controlled vocabulary	Medline CINAHL

Note. CINAHL = Cumulated Index to Nursing and Allied Health Literature; MH = subject heading.

Data Extraction and Critical Appraisal of Selected Articles

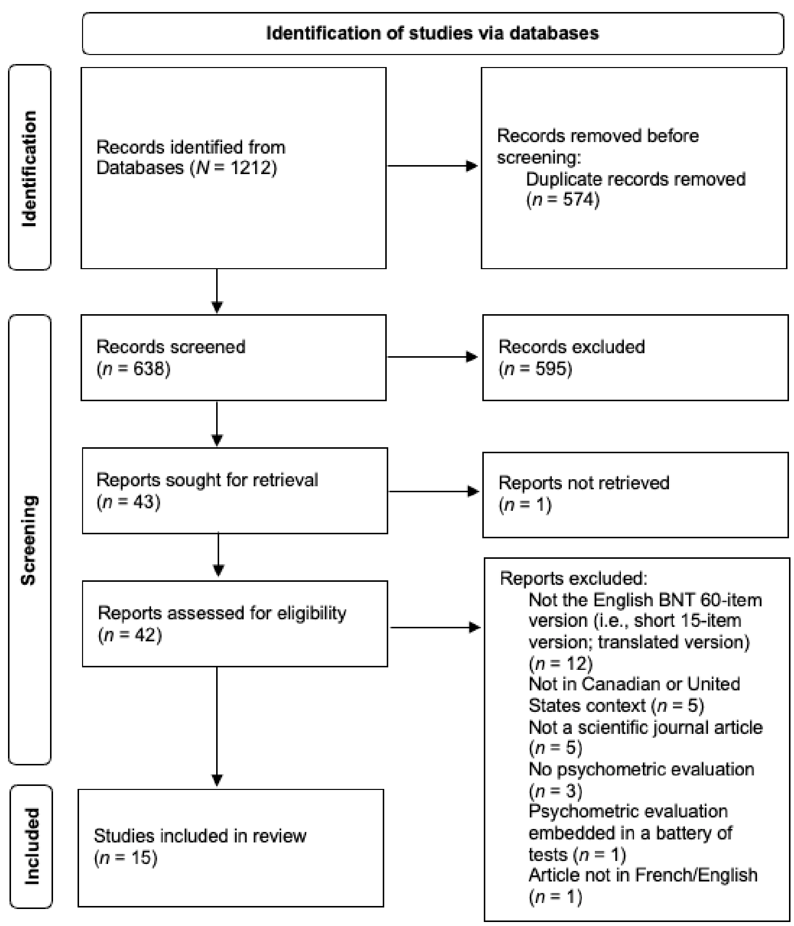
A data extraction grid was developed jointly by the three team members based on the COSMIN guide for assessing psychometric qualities (Mokkink, Prinsen, et al., 2018). From the selected articles, relevant information was systematically identified, such as first author, year of publication, country, study type, objectives, assessment of referred psychometric qualities, sample size, main results, limitations, and relevance of the psychometric qualities investigated. The selection and extraction process were conducted by three authors (MS, DG, AL; PhD students) under the supervision of two professors who are experts in psychometric quality. Each article underwent data extraction by two reviewers. Conflicts were resolved by consensus through discussions among the three reviewers in a dedicated meeting. The extraction results were then compiled and organized in an Excel spreadsheet. To assess the quality of the included articles, a critical appraisal of the articles was conducted. This process was carried out based on the COSMIN risk of bias checklist (Mokkink, de Vet, et al., 2018).

Results

Study Selection and Critical Appraisal of Selected Articles

From the search of the databases, 1212 relevant references were found. Duplicates were removed, leaving 638 articles to be evaluated. The articles were then screened based on title and abstract. Articles that complied with the inclusion criteria previously mentioned were moved to the next assessment step. The remaining 43 articles were then assessed (full-text review) to ensure no exclusion criteria were present. During this final sorting stage, 27 articles were excluded, and a document was not available for one article, leaving a total of 15 included articles. The reasons for exclusion are documented in the PRISMA diagram in **Figure 1**. **Table 2** provides a concise overview of the psychometric properties discussed in the selected articles for analysis. None of the selected articles evaluated content validity, measurement error, or test-retest reliability. We present below results obtained for psychometric properties discussed in the selected articles.

Figure 1



Selection Process PRISMA Diagram

Note. BNT = Boston Naming Test. PRISMA diagram format from Page et al. (2021).

Table 2

Psychometric Properties Studied in the Selected Articles

Property	Definition
Reliability	Reliability is the accuracy and reproducibility of the tool over time and in different contexts. It includes assessment of measurement errors and test-retest reliability (Mokkink, Prinsen, et al., 2018).
Reliability by internal consistency	Internal consistency refers to the homogeneity of a set of statements (items) used solely to measure different aspects of the same variable. The higher the Cronbach's alpha values, the better the internal consistency of a psychometric tool (Moussa, 2008).
Content validity	Content validity refers to the extent to which proposed items reflect the phenomenon targeted by the test (Laveault & Grégoire, 2002).
Structural validity	The examination of structural validity verifies whether the tool's items are just and unidimensional reflections of the construct it claims to measure (Mokkink, Prinsen, et al., 2018).
Discriminant validity	This type of construct validity is concerned with the ability to discriminate between groups with known characteristics (Mokkink, Prinsen, et al., 2018).
Cross-cultural validity	The examination of cross-cultural validity verifies whether performance on the tool's items is consistent with the real abilities of individuals with different characteristics that should not affect test performance (Mokkink, Prinsen, et al., 2018).
Validity by convergence	Validity by convergence involves establishing whether two tests measuring an identical or similar construct produce similar results (Mokkink, Prinsen, et al., 2018).
Validity by divergence	Validity by divergence verifies whether two tests that do not measure the same construct generate uncorrelated measures (Mokkink, Prinsen, et al., 2018).
Sensitivity to change	The prescription of any treatment implies an intent to change the patient's state. The theoretical and statistical concept upon which the demonstration of a modification in a state through a measuring tool is called sensitivity to change (Chomel-Guillaume et al., 2021).

