Published in final edited form as:

J Clin Exp Neuropsychol. 2013 October ; 35(8): 835–845. doi:10.1080/13803395.2013.825234.

A Brief Spanish-English Equivalent Version of the Boston Naming Test: A Project FRONTIER Study

Danielle R. Jahn, M.A.,

Department of Psychology, Texas Tech University

Cortney B. Mauer, M.A.,

Department of Psychology, Texas Tech University

Chloe V. Menon, M.A.,

Department of Psychology, Texas Tech University

Melissa L. Edwards, M.A.,

Department of Psychology, University of North Texas

Jeffrey A. Dressel, Ph.D., and

Human Solutions, Inc.

Sid E. O'Bryant, Ph.D.

Department of Internal Medicine, University of North Texas Health Science Center.

Abstract

The Boston Naming Test is a neuropsychological measure of confrontation naming, short forms of which can be advantageous with various populations. The purpose of this study was to establish a Spanish-English equivalent version of the BNT using item response theory. Data were analyzed from 380 Project FRONTIER participants; 27 items differed between groups and were removed from the measure. Additionally, 18 items did not differ between groups but were poor items. The current 15-item Spanish-English equivalent version of the BNT offers significant advantages. Future work is required to validate the diagnostic utility of the instrument in various settings and populations.

Keywords

Confrontation naming; Item response theory; Language of administration; Boston Naming Test; Hispanic

The Boston Naming Test (BNT; Kaplan, Goodglass, & Weintraub, 1983) is the single most commonly utilized neuropsychological measure of confrontation naming, with various forms included in numerous large-scale clinical studies/batteries of cognitive aging, including the National Alzheimer's Coordinating Center (NACC) Uniform Dataset (UDS), Consortium to Establish A Registry for Alzheimer's Disease (CERAD), and the Alzheimer's Disease Neuroimaging Initiative (ADNI). The BNT has been shown to be clinically useful in a broad range of neurological illnesses including stroke, traumatic brain injury (TBI), cerebrovascular accident (CVA), anoxia, subcortical disease (e.g. Parkinson's disease, multiple sclerosis), Alzheimer's disease, dementia with Lewy bodies (DLB), as well as psychiatric conditions (e.g. schizophrenia; Strauss, Sherman, & Spreen 2006). The BNT has

been translated into many languages including Chinese, Italian, Dutch, Korean, Jamaican, French-Canadian, and Spanish (Strauss et al., 2006).

Appropriate psychological and neuropsychological assessment across those of diverse ethnic and linguistic backgrounds is critical to ensure that neuropsychologists and other clinicians can accurately assess cognitive functioning and interpret results, and provide access to neuropsychological services for a variety of populations (Puente & Puente, 2009). However, simple translation into another language may not adequately reflect the same level of cognitive functioning and therefore may bias results (Puente & Puente, 2009). Conceptual equivalence (e.g., items having the same meaning), functional equivalence (e.g., scores representing the same level of functioning), linguistic equivalence (e.g., words being similar), and condition equivalence (e.g., testing having the same importance and meaning) are all important considerations in testing across languages and cultures, among other things (Puente & Puente, 2009). Considering meaning in various cultures and whether the measure (and items on the measure) assess equivalent cognitive functioning are especially critical (Puente & Ardila, 2000). New norms and standardization are also needed after translation to ensure that interpretation is appropriate in different languages (Puente & Puente, 2009), as scores that are not compared to appropriate norms may misrepresent an individual's true level of cognitive functioning (Puente & Ardila, 2000). Previous authors have highlighted the importance of item analysis when examining measure equivalence across languages (Puente & Puente, 2009).

To this end, Kohnert et al. (1998) have previously sought to obtain accurate interpretation of naming performance in a bilingual sample of English- and Spanish-speakers. The purpose of this examination was to establish preliminary normative data through administration of the BNT for bilingual U.S. educated adults with similar demographic characteristics. The findings of these normative data indicated that performance on a picture naming task was significantly better for the sample in English relative to Spanish. Specifically, the Spanish item analysis demonstrated more variable responses to the items throughout the BNT, indicating that the presentation order of the items was not a valid indicator of correct picture naming. In contrast to examining bilingual test performers on the BNT, Ardila (2007) developed a cross-linguistic naming test consisting of different semantic categories of color pictures to represent basic words. He proposed that this naming test should be readily accessible in different languages and was better because it consists of replaceable photographs instead of ones that are "fixed." Ardila (2007) suggested that this cross-lingual naming test did not require norms and that the frequency of words could possibly be a criterion for difficulty level of an individual's vocabulary. Other measures, such as the Object and Action Naming Battery, have been examined in both Spanish and English utilizing a Spanish-English bilingual sample, and some items did not fit well across both languages (Edmonds & Donovan, 2012). Overall scores on another measure of confrontation naming were also significantly different when the test was administered in Spanish versus English, despite the fact that participants were Spanish-English bilingual (Edmonds & Kiran, 2004). Although these researchers attempted to develop methods to examine confrontation naming items for bilinguals, as well as cross-linguistically, we are unaware of any studies that have utilized item response theory (IRT) analyses to identify a common set of items with equivalent psychometric functioning that can be utilized to assess monolingual individuals across languages.

