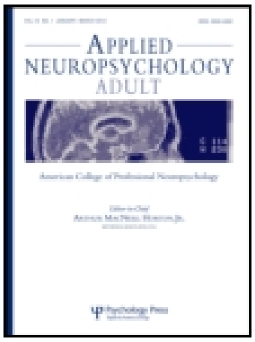

Routledge
Taylor & Francis Group

Applied Neuropsychology: Adult

ISSN: 2327-9095 (Print) 2327-9109 (Online) Journal homepage: http://www.tandfonline.com/loi/hapn21

Construct and Concurrent Validity of the Spanish Adaptation of the Boston Naming Test

Alberto Luis Fernández & Richard Leroy Fulbright

To cite this article: Alberto Luis Fernández & Richard Leroy Fulbright (2015) Construct and Concurrent Validity of the Spanish Adaptation of the Boston Naming Test, Applied Neuropsychology: Adult, 22:5, 355-362, DOI: 10.1080/23279095.2014.939178

To link to this article: http://dx.doi.org/10.1080/23279095.2014.939178

Published online: 10 Feb 2015.

APPLIED NEUROPSYCHOLOGY: ADULT, *22*: 355–362, 2015
Copyright © Taylor & Francis Group, LLC
ISSN: 2327-9095 print/2327-9109 online
DOI: 10.1080/23279095.2014.939178

Construct and Concurrent Validity of the Spanish Adaptation of the Boston Naming Test

Alberto Luis Fernández
Universidad Católica de Córdoba; Cortex Foundation; and Universidad Nacional de Córdoba, Córdoba, Argentina

Richard Leroy Fulbright
Fulbright & Associates, P.C., Dallas, Texas

In 1996, a Spanish edition of the Boston Naming Test (BNT) was published. Changes to the original version were introduced but without an empirical foundation supporting these modifications. This version was later adapted by reordering the items according to their difficulty frequency in a Spanish-speaking population, and norms were developed from this version of the test. However, no validity studies were performed with this new version to determine if the changes introduced had an impact on the test's validity. The current study was performed to assess the construct and concurrent validity of the Spanish-language version of the BNT (SV-BNT). The SV-BNT was administered to 23 patients with Alzheimer's disease (AD) and 36 normal controls. Groups were matched by age and education. Patients with AD showed a significantly lower mean score than the controls on the SV-BNT with a large effect size, which demonstrates adequate construct validity. However, sensitivity was 39% and specificity was 89%. The very low sensitivity shows that the concurrent validity of the test is poor and many patients with AD can score within the normal range. The SV-BNT does not guarantee a proper identification of naming difficulties, and therefore, its clinical use should not be recommended.

Key words: *Boston Naming Test, cross-cultural, Spanish version, validity*

INTRODUCTION

Naming is the ability to produce the names of objects/subjects at will. It is frequently impaired in aphasias and some neurodegenerative diseases. Impaired naming ability is the hallmark of semantic dementia but is also a prominent feature of Alzheimer's disease (AD). Although this decline is not always present in the early stages of the disease, it is very frequent, and therefore, it helps in the discrimination of AD from normal aging.

Confrontation naming tests are the most frequent method used to assess naming ability. These tests involve showing the patient a series of line drawings or photographs of objects, and the patient is asked to say the specific name of the object. The Boston Naming Test (BNT) is, by far, the most widely used test in the world. The original BNT was developed in Boston, MA, and consists of 60 black-and-white line drawings that are hierarchically ordered. There are now two editions of the test, with the updated BNT-Second Edition being the most recent version (Kaplan, Goodglass, & Weintraub, 1983, 2001). It uses the same protocol as the original test.

Perhaps because it was one of the few and earliest tests designed to assess naming ability, the BNT has been adapted to many languages, including: Spanish (Allegri et al., 1997; Pontón et al., 1992; Quiñones-Ubeda, Peña-Casanova, Böhm, Gramunt-Fombuena, & Comas, 2004), Chinese (Cheung, Cheung, & Chan, 2004), Dutch (Mariën,

Address correspondence to Alberto Luis Fernández, Ph.D., Universidad Católica de Córdoba, Cortex Foundation, Universidad Nacional de Córdoba, Chile 279, PB. CP 5000, Córdoba, Argentina. E-mail: neurorehab@onenet.com.ar

Mampaey, Vervaet, Saerens, & De Deyn, 1998), Jamaican (Unverzagt, Morgan, & Thesiger, 1999), Korean (Kim & Na, 1999), French Canadian (Roberts & Doucet, 2011), French Swiss (Thuillard-Colombo & Assal, 1992), Greek (Patricacou, Psallida, Pring, & Dipper, 2007), Italian (Riva, Nichelli, & Devoti, 2000), Malay (Van Dort, Vong, Razak, Kamal, & Meng, 2007), and Swedish (Tallberg, 2005). Because it was developed in an American English environment, it also had to be adapted to other English-speaking countries such as Australia (Cruice, Worrall, & Hickson, 2000) and New Zealand (Barker-Collo, 2007).