The results of the critical appraisal of the included studies are presented in **Table 3**. A sole article focused on reliability, and none discussed sensitivity to change. The data presented on criterion validity were adequate overall, whereas the data on construct validity tended toward the dubious and inadequate. Although little discussed, the data on reliability were judged to be very good.

Reliability by Internal Consistency

Pedraza et al. (2011) obtained a Cronbach's alpha coefficient of 0.91 for the 60 BNT stimuli, suggesting good internal consistency of the tool.

Structural Validity

Pedraza et al. (2011) performed an exploratory factor analysis that indicated sufficient unidimensionality of the BNT. A confirmatory factor analysis based on maximum likelihood estimates generated a compatible one-factor model (comparative fit index = 0.97; root mean square error of approximation = 0.02; Satorra-Bentler scaled x 2 = 1717.61, $p = .001$).

Discriminant Validity (Known Groups)

Several authors highlighted the discrimination qualities of BNT (see **Table 4**).

To determine the diagnostic utility of BNT for hemispheric lateralization of temporal lobe epilepsy, BNT performance was compared to medical imaging (magnetic resonance imaging, electroencephalogram) and to the actual side of the lesion attested during lobectomy. Regression analysis showed the BNT correctly predicted the side of the lesion in 68.8% ($n = 317$) of patients in both cohorts (left and right lateralization; Busch et al., 2005; Busch et al., 2009). Two studies reported sensitivity and specificity values for lateralizing left temporal epilepsy: 57% and 59% respectively (Umfleet et al., 2015) and 58% and 70% (Loring et al., 2008). Nussbaum et al. (2022) highlighted differences on BNT performance between people with mild to moderate traumatic brain injury (TBI) and people with severe TBI.

Three of the studies (Abramson et al., 2022; Erdodi et al., 2018; Nussbaum et al., 2022) investigated the classification

Table 3

Critical Appraisal of Boston Naming Test's Reliability and Validity

Authors	Country	Sample description	Reliability	Validity	
				Criteria	Construct
Abramson et al. (2022)	United States	Patients evaluated in psychiatry and neuropsychology; heterogeneous diagnoses (*N* = 137)		Adequate	Doubtful
Boone et al. (2007)	United States	Patients referred for outpatient neuropsychological assessment in a public hospital (*N* = 161)			Doubtful
Busch et al. (2009)	United States	Patients with left temporal epilepsy who underwent lobectomy (*N* = 104)		Adequate	
Devora et al. (2022)	United States	Adults (56+) investigated for a neurodegenerative condition (*N* = 86)			Doubtful
Durant et al. (2021)	United States	Individuals referred for neuropsychological evaluation for memory complaints. Heterogeneous diagnoses (Alzheimer's, Lewy body dementia and Parkinson's; *N* = 105)		Adequate	Doubtful
Erdodi et al. (2018)	United States	Adults referred for neuropsychological evaluation (*N* = 214)		Very good	
Loring et al. (2008)	United States	Patients with left temporal epilepsy who underwent anterior temporal lobe lobectomy (*n* left = 69; *n* right = 66; total *N* = 135)		Doubtful	Doubtful
Na & King (2019)	United States	Healthy young adults (*N* = 83)			Inadequate
Pedraza et al. (2009)	United States	Older adults without cognitive impairment (*N* = 670)			Very good
Pedraza et al. (2011)	United States	Outpatient neuropsychology referrals (50% referred for dementia assessment, remainder including epilepsy, depression, poststroke status, etc.; *N* = 300)		Very good	Inadequate
Testa et al. (2004)	United States	Individuals with Alzheimer's (*n* = 306), myocardial infarction (*n* = 67), and normal individuals (*n* = 409) who had at least 2 annual assessments (*N* = 782)	Very good		
Umfleet et al. (2015)	United States	Patients with left temporal epilepsy (*N* = 143)		Adequate	Doubtful
Zec et al. (2007)	United States	Adults (*N* = 1090)			Inadequate
Zimmerman et al. (2022)	United States	Adults with head trauma (62%), stroke (17%), or other neurological or psychiatric conditions (21%; *N* = 458)			Inadequate

Table 4

Comparison of Boston Naming Test (BNT) Performance Between Known Groups

Evaluated factor	Group 1	Group 2	Salient results
Epilepsy	Left temporal involvement	Right temporal lobe involvement	Loring et al. (2008) Student's *t* test = 3.3 ($p < .001$), $d = .57$ Sensitivity: .58 Specificity: .70 Umfleet et al. (2015) Student's *t* test = 2.7 ($p = .008$), $d = .46$ Sensitivity: .57 Specificity: .59 Busch et al. (2005, 2009) Regression analysis for correct predictions: .69
Traumatic brain injury	Mild/moderate	Severe	Nussbaum et al. (2022) Fail rate: 2.55–4.07 ($p = .003–.023$)
Alzheimer's disease	With	Without	Testa et al. (2004) Raw score gap: Group 1: $M(SD) = 10.4(3.0)$ Group 2: $M(SD) = 6.6(3.4)$; $p < .001$ Alzheimer diagnosis sensibility: .63 Alzheimer diagnosis specificity: .83 Overall correct classification: .73 (4.6–10.9)
Performance validity	Valid/impaired	Invalid/ cognitive impaired	Abramson et al. (2022) Area under curve (AUC) = 0.59; score adjusted: AUC = 0.60.
	Valid/ Unimpaired	Invalid/ unimpaired	Abramson et al. (2022) AUC = 0.87 $p < .001$; score adjusted: AUC = 0.68 $p < .01$ Optimal cutoff BNT raw score ≤ 50 Sensitivity .61 Specificity .89 Using a demographically adjusted *T* score with optimal cutoff at ≤ 35, Sensitivity: .21 Specificity: .89
	Valid	Invalid	Erdodi et al. (2018) Optimal BNT raw score cutoff ≤ 50: Sensitivity: .15–.41 Specificity: .87–.95 Using a demographically adjusted *T* score with optimal cutoff at ≤ 37, Sensitivity: .15–.35 Specificity: .87–.95

Table 4 (continued)