In addition to these efforts, several abbreviated versions of the BNT have been developed for efficiency when time constraints occur in clinical practice and research. A shortened version of this measure is advantageous for use with patients with less education, lower intellectual functioning, or severe cognitive dysfunction, as they may be more readily bothered or fatigued by a lengthier version (Calero, Arnedo, Navarro, Ruiz-Pedrosa, &

Carnero, 2002). Although advances have been made in validating shortened forms, limited information exists on the specific psychometric properties of such measures, particularly in diverse subgroups (Graves, Bezeau, Fogarty, & Blair, 2004). Recent work also suggests that abbreviated versions of the BNT can be utilized to estimate full administration scores (Hobson et al., 2011).

Item Response Theory (IRT) is the modern approach for evaluating the psychometric properties of assessment instruments, as well as generating new instruments. IRT can also be utilized as a method to generate abbreviated tests through the identification of items that yield optimal as well as sub-optimal psychometric properties (Graves et al., 2004). IRT can be used to estimate item calibrations (e.g., naming difficulty) and to develop a hierarchical item map that organizes items based on level of difficulty assessing three parameters: discrimination, difficulty, and guessing. Abbreviated versions of a test can then be developed based on these item parameters (Edelen & Reeve, 2007; Velozo, Lai, Mallinson, & Hauselman, 2001). Differential item functioning (DIF) analysis in IRT can then be utilized to examine the equivalence of the parameters across groups (e.g., English- versus Spanish-speakers).

IRT analyses have been applied to the BNT in various populations. Graves and colleagues (2004) developed two short forms (e.g., 30-item and 15-item) of the BNT using IRT with older adult outpatients. DIF in the BNT was examined in African-American and Caucasian older adults with no significant cognitive impairments, and 12 items demonstrated DIF in this sample, indicating possible bias in scores due to these items (Pedraza, Graff-Radford, Smith, Ivnik, & Willis, 2009). Yang and colleagues (2011) undertook an examination of DIF in another naming test, the Spanish and English Neuropsychological Assessment Scales naming test, and found that at least five items demonstrated DIF between English- and Spanish-speakers. While these results suggest that items and measures vary across racial/ethnic groups, as well as language of administration, they also point to many items that function equivalently across these groups/administrations. Therefore, we hypothesize that IRT analyses can be utilized to generate an English-Spanish equivalent version of the BNT.

There are many advantages to having an equivalent version that can be administered across multiple languages. First, such an instrument would afford clinicians the ability to implement a common battery across patients. Second, such an instrument would enable direct comparisons of the exact same administration protocol of an instrument across languages, thereby affording more direct comparisons of cultural/linguistic groups in terms of neuropsychological deficits. Finally, such an instrument would enable the generation of normative data across languages for comparable populations. As previously noted, functional equivalence across items and the development of norms are important considerations in neuropsychological assessment across linguistic groups, suggesting that the utilization of IRT to examine equivalence is a critical first step in ensuring that the BNT is an appropriate confrontation naming measure for both English- and Spanish-speaking adults.

To our knowledge, no research to date has utilized IRT methods to generate a brief version of the BNT for monolingual English- and Spanish-speakers that is psychometrically equivalent. It is important that neuropsychological measurements contain items that are both conceptually and psychometrically equivalent across groups who differ in terms of education, ethnicity, and language due to a vast expansion of principally Spanish-speaking individuals in the United States (de la Plata, Arango-Lasprilla, Alegret, Moreno, & Tarraga, 2009). The goal of this study was to develop a valid form of the BNT that could be used with both monolingual English- and monolingual Spanish-speakers and would provide the best information about confrontation naming ability.