Adapting the test involved much more than translating it, because many of the test items proved to be inadequate for the target language/environment. For example, Roberts and Doucet (2011) found that items such as escalator, noose, latch, scroll, yoke, and palette were difficult to score, in part because of a lack of consensus on the specific object names among native French speakers. In addition, 13 other items had two possible names in standard French. In the Greek version, 4 items were replaced (pretzel, doorknocker, stethoscope, scroll) because they were conflicting items that led the participants to give wrong names (Patricacou et al., 2007). Even in the English-speaking countries, item replacements were recommended. Cruice et al. (2000) proposed substituting "beaver" and "pretzel" with "platypus" and "pizza" when the BNT was to be administered to Australian individuals. "Pretzel" is one of the most problematic items, because this object is uncommon in many cultures and there is no translation for it in some languages.

In 1996, a purportedly adapted version of the BNT to Spanish was published (García-Albea & Sánchez Bernardos, 1996). In this version, the following 13 items were replaced: whistle, saw, toothbrush, mushroom, wheelchair, pretzel, seahorse, wreath, escalator, door knocker, latch, scroll, and trellis. However, the authors provided no description of the criteria for excluding these items. The selection of these items seems arbitrary, because there are no previous item analysis data indicating the cultural inappropriateness of these items. Exclusion of the word "pretzel" appears appropriate, because there is no Spanish word for the item and the object is indeed rarely seen in Spanish-speaking countries. By contrast, the exclusion of relatively universal items such as "saw," "wheelchair," or "mushroom" is highly questionable. Moreover, there is no item analysis indicating the appropriateness of the newly included items. Finally, the authors did not specify that changes in the items had been made, and they provided no explanation as to how replacement items had been derived. Consequently, information about the changes that the authors made to the BNT was limited to comparing this adapted Spanish version to the original version of the test.

Allegri et al. (1997) developed norms for Argentina, a Spanish-speaking country, of the Spanish version of the BNT (SV-BNT). Based on their analysis of the correct response frequency for each item, the authors suggested a reordering of the original items. Thus, for example, "scroll," which was in Position 53 in the original order, was moved to the 48th place; "beaver," which was in Position 29, was transferred to the Position 58. As a result, a significant reorganization of the items was made. The authors also developed normative data stratified by education. No validity or reliability data were provided. The authors did not mention problems with language terms, translation, or appropriateness of different items. Nevertheless, the daily work with the test has demonstrated that it presents several difficulties. For example, "protractor" has two possible names: *transportador* and *semicírculo*. Although both names are correct, their usage frequency is very different. The "igloo" is usually mistaken as a mud oven due to its resemblance to the mud ovens built in some parts of South America. Moreover, the problems with the items are not only apparent in the translation from English to Spanish, but also within Spanish usage. For example, Garcia-Albea and Sánchez Bernardos (1996) replaced "pretzel" with *magdalena* (a bakery piece), but in the Argentinean normative study, "ice cream" was considered a correct answer for this item; the "stethoscope" is called *fonendoscopio* in Spain, but *estetoscopio* in Argentina. Therefore, in many cases, it is not clear what answers should be considered correct.

Other authors have provided normative data for the SV-BNT when administering it to Spanish (Quiñones-Ubeda, Peña-Casanova et al., 2004; Rami et al., 2008) and Colombian (Rosselli, Ardila, Florez, & Castro, 1990) samples but using the original ordering of the items.

Because the original items were replaced and significantly reorganized, the psychometric properties of the task might have been affected, especially the validity. It is not clear if these changes have altered the ability of the test to discriminate between participants with and without anomia. The International Test Commission (2005, p. 7) has established guidelines for the cross-cultural adaptation of tests. Among several guidelines, it is pertinent to highlight the following: "Test developers/publishers should provide information on the evaluation of validity in all target populations for whom the adapted versions are intended."

Because of the frequent reports of naming difficulties in patients with AD (Appell, Kertesz, & Fisman, 1982; Chertkow & Bub, 1990; Martin & Fedio, 1983), validity of the BNT has usually been studied by comparing the performance of patients with AD to that of controls. For example, B. W. Williams, Mack, and Henderson (1989) found that a cutoff score of 51 on the BNT correctly classified 80% of patients with AD and 86% of normal control participants. In addition, other studies have demonstrated a correlation between the severity of AD stage and BNT performance (Chenery, Murdoch, & Ingram, 1996; LaBarge, Balota, Storandt, & Smith, 1992). LaBarge et al. (1992) found that the mean scores of control participants were higher than the mean scores of those patients with very mild dementia who, in turn, obtained higher mean scores than those of patients with mild dementia. Chenery et al. (1996) also found that scores on the BNT decreased as severity of AD dementia increased. Moreover, Knesevich, LaBarge, and Edwards (1986) found that anomia, as identified by the BNT scores, predicted a faster progression of AD.