Evaluated factor	Group 1	Group 2	Salient results
			Nussbaum et al. (2022) Optimal cutoff BNT raw score ≤ 47 Sensitivity: .24–.38 Specificity: .89–.94 Using a demographically adjusted *T* score with optimal cutoff at ≤ 33, Sensitivity: .24–.34 Specificity: .92–.93 Abramson et al. (2022) For acceptable specificity with BNT raw score cutoff ≤36, Sensitivity .07 Specificity: .90 Using a demographically adjusted *T* score with cutoff at ≤ 33 for acceptable specificity, Sensitivity: .18 Specificity: .90

accuracy of the BNT for assessing performance validity. Performance validity refers to the extent to which the results of cognitive functioning tests reflect the individual's true level of performance (Larrabee et al., 2019; Nussbaum et al., 2022). The three articles reviewed all compared the BNT to the most common measures of performance validity: Trial 1 of the Test of Memory Malingering (Abramson et al., 2022; Erdodi et al., 2018; Nussbaum et al., 2022), the Validity Index Seven (Nussbaum et al., 2022), or the Medical Symptom Validity Test (Abramson et al., 2022). When considering optimal cut-off score (raw or adjusted T), the BNT was shown to have low precision in discriminating valid from invalid performance. Although the BNT can discriminate performance validity with low-to-moderate accuracy in people without cognitive impairment, it appears it cannot do so in people with cognitive impairment (Abramson et al., 2022). Finally, Testa et al. (2004) explored the BNT's potential for predicting the diagnosis of Alzheimer's disease. When comparing the BNT with an Alzheimer's diagnostic criterion, a sensitivity of .63 and a slightly reduced specificity of .83 were obtained.

Cross-Cultural Validity

Table 5 shows the significant differences in BNT performance between ethnic groups, when assessed after controlling age, education, and sociodemographic parameters. White people generally demonstrated higher mean scores than other ethnic groups.

Two studies highlighted significant differences in several items regarding their discrimination and difficulty. Na and King (2019) demonstrated a differential functioning of 12 items on the BNT between White and African American healthy individuals, namely the items "rhinoceros," "dominoes," "escalator," "muzzle," "unicorn," "noose," "latch," "tripod," "scroll," "tongs," and "palette". They also noted significant differences in items between the two groups of healthy young adults: "sphinx" (23%); "pelican" (22%); "palette" (20%); "asparagus" (18%), and "trellis" (18%). According to the item order of the test (i.e., from least difficult to most difficult), greater variability of score between the groups was observed when assessing the most difficult items. (Na & King, 2019; Pedraza et al., 2011).

Concerns have also been raised about a specific item of the BNT: Zimmerman et al. (2022) explored the possibility of excluding item 48, "noose," from the BNT due to its historical and cultural connotations, including its association with capital punishment and intimidation of African Americans (Potok et al., 2007). The mean differences of 0.17 (SD = 0.4, Mdn = 0.05) obtained in scores when the item was removed demonstrated its significance as part of the test. However, non-White individuals were more likely to fail the item (32.1%) than White individuals (9.5%; $\chi^2(1) = 13.8$, $p < .001$). Considering the differential functioning of the item and its historical connotation, Zimmerman et al. (2022) suggested developing test versions that omit it.

Table 5

Examination of Boston Naming Test Performance Variance by Ethnic Group

Article	Measure	Overall Boston Naming Test performance by group (/60)				
		White	African American	Hispanic	Asian	Differences
Boone et al. (2007)	Score *M* (*SD*) adjusted for age/education	52.40 (1.00)	42.50 (1.70)	42.90 (1.80)	42.40 (2.30)	White > African Americans, Hispanics, Asians ANOVA $F = 13.82$, $p < .001$
Na & King (2019)	Spontaneous response score *M* (*SD*)	54.62 (3.32)	51.85 (3.99)	-	-	$t = 3.43$ $p < 0.001$; Cohen's $d = 0.75$
	Score with stimulus index *M* (*SD*)	54.78 (3.24)	52.15 (3.95)	-	-	$t = 3.31$ $p = 0.001$; Cohen's $d = 0.72$
	Score with phonemic index *M* (*SD*)	57.32 (2.41)	56.18 (2.48)	-	-	$t = 2.08$ $p = .02$ Cohen's $d = 0.47$
	Multiple choice score *M* (*SD*)	59.40 (0.95)	58.91 (1.10)	-	-	$t = 2.08$ $p = .02$ Cohen's $d = 0.47$
Pedraza et al. (2009)	Score *M* (*SD*)	52.90 (4.90)	43.30 (10.10)	-	-	$t(668) = 15.7$; $p < .001$; Cohen's $d = 1.21$

Note. ANOVA = analysis of variance.

Other sociodemographic parameters have been investigated regarding their effects on performance in the BNT. Main results are presented in **Table 6**. Significant differences were observed for the variables of acculturation, education, and age. The BNT scores were not significantly different between men and women.

Validity by Convergence

Five studies discussed this topic. **Table 7** presents the examination of the convergence of the BNT with five tools used in healthy individuals, as well as individuals with dementia or mild cognitive impairment.

Validity by Divergence

No significant correlations were found between the BNT and a fluency test ($p = .21$) and a line orientation judgment test ($p = .19$), highlighting two distinct underlying theoretical contexts (Durant et al., 2021).

Discussion

The Challenges of Inconsistent Use of BNT

The BNT serves multiple purposes beyond its primary role in object naming assessment. It is used as a diagnostic tool and a validity indicator for speech-language tests. Although this versatility underscores the tool's adaptability, it also poses challenges in assessing its specific psychometric qualities. The varied uses contribute to inconsistency in the corpus of data, and recognizing diverse applications is crucial for a comprehensive understanding of the BNT's qualities, as these vary accordingly. In addition to variations across its purposes, there are discrepancies in administration and scoring methods across different clinical environments. Indeed, this variability is noted in contexts where the test is employed for decision support in sensitive cases, such as medico-legal assessments (Nussbaum et al., 2022). This variability supports the need for a better standardization of the administration and scoring of the

Table 6

Effect of Sociodemographic Parameters on Boston Naming Test Performance

Authors	Sociodemographic parameter measured by the authors	Correlation/effect Spearman *r*, *p* value
Boone et al. (2007)	Age of English learning	−.258, .001
	Years of education in the United States	−.272, .001
	Number of years spent in the United States	−.346, .001
Nussbaum et al. (2022)	Education level	.31, < .001
	Age	−.30, < .001
	Gender	Difference of 0.42 word, group effect = .095, *p* = .08