Method

Participants

Participants were 380 monolingual (either English-speaking only or Spanish-speaking only) adults ages 40 to 96 ($M = 64.12$, $SD = 13.01$). All participants lived in a rural area of West Texas. Over half of participants were women (70.3%), with 29.7% men. Average education was 11.49 years ($SD = 3.96$). Participants endorsed the following races: 90.8% White, 6.8% African-American, 3.4% American Indian/Alaskan Native, and 0.5% Other (note: some participants endorsed more than one race, leading to a total percentage greater than 100%). Approximately 24.8% of the sample identified as Hispanic; 92.6% of these participants reported that they were of Mexican-American origin, 1.1% reported that they were of Puerto Rican origin, and 6.4% reported that they were of other Hispanic origin. In terms of cognitive functioning and diagnoses, 253 participants had normal cognitive functioning, 81 were diagnosed with Mild Cognitive Impairment (MCI), 35 were diagnosed with other types of cognitive dysfunction without dementia (e.g., mental retardation, cognitive impairment due to psychiatric issues), 8 were diagnosed with possible or probable dementia of the Alzheimer's type (AD), and 3 were diagnosed with "other" dementia. We also examined cognitive functioning among non-Hispanic versus Hispanic participants to examine any potential differences between these groups. Among non-Hispanic participants, 53 were diagnosed with MCI (18.6%), 27 were diagnosed with other types of cognitive dysfunction without dementia (9.5%), 6 were diagnosed with possible or probable AD (2.2%), and 2 were diagnosed with "other" dementia (0.8%). Among Hispanic participants, 28 were diagnosed with MCI (29.8%), 8 were diagnosed with other types of cognitive dysfunction without dementia (8.5%), 2 were diagnosed with possible or probable AD (2.2%), and 1 was diagnosed with "other" dementia (1.1%). The percentage of participants with various cognitive diagnoses appears to be equivalent across non-Hispanic and Hispanic participants, though Hispanic participants may be slightly overrepresented in MCI diagnoses.

A total of 319 participants completed the BNT in English and 61 in Spanish. Language of testing was decided solely based on participants' preferences. Participants were excluded if they completed the BNT in both Spanish and English ($n = 44$), if they indicated that they were bilingual ($n = 125$), or if they did not complete this measure ($n = 172$). To assess for potential biases in our sample based on completion of the measure, we examined differences in gender, race/ethnicity, cognitive diagnoses, age, and education between participants who completed the BNT and those who did not. The only identified differences were in education, with those who completed the measure having higher mean education than those who did not, $t(503) = 5.339$, $p < .001$, and cognitive diagnoses, with more dementia and MCI cases not completing the measure, $\chi^2 = 52.364$, $p < .001$.

Cognitive diagnoses were roughly comparable between participants who completed the BNT in English and those who completed the test in Spanish, though those who completed the measure in Spanish had a somewhat higher percentage of MCI, relatively lower percentage of other non-dementia cognitive dysfunction, and slightly higher percentage of dementia. For those who completed the measure in English, 65 were diagnosed with MCI (20.4%), 32 were diagnosed with other types of cognitive dysfunction without dementia (10.0%), 6 were diagnosed with possible or probable AD (1.9%), and 3 were diagnosed with "other" dementia (0.9%). Among those who completed the measure in Spanish, 16 were diagnosed with MCI (26.2%), 3 were diagnosed with other types of cognitive dysfunction without dementia (4.9%), and 2 were diagnosed with possible or probable AD (3.3%).

Procedures

Participants were recruited through an ongoing epidemiological study of rural cognition, Project FRONTIER (Facing Rural Obstacles to health Now Through Intervention, Education & Research). Procedures for recruitment and the study protocol were approved by the Texas Tech University Health Sciences Center Institutional Review Board, and all participants provided written consent prior to participation. Project FRONTIER uses community-based participatory research (CBPR) strategy for recruitment and study management (e.g., partnering with local community resources for assessments, using interviewers and health professionals that live in rural areas, hosting events in the community to establish ties and foster collaboration, creating community advisory boards to approve all research). Our previous work has demonstrated the comparability of the recruited sample to that of the larger eligible population by county (O'Bryant et al., 2011).