Changes in naming ability in AD are not only apparent in quantity but also in quality. LaBarge et al. (1992) showed that patients with mild dementia produced more linguistic errors (as compared with visuoperceptual) compared with the other groups. These patients also had a very low recognition rate of the items after the presentation of a multiple-choice test containing the right name of the item in one of the options. This was interpreted as a breakdown in semantic systems (i.e., the loss of semantic information), which results in impairment in the retrieval of proper names.

In addition, the inclusion of patients with AD in the construct validity studies of the BNT has been supported by data obtained in comparisons of test performance of patients with AD dementia and patients with other dementias. Lukatela, Malloy, Jenkins, and Cohen (1998) found that patients with AD made significantly more semantic errors than did patients with vascular dementia and controls, and their errors were primarily superordinate errors (i.e., they tended to name the category to which the item belonged), while the vascular dementia patients had a significant tendency to make more coordinate category member errors (i.e., they frequently produced the name of other items that belonged to the same category [e.g., *rat* instead of *beaver*]). Furthermore, V. G. Williams et al. (2007) showed that patients with AD made significantly more errors than did patients with Lewy-body dementia. Moreover, this study demonstrated that patients with AD made predominantly semantic errors while the other group made mainly visuoperceptual errors.

Taken altogether, these studies show that patients with AD are an appropriate group to use to obtain validity evidence for a confrontation naming test, because naming difficulties are quite specific to this group as compared with controls and other dementia groups.

To date, there are no studies published on the validity of the SV-BNT. Serrano et al. (2001) published data on the validity of a short version (12 pictures) of this test but not on the full version. They administered the full version of the SV-BNT to patients with AD and controls and then selected the most sensitive pictures based on their discriminating power. Their failure to account for the influence of age and education on this short version raises questions about the representativeness of their sample. In view of these shortcomings, this research was designed to evaluate the construct and concurrent validity of the SV-BNT regarding AD.

METHOD

Participants

The current sample was composed of 59 participants divided into two groups. The AD group included 23 participants who had a diagnosis of probable AD according to the National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association (NINCS–ADRDA) criteria (McKhann et al., 1984). Participants who were selected had previously undergone neuropsychological assessment, neurologic and psychiatric evaluation, and imaging (when available) as part of their clinical diagnostic evaluation at a memory clinic. Inclusion as a patient with AD in this study was decided upon by a professional group consensus considering history of current symptoms, medical records, neuropsychological performance, and present life circumstances of the patient. Additional procedures such as brain imaging and laboratory analysis were performed when necessary. Scores on the SV-BNT were not considered for the diagnosis of these patients. The control group was composed of 36 cognitive and neurologically healthy individuals. Potential participants presenting with any of the following conditions were excluded: stroke, loss of consciousness, traumatic head injury, central nervous system diseases, diabetes, chronic renal insufficiency, hepatic encephalopathy, untreated thyroid disease, chronic headache, epilepsy, untreated hypertension, severe cardiac failure, severe sleep disorders, coma, diagnosed psychiatric disease, or consumption of illicit drugs. In addition, participants' Mini Mental State Examination (MMSE) scores had to be greater than 23 to be included in this group. Most of the control participants were recruited at a social club for retired people in the city of Córdoba, Argentina. The rest were community-dwelling individuals who volunteered to participate. Participation of participants was voluntary, and written consent was obtained.

Groups were matched by age, $F(1, 57) = 3.63$, $p = .06$, and education, $F(1, 57) = 0.01$, $p = .94$. Table 1 shows the demographic data and MMSE scores for each group. The AD group was administered the Mattis Dementia Rating Scale (MDRS) as part of the neuropsychological battery. Their mean score was 107.57 ± 10.51, and the score range was 85 to 123. All but one of the patients obtained scores of less than 123 on the MDRS, which is the suggested cutoff for this scale (Fernández & Scheffel, 2003), and 65% of them obtained 110 or less points. These statistics characterize this sample as a group of patients with moderate-to-severe AD.

TABLE 1
Demographic Data of Alzheimer's Disease and Control Groups

	Control (n = 36)	*Alzheimer's Disease Group (n = 23)*
Age	68.81 ± 6.37	72.91 ± 10.22
Education	10.92 ± 4.03	11 ± 5.02
Gender	Male 17%	Male 48%
	Female 83%	Female 52%
MMSE	27.69 ± 1.56	19.61 ± 2.89

MMSE = Mini Mental State Examination.