Table 7

Correlation and Convergence of the Boston Naming Test (BNT) With Other Neuropsychological Tests

Authors	Comparator test	Construct	Results
Durant et al. (2021)	Neuropsychological Assessment Battery	Skills and cognitive functioning	Pearson $r = .68$, $p < .005$ (moderate)
Loring et al. (2008)	Multilingual Aphasia Examination Visual Naming (MAE VN)	Visual naming	BNT Cohen's $d = 0.57$, $p < .001$; MAE VN Cohen's $d = 0.36$, $p = .02$
Na & King (2019)	Woodcock-Johnson Letter-Word Identification	Reading	Pearson $r = .57$, $p < .01$ (moderate)
	Wechsler Abbreviated Scale of Intelligence Vocabulary	Vocabulary	Pearson $r = .61$, $p < .01$ (moderate)
Zimmerman et al. (2022)	BNT-Prorated	Visual naming	Concordance correlation coefficient = 0.99, $p < .001$

BNT. Addressing this concern was one of the motivations behind the development of the second version of the BNT, which included a simplified and clarified user and scoring manual (Kaplan et al., 2001).

Clinical Considerations in Group Differentiation

The BNT has been employed in numerous studies to assess language and cognitive disorders across diverse clinical populations. It exhibits qualities that enable the differentiation of groups according to factors such as the lateralization of brain damage in epilepsy and the severity of TBI. Patients with Alzheimer's disease perform more poorly on the BNT than healthy adults, but the sole result of the test is not enough to ensure adequate diagnostic classification (Testa et al., 2004). The BNT's ability to differentiate the degree of disability may be limited in populations with cognitive disorders, and this limitation poses challenges for speech-language pathologists, as the test's sensitivity to nuanced variations in cognitive impairment may affect the precision of diagnostic assessments and the tailoring of intervention strategies.

Among the various purposes of the BNT, its function as an oral picture-naming test often serves as an initial assessment in speech-language pathology to evaluate anomia and identify affected lexical items, including high-frequency words, low-frequency words, long words or short words. The test plays a crucial role in guiding the selection of additional assessments to be administered to the participant. Having limitations in its ability to comprehensively assess oral naming may therefore lead to interpretation errors for clinicians and compromise the overall evaluation of a patient. Relying solely on this measure makes it difficult to draw definitive conclusions or make precise distinctions regarding health. When the BNT is used with patients of diverse ethnic or cultural backgrounds, the results may also be even more biased due to linguistic, social, or educational differences.

Therefore, the more marginal the uses of the BNT become, the scarcer the scientific data available to substantiate its application, particularly when considering contexts beyond naming assessment, such as clinical diagnosis for conditions like Alzheimer's disease or determining the origin of lesions in epilepsy.

Generalization to Culturally and Ethnically Diverse Populations

Authors acknowledged and highlighted the recognized issues regarding the impact of sociocultural influences on BNT performance (Erdodi et al., 2018). BNT standards derived from studies predominantly involving White population samples may not be universally applicable to other ethnic groups (Boone et al., 2007). This issue is further exacerbated by the limited availability of diverse standards within the same clinical setting. In situations where the tool might be employed for forensic purposes, standards derived from studies involving predominantly White population samples should explicitly outline the boundaries of their applicability (Boone et al., 2007). On this matter, Abramson et al. (2022) pointed out the significant risk of false positives when standards designed for White or African American populations (Heaton, 2004) are applied to other populations. Indeed, Heaton (2004) reported a false-positive rate of 60% when White population standards were employed with a Hispanic population, versus 20% when African American standards were used. The data highlighted the need for caution when interpreting BNT scores across diverse racial, ethnic, and cultural groups. Moreover, replication of studies with more diverse populations is needed to accurately interpret cut-offs in BNT scores. The influence of language proficiency, acculturation variables, age, education, and cultural context must also be considered to ensure cross-cultural validity and an accurate assessment of individuals' naming abilities.

The BNT as an Indicator of Performance Validity

When comparing the BNT with measures of performance validity, the data indicated good specificity but low sensitivity. Thus, a successful completion of the BNT cannot be unequivocally relied upon as a valid indicator of performance. Conversely, BNT failure can be regarded as a significant predictor of neuropsychological issues in a patient classified as healthy (Erdodi et al., 2018; Nussbaum et al., 2022). Nevertheless, sensitivity and specificity are not absolute measures of diagnostic accuracy but rather compromises between the two, influenced by other factors. Thus, it is advisable to interpret these values in conjunction with other tools that have demonstrated more evidence of validity and reliability.

Conclusion: BNT's Strengths and Limitations

Analysis of the BNT's psychometric qualities repositions its diagnostic utility. Although several studies indicated that the BNT effectively distinguishes between groups based on different health conditions, lingering concerns remain regarding its cross-cultural validity and its ability to differentiate when diverse ethnic groups are involved. Additional studies are needed to investigate these aspects and comprehensively explore the psychometric properties of the BNT. Overall, the BNT can be used as a diagnostic tool, but it is important to consider its limitations and interpret it with care in different clinical, cultural, and population contexts. Often, the standards used are not designed for people of diverse ethnic origins and increase the risk of false positives amongst these groups. The overall value of the BNT is thus subject to scrutiny.

One potential path forward could be the development and standardization of a new object naming test that addresses these various concerns. Normalization of tests, considering individuals' cultural backgrounds within the population, is a crucial approach to ensure accuracy in speech-language evaluations. The increasing significance of immigration in Canada and the United States has led to populations becoming more ethnically and culturally diverse. However, the speech-language tests used in these countries often fail to reflect this changing demographic reality. Consequently, speech-language assessments may not adequately account for the needs and cultural contexts of immigrants and ethnic minorities, resulting in biased or unrepresentative results.

As an alternative, clinical professionals should also consider using more culturally sensitive alternatives. For instance, the Neuropsychological Assessment Battery shows strong convergence with the BNT while providing adequate validity and reliability in ethnically diverse populations (Messerly & Marceaux, 2020).