Project FRONTIER examines aging in rural areas and, as part of the study, all participants underwent a standardized interview, medical examination (including blood draw for clinical labs), and neuropsychological testing by trained evaluators. Interviews included questions related to demographics, medical history (personal and family), medications, languages (read, spoken, written) and other topics; medical exams included a review of systems, neurological examination, and Hachinski Ischemic Index scale. Clinical labs included a lipid panel, thyroid levels, CBC, CMP and HbA1c; neuropsychological testing include the BNT, Repeatable Battery for the Assessment of Neuropsychological Status, Trail Making Test, Controlled Oral Word Association Task, Executive Interview, CLOX: An Executive Clock-Drawing Task, Animal Naming, Mini Mental Status Exam, and Clinical Dementia Rating scale. Interviews/assessments were conducted either at a hospital office or at the participant's home. Neuropsychological measures, including the BNT, were administered by trained psychological technicians (each of whom had at least a high school diploma) who were from the rural area. Technicians were fluent in both Spanish and English and, as they were from the same communities as participants, understood the cultural considerations within these communities. The entire process took approximately four hours and was completed within two weeks. Participants were paid \$40 for participation in the study. All data was presented at a weekly consensus committee consisting of physicians, psychologists (including neuropsychologists), and research staff to assign medical and cognitive diagnoses, according to published criteria (i.e., NINCDS-ADRDA possible/probable Alzheimer's disease, vascular dementia, "other" dementia [do not meet criteria for AD or VaD], mild cognitive impairment [Mayo Criteria], cognitive impairment no dementia [CIND; included life-long impairments such as mental retardation rather than recent cognitive loss], age-associated cognitive impairment [AACI, cognitively normal with complaints of impairment], or cognitively normal).

Measures

Boston Naming Test (BNT; Kaplan et al., 1983). The BNT is a 60-item measure of confrontation naming or confrontation word retrieval. A series of black and white pictures, which become progressively more difficult (e.g., show more rare or uncommon items), are presented individually to participants, and participants are asked to verbally name the item in each picture when it is presented. If participants clearly incorrectly perceive the stimulus picture (e.g., believe a mushroom is an umbrella), a stimulus cue is provided, which gives the participant a brief definition of the item. If the participant responds correctly after the stimulus clue, the item is scored as correct. Participants can also be given a phonemic cue if they provide an incorrect answer that is not simply a misperception of the picture. The phonemic cue is the first syllable of the correct response; however, if the participant responds correctly after the phonemic cue is given, the item is still scored as incorrect. Scoring was consistent across both Spanish and English speakers; no changes in

administration or scoring were made, other than language to ensure accurate comparison across languages. Administration began at item number one for all participants. Participants were tested in their preferred language and technicians ensured that participants were proficient in that language through conversing in that language prior to administration of neuropsychological testing.

Data Analysis

Item response theory likelihood ratio differential item functioning was employed to examine items that potentially evidenced DIF across Spanish- and English-speakers. DIF means that respondents in different groups who have equal ability would not perform equally on the item, indicating potential item bias. As stated previously, IRT analyses examine three primary constructs for each item included in the analysis: discrimination, difficulty, and guessing. An assumption of IRT is that each item is related to an underlying latent trait (in the case of the BNT, confrontation naming), and the analysis identifies how well each item discriminates between varying levels of that trait. High discrimination, reflected as higher discrimination scores, indicates that items differentiate well between varying levels of the underlying trait. However, low discrimination indicates that the item does not provide good information about the underlying trait, as incorrect versus correct responses do not differentiate between ability on the latent trait. Difficulty is a measure of the point at which each item discriminates along the continuum of the underlying trait. Ideally, the items identified as having good discrimination would have a range of difficulties. This would allow us to pinpoint where a respondent falls on the ability axis, as we would have items with good discrimination at various points along the ability trait. The guessing parameter is a threshold parameter indicating the likelihood of guessing correctly on the item.

Differential item functioning examines these parameters in two different groups of participants, and identifies any significant differences in those scores for the two groups. The reference group is generally the larger, majority group and the focal group is usually the smaller, minority group. These differences essentially mean that items function differently and cannot be considered equivalent in the groups. Therefore, the overall goal of a differential item functioning was to create a briefer version of the measure utilizing items that discriminate well across a range of difficulties and are equivalent across groups. Each item can exhibit DIF in one of two ways: uniform DIF indicates that one of the two groups is more likely to provide a correct response across all levels of the ability. In contrast, nonuniform DIF indicates that the chance of giving a correct response differs across subgroups at all levels of the ability spectrum (e.g., the probability of a correct response for a particular item is higher for one group at lower levels of the latent trait, but higher for the other group at higher levels of the latent trait).