TABLE 2
Items Included in the Spanish Version of the Boston Naming Test

1. cama	bed
2. árbol	tree
3. lápiz	pencil
4. **reloj**	**watch**
5. tijera	scissors
6. peine	comb
7. flor	flower
8. **martillo**	**hammer**
9. escoba	broom
10. **zanahoria**	**carrot**
11. percha	hanger
12. **corona**	**crown**
13. **regadera**	**watering can**
14. **termómetro**	**thermometer**
15. camello	camel
16. banco	bench
17. raqueta	racket
18. volcán	volcano
19. **escalera**	**ladder**
20. pirámide	pyramid
21. **chupete**	**pacifier**
22. **sacapuntas**	**pencil sharpener**
23. pulpo	octopus
24. caracol	snail
25. acordeón	accordion
26. helicóptero	helicopter
27. máscara	mask
28. **pez espada**	**sword fish**
29. arpa	harp
30. casa	house
31. canoa	canoe
32. embudo	funnel
33. zancos	stilts
34. compás	compass
35. cactus	cactus
36. pinzas	tweezers
37. hamaca	hammock
38. **aguja**	**needle**
39. bozal	muzzle
40. **helado**	**ice cream**
41. **cerradura**	**lock**
42. rinoceronte	rhinoceros
43. iglú	igloo
44. ábaco	abacus
45. paleta	palette
46. trípode	tripod
47. dominó	dominoes
48. pergamino	parchment
49. globo	globe
50. pelícano	pelican
51. dardo	dart
52. espárrago	asparagus
53. estetoscopio	stethoscope
54. esfinge	sphinx
55. armónica	harmonica
56. unicornio	unicorn
57. transportador	protractor
58. castor	beaver
59. bellota	acorn
60. yugo	yoke

Note. The newly included items in the Spanish version are in bold (García-Albea & Sánchez Bernardos, 1996). Items are ordered following Allegri et al. (1997).

Measures

All participants were administered the MMSE and the SV-BNT. Both tests and the screening questionnaire for health records were administered in a single session lasting approximately 30 min. The SV-BNT was administered using the standard instructions and procedures, but with the items reordered according to the suggestions of Allegri et al. (1997). Table 2 shows the items included in the SV-BNT.

Procedures

To assess the construct validity to the SV-BNT, an analysis of variance (ANOVA) was performed comparing both groups.

To evaluate the concurrent validity, patients with AD were classified as AD or non-AD according to their score on the SV-BNT (predictor variable). This classification was made taking into account the cutoff scores suggested by Allegri et al. (1997) for each educational group: 38 for elementary school (less than 8 years of education), 44 for high school (8 to 12 years of education), and 48 for college (more than 12 years of education). These cutoff scores represent 2 standard deviations below the mean of each group. Using the DAG [Diagnostic and AGreement Statistics]-Stat spreadsheet (Mackinnon, 2000), several indexes were calculated. Table 3 shows every index and the formulas used to compute them.

RESULTS

The ANOVA showed a significant difference between the mean scores on the SV-BNT of each group, $F(1, 57) = 23.58$, $p = .000$. The eta squared was .41, meaning that the effect size is large (Cohen, 1988). Figure 1 shows the mean scores of both groups.

Table 4 exhibits the results of the concurrent validity analysis. The resulting indexes are as follows: sensitivity, 39%; specificity, 89%, efficiency, 70%; predictive value of a positive test, 69%; predictive value of a negative test, 70%; likelihood ratio of a positive test, 3.52; and likelihood ratio of a negative test, 0.69.

DISCUSSION

Data regarding the validity of the SV-BNT yielded mixed results. The construct validity data produced satisfactory results because the difference between both groups is significant and the effect size is large. Therefore, the test indeed discriminates between a group of patients that includes individuals with varying degrees of naming difficulties and a group of individuals without naming difficulties. This finding was not unexpected given the prominent general cognitive impairment of the patients included in the AD group, as

TABLE 3
Indexes and Formulas Used to Calculate Diagnostic Accuracy Statistics

Indexes	*Formulas*
Sensitivity: proportion of true positives classified as positives by the test	True positives/true positives + false negatives
Specificity: proportion of true negatives classified as negatives by the test	True negatives/true negatives + false positives
Efficiency: proportion of positives and negatives classified correctly by the test	True positives + true negatives/total
Predictive value of a positive test: probability that an observation with a positive test will be positive on the criterion	True positives/true positives + false positives
Predictive value of a negative test: probability that an observation with a negative test will be negative on the criterion	True negatives/true negatives + false negatives
Likelihood ratio of a positive test: how many times the test is more likely to obtain an abnormal score among people with disease as compared to people without disease	Sensitivity/(1 – specificity)
Likelihood ratio of a negative test: how many times the test is more likely to obtain a normal score among people without as compared to people with disease	*Likelihood ratio of a negative test* = (1 – sensitivity)/specificity

shown by their performance on the MMSE and MDRS. On the other hand, the data indicated rather weak concurrent validity. According to the criteria established by Cicchetti (2001), the indexes of the probability data in this case can be valued as follows: sensitivity = poor, specificity = good, efficiency = fair, predictive value of a positive test = poor, and predictive value of a negative test = fair. Due to its very low sensitivity, the SV-BNT produces many false negatives (i.e., many patients with AD obtain results within normal limits). The SV-BNT can acceptably identify those patients without AD but underidentifies patients with AD. Accordingly, the likelihood ratio of a negative test is very low, indicating that a patient without dementia is barely more likely to obtain a normal score than is a patient with dementia. Hence, a normal score on the SV-BNT does not rule out the possibility of AD because many patients with AD obtain normal scores. Although the efficiency is qualified as fair, the percentage of incorrectly diagnosed patients is high at 30%. This misidentification percentage is quite high for a diagnostic test because 30 individuals out of 100 could receive a wrong diagnosis. In the clinical setting, the consequences of such a rate of incorrect diagnosis could result in patients being incorrectly diagnosed and potentially provided with treatment less specific to their conditions. Cicchetti and Rourke (2004) have concluded that false-negative errors are much more important than false-positive errors. They even consider that "… a False Negative error is three times as serious as a False Positive error …" (pp. 104). This judgment is based on the assumption that a false-positive case will be detected by confirmatory diagnosis but a false-negative case will remain with an undetected disorder.