Strengths and Limitations of This Review

Among this review's strengths, the use of the systematic review method, recommendations, and search filters proposed by the COSMIN initiative (Mokkink, de Vet, et al., 2018; Mokkink, Prinsen, et al., 2018) contributed to the rigour and reproducibility of the results. Collaboration with a consulting librarian specializing in review methods further validated the quality and relevance of the search strategy. The independent evaluation of articles by two reviewers and by three reviewers during conflicts contributed to the accuracy of the selection and extraction process. Finally, the critical evaluation of article quality using the COSMIN grid promoted transparency in data interpretation.

Several limitations need to be highlighted, however. First, the number of articles included was relatively small, reducing the generalizability of the results. All the studies were carried out in Canada or the United States, limiting the applicability of the results to other contexts. Only one study included in this review examined the Canadian context, while 14 were conducted in the United States. This disproportion may impact the generalizability of our results within North America, given that populations differ between these two countries. Finally, most of the BNT's psychometric qualities were evaluated by a limited number of authors and for different uses of the BNT.

Although these results provide important food for thought on the qualities of the tool, further studies are needed to make conclusions about the qualities of the BNT when used in Canada or the United States. Nonetheless, by uncovering potential challenges and strengths related to the administration, scoring, and interpretation of the BNT in diverse settings, this review seeks to offer insights that will inform clinicians, raise awareness, and provide valuable suggestions for refining neuropsychological assessment practices. Indeed, the existence of validated versions for different populations is no guarantee of their effective application in clinical practice. As highlighted in this study, issues of access to such versions, a heightened perception of validity, or a lack of awareness of limitations can lead to accuracy challenges in clinical evaluation for culturally diverse populations (Berry, 2007; Enobi et al., 2022; Manly et al., 2004). Using appropriate alternatives, such as a version of the BNT suitable for bilinguals (Ali et al., 2022), or triangulating results with other convergent tests, can make a significant contribution to the quality of neuropsychological assessments.

References

Abramson, D. A., Resch, Z. J., Ovsiew, G. P., White, D. J., Bernstein, M. T., Basurto, K. S., & Soble, J. R. (2022). Impaired or invalid? Limitations of assessing performance validity using the Boston Naming Test. *Applied Neuropsychology: Adult, 29*(4), 486–491. https://doi.org/10.1080/23279095.2020.1774378

Albert, M. S. (1988). General issues in geriatric neuropsychology. In M. S. Albert & M. B. Moss (Eds.), *Geriatric neuropsychology* (pp. 3–10). The Guilford Press.

Alexander, M. P., & Hillis, A. E. (2008). Aphasia. In M. Aminoff, F. Boller, D. Swaab, G. Goldenberg, & B. Miller (Eds.), *Handbook of clinical neurology* (Vol. 88, pp. 287–309). Elsevier. https://doi.org/10.1016/S0072-9752(07)88014-6

Ali, S., Elliott, L., Biss, R. K., Abumeeiz, M., Brantuo, M., Kuzmenka, P., Odenigbo, P., & Erdodi, L. A. (2022). The BNT-15 provides an accurate measure of English proficiency in cognitively intact bilinguals–A study in cross-cultural assessment. *Applied Neuropsychology: Adult, 29*(3), 351–363. https://doi.org/10.1080/23279095.2020.1760277

Allegri, R. F., Villavicencio, A. F., Taragano, F. E., Rymberg, S., Mangone, C. A., & Baumann, D. (1997). Spanish Boston Naming Test norms. *The Clinical Neuropsychologist, 11*(4), 416-420. https://doi.org/10.1080/13854049708400471

Barker-Collo, S. L. (2001). The 60-item Boston naming test: Cultural bias and possible adaptations for New Zealand. *Aphasiology, 15*(1), 85–92. https://doi.org/10.1080/02687040042000124

Benton, A. L., Hamsher, K. D., & Sivan, A. B. (1994). *Multilingual Aphasia Examination, Third Edition*. Psychological Assessment Resources. https://www.parinc.com/Products/Pkey/222

Berry, J. W. (2007). Acculturation. In J. E. Grusec & P. D. Hastings (Eds.), *Handbook of socialization: Theory and research* (pp. 543–558). Guilford Press.

Boone, K. B., Victor, T. L., Wen, J., Razani, J., & Pontón, M. (2007). The association between neuropsychological scores and ethnicity, language, and acculturation variables in a large patient population. *Archives of Clinical Neuropsychology, 22*(3), 355–365. https://doi.org/10.1016/j.acn.2007.01.010

Bortnik, K. E., Boone, K. B., Wen, J., Lu, P., Mitrushina, M., Razani, J., & Maury, T. (2013). Survey results regarding use of the Boston Naming Test: Houston, we have a problem. *Journal of Clinical and Experimental Neuropsychology, 35*(8), 857–866. https://doi.org/10.1080/13803395.2013.826182

Brooks, B. L., Strauss, E., Sherman, E., Iverson, G. L., & Slick, D. J. (2009). Developments in neuropsychological assessment: Refining psychometric and clinical interpretive methods. *Canadian Psychology/Psychologie Canadienne, 50*(3), 196–209. https://doi.org/10.1037/a0016066

Busch, R. M., Frazier, T. W., Haggerty, K. A., & Kubu, C. S. (2005). Utility of the Boston Naming Test in predicting ultimate side of surgery in patients with medically intractable temporal lobe epilepsy. *Epilepsia, 46*(11), 1773–1779. https://doi.org/10.1111/j.1528-1167.2005.00300.x

Busch, R. M., Frazier, T. W., Iampietro, M. C., Chapin, J. S., & Kubu, C. S. (2009). Clinical utility of the Boston Naming Test in predicting ultimate side of surgery in patients with medically intractable temporal lobe epilepsy: A double cross-validation study. *Epilepsia, 50*(5), 1270–1273. https://doi.org/10.1111/j.1528-1167.2008.01865.x

Campbell, A. L., Ocampo, C., Rorie, K. D., Lewis, S., Combs, S., Ford-Booker, P., Briscoe, J., Lewis-Jack, O., Brown, A., & Wood, D. (2002). Caveats in the neuropsychological assessment of African Americans. *Journal of the National Medical Association, 94*(7), 591–601. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2594299/