The software package IRTLRDIF 2.0b (Thissen, 2001) was utilized for the analysis; English-speakers were the reference group and Spanish-speakers were the focal group. We examined differential item functioning across three parameters: a (the slope or discrimination parameter), b (the threshold or difficulty parameter), and g (the asymptote or “guessing” parameter). As each item is analyzed, all parameters are left unconstrained (with parameters for all other items constrained, and with the assumption of no DIF in other items). If the critical value for the null hypothesis (i.e., that all unconstrained parameters are equal across both groups) exceeds the critical value, then the null hypothesis is rejected and each parameter is examined individually to determine which differ(s) between the groups.

Results

The differential item functioning analyses identified 27 BNT items that differed between English- and Spanish-speakers, and were therefore removed. Differences were considered

significant if G^2 (the test of the hypothesis that all parameters were equal across groups) was greater 3.48 (Thissen, 2001). This value is the critical value for the χ^2 distribution at $\alpha = .05$, with one degree of freedom (Thissen, 2001). Of the items excluded for evidencing DIF, six items evidenced non-uniform DIF, meaning that the asymptote parameter (also called the discrimination parameter) significantly differed between the English-speaking and Spanish-speaking groups. Items excluded for non-uniform DIF were racquet (number 21), harp (number 38), accordion (number 47), noose (number 48), compass (number 50), and tongs (number 54). In addition, four items were excluded due to uniform DIF (i.e., the difficulty parameter significantly differing between groups). These items were wreath (number 28), acorn (number 32), muzzle (number 44), and unicorn (number 45).

In addition, two items, stilts (number 34) and hammock (number 39) evidenced differences in the discrimination and difficulty parameters. One item, harmonica (number 30) evidenced differences in the discrimination and threshold parameters. A total of 11 items evidenced DIF on the threshold (or guessing) parameter (i.e., bed [number 1], tree [number 2], camel [number 17], snail [number 22], volcano [number 23], globe [number 27], beaver [number 29], igloo [number 33], knocker [number 40], stethoscope [number 42], and asparagus [number 49]). Finally, three items evidenced differences between groups on the difficulty and threshold parameters. These items were dart (number 25), escalator (number 37), and pyramid (number 43).

In addition to the items excluded for evidencing DIF, there were items that did not differ between groups, but evidenced poor discrimination (i.e., discrimination scores at or below 1.00) in both groups, and were also removed because they did not provide good information about the underlying trait of confrontation naming. These items were pencil (number 3), house (number 4), scissors (number 6), comb (number 7), flower (number 8), saw (number 9), toothbrush (number 10), broom (number 12), hanger (number 15), bench (number 20), dominoes (number 35), latch (number 51), and abacus (number 60). Two items (i.e., mask [number 18] and funnel [number 46]) had poor discrimination among Spanish-speakers but adequate discrimination among English-speakers and were consequently kept in the final model. Finally, we examined the difficulty of the remaining items to ensure that the difficulty of these items did not closely overlapped with other items, particularly those that evidenced better discrimination. If the difficulty of items is very similar, they result in redundant information about latent traits. We determined that, across both groups, five items had difficulty scores that overlapped with difficulty scores of items that evidenced better discrimination. Therefore to reduce redundancy of items, these five items (i.e., seahorse [number 24], canoe [number 26], pelican [number 41], yoke [number 56], and protractor [number 59]) were removed from the measure. Additionally, difficulties of some items overlapped in either the reference or the focal group, but not both groups. These items were helicopter (number 11), octopus (number 13), mushroom (number 14), funnel (number 46), sphinx (number 55), and trellis (number 57); these items were retained in the measure because they provided non-redundant difficulty information in one of the groups. This resulted in 15 items included in the final scale: whistle (number 5), helicopter (number 11), octopus (number 13), mushroom (number 14), wheelchair (number 16), mask (number 18), pretzel (number 19), rhinoceros (number 31), cactus (number 36), funnel (number 46), tripod (number 52), scroll (number 53), sphinx (number 55), trellis (number 57), and palette (number 58). These items did not evidence DIF, discriminated well among varying levels of the latent construct, and spanned a broad range of difficulties without providing overlapping information in the two groups. See Table 1 for complete results.