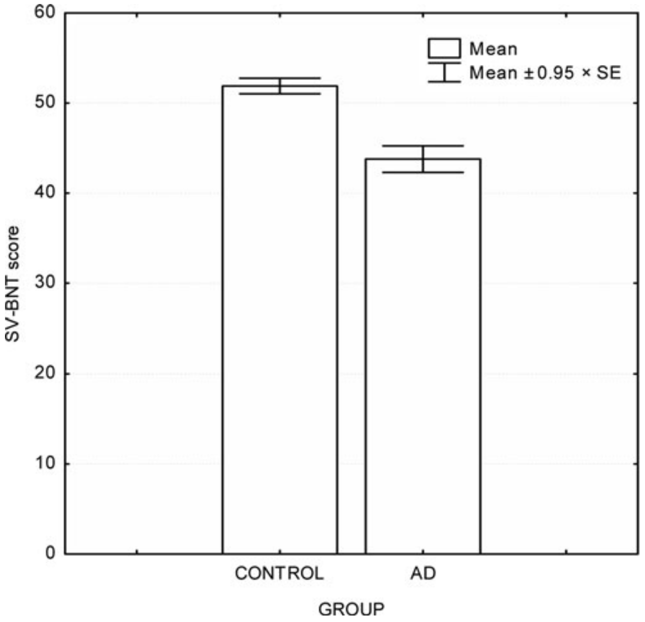

FIGURE 1 Mean scores on the Spanish-language version of the Boston Naming Test (SV-BNT) for each group.

There are several possible explanations for the poor concurrent validity observed in this study. First, the adaptation of the BNT to the Spanish language should be discussed. As mentioned before, García-Albea and Sánchez Bernardos (1996) replaced some of the original items seemingly without any empirical data supporting their exclusion/inclusion of items. As a consequence, the test might have turned into an easier task, and a ceiling effect might be the consequence of those changes. Because of this, it is possible that patients with AD might earn high scores on the test.

Secondly, as mentioned earlier, some of the pictures have more than one correct name. The usage frequency of each acceptable word is dissimilar, resulting in the same item

TABLE 4
Classification of Cases According to the Scores of Each Patient on the SV-BNT

	Test		
Criterion	*AD*	*Control*	*Total*
AD	9	14	23
Control	4	32	36
Total	13	46	59

SV-BNT = Spanish-language version of the Boston Naming Test; AD = Alzheimer's disease group.

being classified as easy or hard based on which term for it is used. It seems that this would have the potential of inflating or deflating test scores depending on which of the two terms the patient produces. For example, in this study, two correct words for "abacus" in Spanish were registered: *ábaco* and *contador*, but *contador* was much more frequent than *ábaco*. As a consequence, this could be considered an easy item if *contador* is considered the right word for it, but it is a difficult item if *ábaco* is expected to be retrieved. Indeed, Busca-Palabras (B-Pal), a software program that contains a large database of the frequency of use of Spanish words, establishes that *contador* has a frequency of 5 times per million words while *ábaco* appears 0.18 times per million words (Davis & Perea, 2005).

A similar problem arises with nonunivocal items (i.e., those items that do not elicit a uniform answer). Such is the case of "igloo." As described before, this picture is usually interpreted as a mud oven in this region. Although this is not a correct answer on the English-language BNT, it is not incorrect either, because the problem is not the difficulty to retrieve the name of the item but a misidentification produced by a culturally inappropriate item. As a result, the correct answer for such an item is difficult to determine in populations of different cultures and languages. Likewise, if both answers are judged as correct, their frequency of appearance per million words is very different according to B-Pal (as discussed earlier): 1.07 for *iglú* (the Spanish adaptation of "igloo") and 7.14 for *horno* (oven). Therefore, the difficulty level of the items is entirely dependent on how the object is perceived. *Helado*/*Magdalena* is another conflicting item due to a lack of concordance in the perception of the picture; some individuals interpret the item as a *magdalena* (bakery piece) while others view it as an *helado* (ice cream).

Third, some intrinsic psychometric properties of the test may also contribute to the low concurrent validity. Although no studies on the concurrent validity have been published with the original version of the BNT, there are some data indicating that its discriminatory power is low. In fact, research has shown that the original version of the BNT has a ceiling effect and that most of the scores are skewed toward the high end of the range (Hawkins & Bender, 2002). This phenomenon was also observed with the SV-BNT (Allegri et al., 1997). Because most of the normal controls obtain high scores, it seems that one must demonstrate considerable impairment in naming ability to score below the normal range. Consequently, individuals with an incipient naming deficit may not be detected, because they score within the normal range on the BNT.