Caramazza, A., & Hillis, A. E. (1991). Lexical organization of nouns and verbs in the brain. *Nature, 349* (6312), 788–790. https://doi.org/10.1038/349788a0

Chedid, G., Wilson, M. A., Bedetti, C., Rey, A. E., Vallet, G. T., & Brambati, S. M. (2019). Norms of conceptual familiarity for 3,596 French nouns and their contribution in lexical decision. *Behavior Research Methods, 51*, 2238–2247. https://doi.org/10.3758/s13428-018-1106-8

Cheung, R. W., Cheung, M.-C., & Chan, A. S. (2004). Confrontation naming in Chinese patients with left, right or bilateral brain damage. *Journal of the International Neuropsychological Society, 10*(1), 46–53. https://doi.org/10.1017/s1355617704101069

Chomel-Guillaume, S., Leloup, G., & Bernard, I. (2021). *Les aphasies : évaluation et rééducation.* Elsevier Health Sciences. https://www.elsevier-masson.fr/les-aphasies-9782294756108.html

COSMIN. (n.d.). About the COSMIN initiative. https://www.cosmin.nl/about/

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

Devora, P. V., O'Mahar, K., Karboski, S. M., Benge, J. F., & Hilsabeck, R. C. (2022). Correspondence of the Boston Naming Test and Multilingual Naming Test in identifying naming impairments in a geriatric cognitive disorders clinic. *Applied Neuropsychology: Adult,* 1–7. https://doi.org/10.1080/23279095.2022.2130318

Durant, J., Berg, J. L., Banks, S. J., Kaylegian, J., & Miller, J. B. (2021). Comparing the Boston Naming Test with the Neuropsychological Assessment Battery-Naming Subtest in a neurodegenerative disease clinic population. *Assessment, 28*(5), 1256–1266. https://doi.org/10.1177/1073191119872253

Enobi, Y., Kemmotsu, N., Robinson, E., & Murphy, C. (2022). Effects of language and acculturation on neurocognitive performance of Japanese Americans. *Neuropsychology, 36*(7), 651–663. https://doi.org/10.1037/neu0000839

Erdodi, L. A., Dunn, A. G., Seke, K. R., Charron, C., McDermott, A., Enache, A., Maytham, C., & Hurtubise, J. L. (2018). The Boston Naming Test as a measure of performance validity. *Psychological Injury & Law, 11*(1), 1–8. https://doi.org/10.1007/s12207-017-9309-3

Fastenau, P. S., Denburg, N. L., & Mauer, B. A. (1998). Parallel short forms for the Boston Naming Test: Psychometric properties and norms for older adults. *Journal of Clinical and Experimental Neuropsychology, 20*(6), 828–834. https://doi.org/10.1076/jcen.20.6.828.1105

Ferman, T. J., Ivnik, R. J., & Lucas, J. A. (1998). Boston Naming Test discontinuation rule: Rigorous versus lenient interpretations. *Assessment, 5*(1), 13–18. https://doi.org/10.1177/107319119800500103

Fernández, A. L., & Fulbright, R. L. (2015). Construct and concurrent validity of the Spanish adaptation of the Boston Naming Test. *Applied Neuropsychology: Adult, 22*(5), 355–362. https://doi.org/10.1080/23279095.2014.939178

Frey, W. H. (2019, June 24). *Less than half of US children are White, census shows.* Brookings. https://www.brookings.edu/articles/less-than-half-of-us-children-under-15-are-white-census-shows/

Ghasisin, L., Yadegari, F., Rahgozar, M., Nazari, A., & Rastegarianzade, N. (2015). A new set of 272 pictures for psycholinguistic studies: Persian norms for name agreement, image agreement, conceptual familiarity, visual complexity, and age of acquisition. *Behavior Research Methods, 47*, 1148–1158. https://doi.org/10.3758/s13428-014-0537-0

Gladsjo, J. A., Schuman, C. C., Evans, J. D., Peavy, G. M., Miller, S. W., & Heaton, R. K. (1999). Norms for letter and category fluency: Demographic corrections for age, education, and ethnicity. *Assessment, 6*(2), 147–178. https://doi.org/10.1177/107319119900600204

Goodglass, H., Kaplan, E., & Barresi, S. (2001). *BDAE-3: The Boston Diagnostic Aphasia Examination:* Third edition. Pro-ed. https://www.proedinc.com/Products/11850/bdae3-boston-diagnostic-aphasia-examinationthir.aspx

Graves, R., Bezeau, S., Fogarty, J., & Blair, R. (2004). Boston Naming Test short forms: A comparison of previous forms with new item response theory based forms. *Journal of Clinical and Experimental Neuropsychology, 26*(7), 891–902. https://doi.org/10.1080/13803390490510716

Harry, A., & Crowe, S. F. (2014). Is the Boston Naming Test still fit for purpose? *The Clinical Neuropsychologist, 28*(3), 486–504. https://doi.org/10.1080/13854046.2014.892155

Heaton, R. K., Walden Miller, S., Taylor, M. J., & Grant, I. (2004). *Revised comprehensive norms for an expanded Halstead-Reitan Battery: Demographically adjusted neuropsychological norms for African American and Caucasian adults, professional manual.* Psychological Assessment Resources. https://www.parinc.com/Products/Pkey/357

Hirsh, K. W., & Funnell, E. (1995). Those old, familiar things: Age of acquisition, familiarity and lexical access in progressive aphasia. *Journal of Neurolinguistics, 9*(1), 23–32. https://doi.org/10.1016/0911-6044(95)00003-8

Jørgensen, K., Johannsen, P., & Vogel, A. (2017). A Danish adaptation of the Boston Naming Test: Preliminary norms for older adults and validity in mild Alzheimer's disease. *Clinical Neuropsychologist, 31*(sup1), 72–87. https://doi.org/10.1080/13854046.2017.1371337

Kang, Y., Kim, H., & NA, D.-L. (2000). Parallel short forms for the Korean-Boston Naming Test (K-BNT). *Journal of the Korean Neurological Association, 18*(2)144–150. https://www.jkna.org/journal/view.php?number=4541

Kaplan, E., Goodglass, H., & Weintraub, S. (1983). *Boston Naming Test.* Lea & Febiger.