Discussion

The present study generated the first Spanish-English, psychometrically equivalent, brief version of the BNT. IRT-based analyses identified 27 items with DIF, which indicates that the functioning of these items was not equivalent across the language groups. This means that participants with equal abilities in different groups would not perform equivalently on the items. A total of 15 items were retained in the final model, as these were functionally equivalent across groups while also demonstrating acceptable discrimination at varying levels of difficulty along the naming ability continuum. These 15 items represent a brief form of the BNT that is comparable across monolingual English- and Spanish-speakers. The results underscore the importance of examining the finer psychometric properties of neuropsychological tests in diverse populations. In particular, these results suggest that administering the full BNT to Spanish-speaking adults does not necessarily provide valid information about their confrontation naming ability. More specifically, some items may be more difficult in Spanish than in English and therefore may underestimate naming abilities if the norms for English speakers are used. Therefore, practitioners should consider administering the 15-item version to ensure equivalence among all patients. Among English-speaking patients, some BNT items do not provide good information about naming abilities and could be removed to reduce patient burden and the time required to complete the measure. Our findings also have implications in terms of future research. For example, researchers may validate the 15-item version we identified in other samples. Additionally, it may be beneficial to administer the 15-item version of the BNT if research designs include both English and Spanish speakers, to ensure that results are equivalent across both groups and do not skew results. Finally, the current study provides a foundation on which additional research can be conducted, such as the development of norms for the revised version of the BNT.

The presence of DIF on the BNT raises concern about the construct validity of this measure when applied to this population. A total of 27 items demonstrated significant differences in parameter estimates (i.e., discrimination, difficulty, and/or guessing), which implies that the probability of correctly responding to these items dramatically changed as a function of language group membership. This pattern of results is consistent with previous reports that translated neuropsychological measures may not appropriately assess the construct they were intended to measure. For example, Lowenstein, Rubert, Arugelles, and Duara (1995) evaluated patients diagnosed with Alzheimer's disease using several English- and Spanish-translated neuropsychological measures and reported that Spanish-translated neuropsychological measures were less strongly associated with functional ability than were the English assessments.

Lack of equivalence across testing languages is in part due to the problem of translating a word/concept from one language to another while attempting to conserve the original meaning (Geisinger, 1994). Although translation-back translation procedures are typically utilized to conserve a test's content validity (Brislin, 1980; Geisinger, 1994), construct validity may not always be preserved in that process (Gutierrez, 2002). However, as the current project demonstrates, there are items that can be utilized across languages that accurately assess the underlying construct of interest (i.e., naming ability).

A growing number of studies have focused on the development of neuropsychological test norms to facilitate the interpretation of test scores in linguistically and culturally diverse populations. In particular, a number of studies have sought to provide normative data for the BNT, including in its shortened versions (e.g., Fastenau, Denburg, & Mauer, 1998; Goodglass, Kaplan, & Barresi, 2001; Kent & Luszcz, 2002; Mitrushina, Boone, & D'Elia, 1999). It is possible that the current 15-item BNT will provide a way to re-analyze these

existing norms to generate meta-norms for the BNT that can be utilized for both English- and Spanish-speakers. Additional research is needed to examine the diagnostic accuracy of this 15-item BNT across both English- and Spanish-speaking individuals with and without cognitive impairment. The current research group is currently investigating this topic among Mexican-Americans and non-Hispanics diagnosed with MCI and AD.

While the current study contributes significantly to the literature, there are limitations. First, the sample size (particularly for the Spanish-speaking subgroup) was relatively small and the findings need to be cross-validated among a larger cohort, which the current research group is in the process of completing. Second, language group classification and ethnicity are confounded in the current cohort as nearly all Spanish-speakers were Hispanic, and we did not measure acculturation in the current study. Although this does not represent a threat to the integrity of our results since groups were matched on ability level, it does raise concern about the generalizability of our conclusions to other Spanish-speakers of differing sociocultural and linguistic backgrounds. Future work should also incorporate bilingual populations; we excluded bilingual participants to ensure a pure comparison between English-speakers and Spanish-speakers, and because our bilingual participants who spoke primarily Spanish evidenced differences in reading, writing, and speaking fluency from our bilingual participants who spoke primarily English. Future research should also seek to replicate these findings using corroborating computational procedures for DIF detection, as debate remains within the DIF literature about the extent to which item parameter estimates and DIF detection are dependent on the methods used to calculate those parameters. We used only a likelihood ratio test to assess for DIF.

Despite these limitations, the current study offers a significant advancement with regard to utilization of the BNT by generating a psychometrically equivalent Spanish-English brief version. This investigation represents an important first step in moving towards more socioculturally adapted and valid diagnostic decision-making. The results of this study suggest that BNT item functioning is strongly affected by language-group membership beyond basic differences in confrontation naming ability. The current research group has ongoing studies designed to (1) cross-validate the current 15-item BNT among both English and Spanish-speakers, (2) examine the diagnostic accuracy of the 15-item version among MCI and AD, and (3) examine the utility of the 15-version BNT in predicting cognitive decline/dysfunction (i.e., progression from MCI to AD) among English and Spanish-speakers. Eventually, research should also seek to extend the use of this modern psychometric methodology to other neuropsychological measures in both cognitively intact and impaired populations.