Finally, the inconsistent results between construct and concurrent validity are not unusual among neuropsychological tests. As Ivnik et al. (2000) have shown, many neuropsychological tests show good evidence of construct validity when mean scores of control and patient groups are compared; however, that evidence is not always paired with concurrent validity (i.e., the test's accuracy at identifying participants with the target condition). Indeed, in their research, most of the diagnostic accuracy indexes of their factor scores (composed of a collection of scores on several different neuropsychological tests) did not reach 70%, which is the lowest acceptable limit when evaluating the validity of a given test (Cicchetti, 2001). Although this feature of neuropsychological tests might be interpreted as a psychometric weakness, it is important to remember that most neuropsychological tests are developed to evaluate a given ability. Although impairment in a particular cognitive skill might be frequently observed in a particular disease, impairment in that particular ability may not be pathognomonic of that disease. In short, impairment on a single neuropsychological test is never diagnostic of a certain disease process. For example, although naming impairment is frequently observed in patients with AD, not all patients with AD suffer from anomia. Hence, a naming test will probably be effective at identifying patients with AD who have significantly impaired naming abilities. Consequently, the value of this test (and many neuropsychological measures) lies in its ability to measure expressive language functioning rather than its standalone ability to identify specific patients with AD. Furthermore, it is important to emphasize that no competent neuropsychologist or other neurodiagnostician would diagnose AD on the basis of the sole neuropsychological measure. Despite these considerations, some other confrontation naming tests have shown a better balance between specificity and sensitivity (Fernández, 2013).

One other issue should be addressed regarding the clear difference between the construct and concurrent validity found in this research. Most of the patients in this study had moderate-to-severe dementia, as demonstrated by their scores on the MMSE and MDRS. Consequently, it is quite bewildering that many of them perform within normal limits on the SV-BNT. This effect might be explained by a combination of two factors: (a) the very low cutoff scores obtained by Allegri et al. (1997; 38, 44, and 48 out of 60, for elementary, high school, and college, respectively), and (b) the ceiling effect produced by this task.

The other normative studies published with Spanish-speaking populations using the original item ordering of the BNT need to be addressed. In the first place, to our knowledge, there is no research published on the validity of this version in Spanish-speaking populations; hence, its validity remains uncertain. In the second place, Allegri et al. (1997) have demonstrated that the original ordering of the items is not appropriate for Spanish speakers; consequently, using this version with the Spanish-speaking population might potentially lead to incorrect measurements of the trait under observation (naming), which, in turn, will lead to wrong conclusions about the patient's cognitive performance. Regarding this issue, the International Test Commission (2005) guidelines mentioned earlier have been established on the basis of a very sound bulk of data showing the risks for cross-cultural administration of tests (Ardila, 2007; Hambleton, 2005; Nell, 2000; van de Vijver & Tanzer, 2004). Recently, Fernández-Blázquez et al. (2012) have published a short version of the BNT (15 items) obtained through the application of the item response theory methodology after the administration of the original item ordering. Although the selection of the final items results from a fairly elaborate process, the validity evidence is rather weak. The authors reported that the anomic participants obtained a lower percentage of correct answers on every item; however, the classification of the participants into anomic or nonanomic is not based on an independent criterion. Participants were classified into each group according to their performance on the test, therefore contaminating the criterion. Moreover, no mean differences test between groups was performed, nor were indexes of sensitivity or specificity provided.

Ethical Implications

The use of psychological tests is regulated by ethical standards to ensure its proper use. Professional associations regularly release their ethics code with recommendations for an adequate use of psychological tests (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999). In addition, as mentioned earlier, the International Test Commission (2005) has established guidelines for the cross-cultural adaptation of tests to allow for proper use of tests. The "Standards for Educational and Psychological Testing" (AERA, APA, & NCME, 1999) is one of the most widely accepted set of ethical regulations in the use of psychological tests. Following this publication, the results of the present study have serious ethical implications for the appropriate use of the SV-BNT. These results involve a series of conflicts with many of the standards involving validity (1.5, 1.6, 1.16), fairness in testing (7.3, 7.10), testing individuals of diverse linguistic backgrounds (9.2, 9.4, 9.6, 9.7, 9.9), and the responsibilities of test users (11.1, 11.4; AERA, APA, & NCME, 1999). Describing each one of these standards and the conflicts with the results presented here is beyond the scope of this article, but they can be summarized in the following items: (a) Appropriate data on the validity evidence of the SV-BNT were not provided by publishers/researchers who adapted the test. This fact has obvious ethical implications because the data collected with this test are used to make decisions regarding the patients, and as mentioned earlier, this may lead to misdiagnosis. (b) Many neuropsychologists have included the SV-BNT in their batteries for the assessment of Spanish-speaking patients despite the lack of validity evidence. This fact was probably based on the unawareness of the impact of culture and regional/idiomatic variations within Spanish or on a-priori assumption that culture may not have a significant impact in a confrontation naming test. However, there is a large body of data showing that even tests measuring universal constructs such as attention are not free of cross-cultural bias (Ardila, 2007; Ardila & Moreno, 2001; Fernández & Marcopulos, 2008; Nell, 2000).