Kaplan, E., Goodglass, H., & Weintraub, S. (2001). *Boston Naming Test: Second edition*. Pro-ed. https://www.proedinc.com/Products/11870/boston-naming-testsecond-edition.aspx

Kertesz, A. (2007). *Western Aphasia Battery-Revised.* Pearson. https://www.pearsonassessments.com/store/usassessments/en/Store/Professional-Assessments/Speech-%26-Language/Western-Aphasia-Battery-Revised/p/100000194.html

Kohnert, K. J., Hernandez, A. E., & Bates, E. (1998). Bilingual performance on the Boston Naming Test: Preliminary norms in Spanish and English. *Brain and Language, 65*(3), 422–440. https://doi.org/10.1006/brln.1998.2001

Lansing, A. E., Ivnik, R. J., Cullum, C. M., & Randolph, C. (1999). An empirically derived short form of the Boston Naming Test. *Archives of Clinical Neuropsychology, 14*(6), 481–487. https://doi.org/10.1016/S0887-6177(98)00022-5

Larrabee, G. J., Rohling, M. L., & Meyers, J. E. (2019). Use of multiple performance and symptom validity measures: Determining the optimal per test cutoff for determination of invalidity, analysis of skew, and inter-test correlations in valid and invalid performance groups. *The Clinical Neuropsychologist, 33*(8), 1354–1372. https://doi.org/10.1080/13854046.2019.1614227

Laveault, D., & Grégoire, J. (2002). *Introduction aux théories des tests en psychologie et en sciences de l'éducation* (2e éd.). De Boeck Supérieur.

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

Loring, D. W., Strauss, E., Hermann, B. P., Barr, W. B., Perrine, K., Trenerry, M. R., Chelune, G., Westerveld, M., Lee, G. P., Meador, K. J., & Bowden, S. C. (2008). Differential neuropsychological test sensitivity to left temporal lobe epilepsy. *Journal of the International Neuropsychological Society, 14*(3), 394–400. https://pubmed.ncbi.nlm.nih.gov/18419838/

Mack, W. J., Freed, D. M., Williams, B. W., & Henderson, V. W. (1992). Boston Naming Test: Shortened versions for use in Alzheimer's disease. *Journal of Gerontology, 47*(3), P154–P158. https://doi.org/10.1093/geronj/47.3.p154

Manly, J. J., Byrd, D. A., Touradji, P., & Stern, Y. (2004). Acculturation, reading level, and neuropsychological test performance among African American elders. *Applied Neuropsychology, 11*(1), 37–46. https://doi.org/10.1207/s15324826an1101_5

Messerly, J., & Marceaux, J. C. (2020). Examination of the reliability and validity of the NAB Naming Test in a diverse clinical sample. *The Clinical Neuropsychologist, 34*(2), 406–422. https://doi.org/10.1080/13854046.2019.1635647

Mesulam, M. M. (2016). Aphasia: Sudden and progressive. In J. Stein, D. Bennett, C. Coen, R. Dunbar, G. Goodwin, M. Husain, E. Mann, J. Morris, E. Rolls, J. S. H. Taylor, & V. Walsh (Eds.), *The curated reference collection in neuroscience and biobehavioral psychology* (pp. 517–521). Elsevier Science.

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

Mokkink, L. B., de Vet, H. C., Prinsen, C. A. C., Patrick, D. L., Alonso, J., Bouter, L. M., & Terwee, C. B. (2018). COSMIN risk of bias checklist for systematic reviews of patient-reported outcome measures. *Quality of Life Research, 27*, 1171–1179. https://doi.org/10.1007/s11136-017-1765-4

Mokkink, L. B., Prinsen, C. A. C., Patrick, D. L., Alonso, J., Bouter, L. M., de Vet, H. C., Terwee, C. B., & Mokkink, L. (2018). *COSMIN methodology for systematic reviews of patient-reported outcome measures (PROMs)* [User manual]. COSMIN. https://cosmin.nl/wp-content/uploads/COSMIN-syst-review-for-PROMs-manual_version-1_feb-2018.pdf

Moussa, S. (2008, March 7–9). L'alpha de Cronbach et l'estimation de son intervalle de confiance : L'étude de deux approches [Paper presentation; Cronbach's alpha and the estimation of its confidence interval: The study of two approaches]. *Les 4èmes journées de l'entreprise organisées par l'Association de Recherches pour le Développement des Affaires,* Gafsa, Tunisie. https://www.researchgate.net/publication/333308913_L'alpha_de_Cronbach_et_l'estimation_de_son_intervalle_de_confiance_L'etude_de_deux_approches

Na, S., & King, T. Z. (2019). Performance discrepancies on the Boston Naming Test in African-American and non-Hispanic White American young adults. *Applied Neuropsychology: Adult, 26*(3), 236–246. https://doi.org/10.1080/23279095.2017.1393427

Norman, M. A., Moore, D. J., Taylor, M., Franklin Jr, D., Cysique, L., Ake, C., Lazarretto, D., Vaida, F., Heaton, R. K., & the HNRC Group. (2011). Demographically corrected norms for African Americans and Caucasians on the Hopkins Verbal Learning Test–Revised, Brief Visuospatial Memory Test–Revised, Stroop Color and Word Test, and Wisconsin Card Sorting Test 64-card version. *Journal of Clinical and Experimental Neuropsychology, 33*(7), 793–804. https://doi.org/10.1080/13803395.2011.559157

Nussbaum, S., May, N., Cutler, L., Abeare, C. A., Watson, M., & Erdodi, L. A. (2022). Failing performance validity cutoffs on the Boston Naming Test (BNT) is specific, but insensitive to non-credible responding. *Developmental Neuropsychology, 47*(1), 17–31. https://doi.org/10.1080/87565641.2022.2038602

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, Article 71. https://doi.org/10.1136/bmj.n71

Pedraza, O., Graff-Radford, N. R., Smith, G. E., Ivnik, R. J., Willis, F. B., Petersen, R. C., & Lucas, J. A. (2009). Differential item functioning of the Boston Naming Test in cognitively normal African American and Caucasian older adults. *Journal of the International Neuropsychological Society, 15*(5), 758–768. https://doi.org/10.1017/S1355617709990361