Acknowledgments

Research reported in this publication was supported by the National Institutes of Health Award Numbers R01AG039389 and L60MD001849. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This research was also funded in part by grants from the Hogg Foundation for Mental Health (JRG-040 & JRG-149), the Environmental Protection Agency (RD834794), and the National Academy of Neuropsychology. The authors would additionally like to thank the Project FRONTIER participants and research team.

References

- Ardila A. Toward the development of a cross-linguistic naming test. *Archives of Clinical Neuropsychology*. 2007; 22:297–307. doi:10.1016/j.acn.2007.01.016. [PubMed: 17303376]
- Brislin, RW. Translation and content analysis of oral and written materials.. In: Triandis, HC.; Berry, JW., editors. *Handbook of cross-cultural psychology-methodology*. Allyn & Bacon; Boston, ME: 1980. p. 389-444.

- Calero MD, Arnedo ML, Navarro E, Ruiz-Pedrosa M, Carnero C. Usefulness of a 15-item version of the Boston Naming Test in Neuropsychological assessment of low-educational elders with dementia. *The Journal of Gerontology: Psychological Sciences*. 2002; 57:187–191. doi: 10.1093/geronb/57.2.P187.
- de la Plata CM, Arango-Lasprilla JC, Alegret M, Moreno A, Tarraga L, Lara M, Hewlitt M, Hynan L, Cullum CM. Item analysis of three Spanish naming tests: A cross-cultural investigation. *NeuroRehabilitation*. 2009; 24:75–85. doi: 10.3233/NRE-2009-0456. [PubMed: 19208960]
- Edelen MO, Reeve BB. Applying item level response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research*. 2007; 16(Suppl 1):5–18. doi: 10.1007/s11136-007-9198-0. [PubMed: 17375372]
- Edmonds LA, Donovan NJ. Item-level psychometrics and predictors of performance for Spanish/English bilingual speakers on An Object and Action Naming Battery. *Journal of Speech, Language, and Hearing Research*. 2012; 55:359–381. doi:10.1044/1092-4388(2011/10-0307.
- Edmonds LA, Kiran S. Confrontation naming and semantic relatedness judgments in Spanish/English bilinguals. *Aphasiology*. 2004; 18:567–579. doi:10.1080/02687030444000057.
- Fastenau PS, Denburg NL, Mauer BA. Parallel short forms of the Boston Naming Test: Psychometric properties and norms for older adults. *Journal of Clinical and Experimental Neuropsychology*. 1998; 20:828–834. doi:10.1076/jcen.20.6.828.1105. [PubMed: 10484693]
- Geisinger KF. Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment*. 1994; 6:304–312. doi:10.1037/1040-3590.6.4.304.
- Goodglass, H.; Kaplan, E.; Barresi, B. *The assessment of aphasia and related disorders*. 3rd ed.. Lippincott Williams & Wilkins; Philadelphia, PA: 2001.
- Graves RE, Bezeau SC, Fogarty J, Blair R. Boston Naming Test short forms: A comparison of previous forms with new item response theory based forms. *Journal of Clinical and Experimental Neuropsychology*. 2004; 26:891–902. doi:10.1080/13803390490510716. [PubMed: 15742540]
- Gutierrez, G. The empirical development of a neuropsychological screening instrument for Mexican-Americans.. In: Ferraro, FR., editor. *Minority and cross-cultural aspects of neuropsychological assessment: Studies on neuropsychology, development, and cognition*. Swets & Zeitlinger; Lisse, Netherlands: 2002. p. 205-224.
- Hobson VL, Hall JR, Harvey M, Cullum CM, Lacritz L, Massman PJ, O'Bryant SE. An examination of the Boston Naming Test: Calculation of “estimated” 60-item score from 30- and 15-item scores in a cognitively impaired population. *International Journal of Geriatric Psychiatry*. 2011; 26:351–355. doi:10.1002/gps.2533. [PubMed: 20690129]
- Kaplan, E.; Goodglass, H.; Weintraub, S. *Boston Naming Test*. Lea & Feibiger; Philadelphia: 1983.
- Kent PS, Luszcz MA. A review of the Boston Naming Test and multiple-occasion normative data for older adults on 15-item versions. *The Clinical Neuropsychologist*. 2002; 16:555–574. doi:10.1076/clin.16.4.555.13916. [PubMed: 12822064]
- Kohnert KJ, Hernandez AE, Bates E. Bilingual performance on the Boston Naming Test: Preliminary norms in Spanish and English. *Brain and Language*. 1998; 65:422–440. doi:10.1006/brln.1998.2001. [PubMed: 9843612]
- Lowenstein DA, Rubert MP, Arguelles T, Duara R. Neuropsychological test performance and prediction of functional capacities among Spanish-speaking and English-speaking patients with dementia. *Archives of Clinical Neuropsychology*. 1995; 10:75–88. doi: 10.1016/0887-6177(93)E0005-V. [PubMed: 14589730]
- Mitrushina, MN.; Boone, KB.; D'Elia, LF. *Handbook of normative data for neuropsychological assessment*. Oxford University Press; New York, NY: 1999.
- O'Bryant SE, Edwards M, Menon CV, Gong G, Barber RC. Long-term low-level arsenic exposure is associated with poorer neuropsychological functioning: A Project FRONTIER study. *International Journal of Environmental Research and Public Health*. 2011; 8:861–874. doi:10.3390/ijerph8030861. [PubMed: 21556183]
- Pedraza O, Graff-Radford NR, Smith GE, Ivnik RJ, Willis FB, Petersen RC, Lucas JA. Differential item functioning of the Boston Naming Test in cognitively normal African American and