In summary, the SV-BNT showed good construct validity but poor concurrent validity, especially a very low sensitivity. The most likely reasons for the poor concurrent validity are manifold: flawed adaptation process, confounds caused by the inclusion of culturally inappropriate items, a ceiling effect, and a psychometric characteristic of many neuropsychological tests. Given these drawbacks, the SV-BNT does not guarantee a proper identification of naming difficulties, and therefore, its clinical use should not be recommended.

REFERENCES

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

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

Appell, J., Kertesz, A., & Fisman, M. (1982). A study of language functioning in Alzheimer patients. *Brain & Language*, *17*, 73–91.

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

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

Barker-Collo, S. (2007). Boston Naming Test performance of older New Zealand adults. *Aphasiology*, *21*, 1171–1180. doi:10.1080/02687030600821600

Chenery, H. J., Murdoch, B. E., & Ingram, J. C. L. (1996). An investigation of confrontation naming performance in Alzheimer's dementia as a function of disease severity. *Aphasiology*, *10*, 423–441.

Chertkow, H., & Bub, D. (1990). Semantic memory loss in dementia of Alzheimer's type. What do various measures measure? *Brain*, *113*, 397–417. doi:10.1093/brain/113.2.397

Cheung, R. W., Cheung, M. C., & Chan, A. S. (2004). Confrontation naming in Chinese patients with left, right, or bilateral damage. *Journal of the International Neuropsychological Society*, *10*, 46–53. doi:10.10170S1355617704101069

Cicchetti, D. V. (2001). The precision of reliability and validity estimates re-visited: Distinguishing between clinical and statistical significance of sample size requirements. *Journal of Clinical and Experimental Neuropsychology*, *23*, 695–700. doi:10.1076/jcen.23.5.695.1249

Cicchetti, D. V., & Rourke, B. P. (Eds.). (2004). *Methodological and biostatistical foundations of clinical neuropsychology and medical and health disciplines* (2nd ed.). New York, NY: Psychology Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Cruice, M. N., Worrall, L. E., & Hickson, L. M. H. (2000). Boston Naming Test results for healthy older Australians: A longitudinal and cross-sectional study. *Aphasiology*, *14*, 143–155. doi:10.1080/026870300401522

Davis, C. J., & Perea, M. (2005). BuscaPalabras: A program for deriving orthographic and phonological neighborhood statistics and other psycholinguistic indices in Spanish. *Behavior Research Methods*, *37*, 665–671.

Fernández, A. L. (2013). Development of a confrontational naming test for Spanish-speakers: The Cordoba Naming Test. *The Clinical Neuropsychologist*, *27*, 1179–1198. doi:10.1080/13854046.2013.822931

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

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

Fernández-Blázquez, M. A., Ruiz-Sánchez de León, J. M., López-Pina, J. A., Llanero-Luque, M., Montenegro-Peña, M., & Montejo-Carrasco, P. (2012). Nueva versión reducida del test de denominación de Boston para mayores de 65 años: aproximación desde la teoría de respuesta al ítem [A new shortened version of the Boston Naming Test for those aged over 65: An approach from item response theory]. *Revista de Neurología*, *55*, 399–407.

García-Albea, J. E., & Sánchez Bernardos, M. L. (1996). Test de Boston para el diagnóstico de la afasia: Adaptación española [Boston Naming Test for aphasia diagnosis: Spanish version]. In H. Goodglass & E. Kaplan (Eds.), *La evaluación de la afasia y de trastornos relacionados* [Assessment of aphasia and related disorders] (2nd ed., pp. 129–198). Madrid, Spain: Editorial Médica Panamericana.

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

Hawkins, K. A., & Bender, S. (2002). Norms and the relationship of Boston Naming Test performance to vocabulary and education: A review. *Aphasiology*, *16*, 1143–1153.

International Test Commission. (2005). *International Test Commission guidelines for translating and adapting tests*. Retrieved from http://www.intestcom.org

Ivnik, R. J., Smith, G. E., Petersen, R. C., Boeve, B. F., Kokmen, E., & Tangalos, E. G. (2000). Diagnostic accuracy of four approaches to interpreting neuropsychological test data. *Neuropsychology*, *14*, 163–177. doi:10.1037/0894-4105.14.2.163

Kaplan, E. F., Goodglass, H., & Weintraub, S. (1978, 1983). *The Boston Naming Test: Experimental edition* (*1978*). Boston, MA: Kaplan & Goodglass. (2nd ed., Philadelphia, PA: Lea & Febiger)

Kaplan, E. F., Goodglass, H., & Weintraub, S. (2001). *The Boston Naming Test* (2nd ed.). Philadelphia, PA: Lippincott Williams & Wilkins.