Pedraza, O., Sachs, B. C., Ferman, T. J., Rush, B. K., & Lucas, J. A. (2011). Difficulty and discrimination parameters of Boston Naming Test items in a consecutive clinical series. *Archives of Clinical Neuropsychology, 26*(5), 434–444. https://doi.org/10.1093/arclin/acr042

Pompéia, S., Miranda, M. C., & Bueno, O. F. A. (2001). A set of 400 pictures standardised for Portuguese: Norms for name agreement, familiarity and visual complexity for children and adults. *Arquivos de Neuro-Psiquiatria, 59*(2B), 330–337. https://doi.org/10.1590/S0004-282X2001000300004

Potok, M., Visconti, L., Frankel, B., & Holmes, N. (2007, November 25). The geography of hate. *The New York Times*. https://www.nytimes.com/2007/11/25/opinion/25potok.html

Rabin, L. A., Barr, W. B., & Burton, L. A. (2005). Assessment practices of clinical neuropsychologists in the United States and Canada: A survey of INS, NAN, and APA Division 40 members. *Archives of Clinical Neuropsychology, 20*(1), 33–65. https://doi.org/10.1016/j.acn.2004.02.005

Rabin, L. A., Paolillo, E., & Barr, W. B. (2016). Stability in test-usage practices of clinical neuropsychologists in the United States and Canada over a 10-year period: A follow-up survey of INS and NAN members. *Archives of Clinical Neuropsychology, 31*(3), 206–230. https://doi.org/10.1093/arclin/acw007

Roberts, P. M., & Doucet, N. (2011). Performance of French-speaking Quebec adults on the Boston Naming Test. *Canadian Journal of Speech-Language Pathology & Audiology, 35*(3), 254–267. https://www.cjslpa.ca/detail.php?ID=1078

Roberts, P. M., Garcia, L. J., Desrochers, A., & Hernandez, D. (2002). English performance of proficient bilingual adults on the Boston Naming Test. *Aphasiology, 16*(4-6), 635–645. https://doi.org/10.1080/02687030244000220

Saxton, J., Ratcliff, G., Munro, C. A., Coffey, E. C., Becker, J. T., Fried, L., & Kuller, L. (2000). Normative data on the Boston Naming Test and two equivalent 30-item short forms. *The Clinical Neuropsychologist, 14*(4), 526–534. https://doi.org/10.1076/clin.14.4.526.7204

Sheppard, C., Kousaie, S., Monetta, L., & Taler, V. (2016). Performance on the Boston Naming Test in bilinguals. *Journal of the International Neuropsychological Society, 22*(3), 350–363. https://doi.org/10.1017/s135561771500123x

Statistics Canada. (2022, October 26). *The Canadian census: A rich portrait of the country's religious and ethnocultural diversity*. https://www150.statcan.gc.ca/n1/daily-quotidien/221026/dq221026b-eng.htm

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

Teng, E. L., Wimer, C., Roberts, E., Damasio, A. R., Eslinger, P. J., Folstein, M. F., Tune, L. E., Whitehouse, P. J., Bardolph, E. L., & Chui, H. C. (1989). Alzheimer's dementia: Performance on parallel forms of the dementia assessment battery. *Journal of Clinical and Experimental Neuropsychology, 11*(6), 899–912. https://doi.org/10.1080/01688638908400943

Terwee, C. B., Jansma, E. P., Riphagen, I. I., & de Vet, H. C. (2009). Development of a methodological PubMed search filter for finding studies on measurement properties of measurement instruments. *Quality of Life Research, 18*, 1115–1123. https://doi.org/10.1007/s11136-009-9528-5

Testa, J. A., Ivnik, R. J., Boeve, B., Petersen, R. C., Pankratz, V. S., Knopman, D., Tangalos, E., & Smith, G. E. (2004). Confrontation naming does not add incremental diagnostic utility in MCI and Alzheimer's disease. *Journal of the International Neuropsychological Society, 10*(4), 504–512. https://doi.org/10.1017/S1355617704104177

Thuillard Colombo, F. T., & Assal, G. (1992). Adaptation française du test de dénomination de Boston. Versions abrégées. *Revue européenne de psychologie appliquée, 42*(1), 67–73.

Umfleet, L. G., Janecek, J. K., Quasney, E., Sabsevitz, D. S., Ryan, J. J., Binder, J. R., & Swanson, S. J. (2015). Sensitivity and specificity of memory and naming tests for identifying left temporal-lobe epilepsy. *Applied Neuropsychology: Adult, 22*(3), 189–196. https://doi.org/10.1080/23279095.2014.895366

United States Census Bureau. (n.d.). *QuickFacts: United States*. Retrieved March 11, 2023 from https://www.census.gov/quickfacts/fact/table/US/PST045222

Werry, A. E., Daniel, M., & Bergström, B. (2019). Group differences in normal neuropsychological test performance for older non-Hispanic White and Black/African American adults. *Neuropsychology, 33*(8), 1089–1100. https://doi.org/10.1037/neu0000579

Yochim, B. P., Rashid, K., Raymond, N. C., & Beaudreau, S. A. (2013). How frequently are words used on naming tests used in spoken conversation? *The Clinical Neuropsychologist, 27*(6), 973–987. https://doi.org/10.1080/13854046.2013.797501

Zec, R. F., Burkett, N. R., Markwell, S. J., & Larsen, D. L. (2007). A cross-sectional study of the effects of age, education, and gender on the Boston Naming Test. *The Clinical Neuropsychologist, 21*(4), 587–616. https://doi.org/10.1080/13854040701220028

Zimmerman, D., Attridge, J., Rolin, S., & Davis, J. (2022). Psychometric equivalence of standard and prorated Boston Naming Test scores. *Assessment, 29*(3), 527–534. https://doi.org/10.1177/1073191120983925

Authors' Note

Correspondence concerning this article should be addressed to Manon Spigarelli, 525 Bd. Wilfrid-Hamel, Québec, QC, CANADA, G1M 2S8.
Email: **manon.spigarelli.1@ulaval.ca**

Acknowledgments

This project was supported by a scholarship (MS) from the Fonds de recherche du Québec–Santé.

Disclosures

No conflicts of interest, financial or otherwise, are declared by the authors.