Kim, H. L., & Na, D. L. (1999). Normative data on the Korean version of the Boston Naming Test. *Journal of Clinical and Experimental Neuropsychology*, *21*, 127–133.

Knesevich, J. W., LaBarge, E., & Edwards, D. (1986). Predictive value of the Boston Naming Test in mild senile dementia of the Alzheimer type. *Psychiatry Research*, *19*, 155–161. doi:10.1016/0165-1781(86)90008-9

LaBarge, E., Balota, D. A., Storandt, M., & Smith, D. S. (1992). An analysis of confrontation naming errors in senile dementia of the Alzheimer type. *Neuropsychology*, *6*, 77–95. doi:10.1037/0894-4105.6.1.77

Lukatela, K., Malloy, P., Jenkins, M., & Cohen, R. (1998). The naming deficit in early Alzheimer's and vascular dementia. *Neuropsychology*, *12*, 565–572.

Mackinnon, A. (2000). A spreadsheet for the calculation of comprehensive statistics for the assessment of diagnostic tests and inter-rater agreement. *Computers in Biology and Medicine*, *30*, 127–134. doi:10.1016/S0010-4825(00)00006-8

Mariën, 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. doi:10.1006/brln.1998.2000

Martin, A., & Fedio, P. (1983). Word production and comprehension in Alzheimer's disease: The breakdown of semantic knowledge. *Brain and Language*, *19*, 124–141.

McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., & Stadlan, E. (1984). Clinical diagnosis of Alzheimer's disease: Report of the NINCS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's disease. *Neurology*, *34*, 939–944.

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

Patricacou, A., Psallida, E., Pring, T., & Dipper, L. (2007). The Boston Naming Test in Greek: Normative data and the effects of age and education on naming. *Aphasiology*, *21*, 1157–1170. doi:10.1080/02687030600670643

Pontón, M. O., Satz, P., Herrera, L., Young, R., Ortiz, F., D'Elia, L., … Namerow, N. A. (1992). Modified Spanish Version of the Boston Naming Test. *The Clinical Neuropsychologist*, *6*, 334.

Quiñones-Ubeda, S., Peña-Casanova, J., Böhm, P., Gramunt-Fombuena, N., & Comas, L. (2004). Estudio normativo piloto de la segunda edición del Boston Naming Test en una muestra española de adultos jóvenes (20–49 años) [Preliminary normative data for the second edition of the Boston Naming Test for young Spanish adults]. *Neurología*, *19*, 248–253.

Rami, L., Serradell, M., Bosch, B., Caprile, C., Sekler, A., Villar, A., … Molinuevo, J. L. (2004) Normative data for the Boston Naming Test and the Pyramids and Palm Trees Test in the elderly Spanish population. *Journal of Clinical and Experimental Neuropsychology*, *30*(1), 1–6.

Riva, D., Nichelli, F., & Devoti, M. (2000). Developmental aspects of verbal fluency and confrontation naming in children. *Brain and Language*, *71*, 267–284. doi:10.1006/brln.1999.2166

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

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.

Serrano, C., Allegri, R. F., Drake, M., Butman, J., Harris, P., Nagle, C., & Ranalli, C. (2001). Versión corta en español del test de denominación de Boston: su utilidad en el diagnóstico diferencial de la enfermedad de Alzheimer [A shortened form of the Spanish Boston Naming Test: A useful tool for the diagnosis of Alzheimer's disease]. *Revista de Neurologí*a, *33*, 624–627.

Tallberg, I. M. (2005). The Boston Naming Test in Swedish: Normative data. *Brain and Language*, *94*, 19–31. doi:10.1016/j.bandl.2004.11.004

Thuillard-Colombo, F., & Assal, G. (1992). Adaptation française du test de dénomination de Boston versions abrégées [French adaptation of the Boston naming test-short version]. *Revue Européenne de Psychology Appliquée*, *42*, 67–71.

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

van de Vijver, F., & Tanzer, N. (2004). Bias and equivalence in cross-cultural assessment: An overview. *Revue européenne de psychologie appliqu*ée, *54*, 119–135.

Van Dort, S., Vong, E., Razak, R. A., Kamal, R. M., & Meng, H. P. (2007). Normative data on a Malay version of the Boston Naming Test. *Journal Sains Kesihatan Malaysia*, *5*, 2736.

Williams, B. W., Mack, W., & Henderson, V. W. (1989). Boston Naming Test in Alzheimer's disease. *Neuropsychologia*, *27*, 1073–1079. doi:10.1016/0028-3932(89)90186-3

Williams, V. G., Bruce, J. M., Westervelt, H. J., Davis, J. D., Grace, J., Malloy, P. F., & Tremont, G. (2007). Boston Naming performance distinguishes between Lewy body and Alzheimer's dementias. *Archives of Clinical Neuropsychology*, *22*, 925–931.