- Caucasian older adults. *Journal of the International Neuropsychological Society*. 2009; 15:758. doi:10.1017/S1355617709990361. [PubMed: 19570311]
- Puente, AE.; Ardila, A. Neuropsychological assessment of Hispanics.. In: Fletcher-Janzen, E.; Strickland, TL.; Reynolds, CR., editors. *Handbook of cross-cultural neuropsychology*. Plenum; New York: 2000. p. 87-104.
- Puente, AE.; Puente, AN. The challenge of measuring abilities and competencies in Hispanics/Latinos.. In: Grigorenko, EL., editor. *Multicultural psychoeducational assessment*. Springer; New York: 2009. p. 417-442.
- Strauss, E.; Sherman, EMS.; Spreen, O. *A compendium of neuropsychological tests: Administration, norms, and commentary*. 3rd ed.. Oxford University Press; New York, NY: 2006.
- Thissen, D. IRTLRDIF v.2.0b: Software for the computation of the statistics involved in item response theory likelihood-ratio tests for differential item functioning. 2001. Retrieved from <http://www.unc.edu/~dthissen/dl.html>
- Veloza CA, Lai JS, Mallinson T, Hauselman E. Maintaining instrument quality while reducing items: Application of rasch analysis to a self-report of visual function. *Journal of Outcome Measurement*. 2001; 4:667–680. Retrieved from <http://www.jampress.org/JOM.htm>. [PubMed: 11253902]
- Yang FM, Heslin KC, Mehta KM, Yang CW, Ocepek-Welikson K, Kleinman M, Morales LS, Hays RD, Stewart AL, Mungas D, Jones RN, Teresi JA. A comparison of item response theory-based methods for examining differential item functioning in object naming test by language of assessment among older Latinos. *Psychological Test and Assessment Modeling*. 2011; 53:440–460. Retrieved from <http://www.psychologie-aktuell.com/index.php?id=200>. [PubMed: 23471423]

Table 1
Differential Item Functioning for Boston Naming Test items

Item	Hypothesis	G^2	d.f.	Reference Group			Focal Group		
				a	b	g	a	B	g
01 - bed	Unconstrained	11.5 [*]	3	-0.38	3.38	0.49	0.27	-31.37	0.50
	1	0.0	1	-0.38	3.36	0.49	0.33	-23.65	0.49
	2	0.0	1	-0.38	3.37	0.49	-0.38	22.77	0.49
	3	11.5 [*]	1	0.34	-6.71	0.50	0.34	-6.71	0.50
02 - tree	Unconstrained	9.2 [*]	3	-37.11	1.33	0.47	-1.16	8.18	0.50
	1	0.1	1	-37.02	1.33	0.46	-1.89	8.08	0.46
	2	0.0	1	-30.92	1.35	0.46	-30.92	4.59	0.46
	3	9.2 [*]	1	-7.02	2.89	0.50	-7.02	2.89	0.50
03 - pencil	Unconstrained	0.0	3	0.18	-99.04	0.50	0.18	-99.04	0.50
04 - house	Unconstrained	1.6	3	0.23	-38.28	0.50	0.55	-4.01	0.50
05 - whistle	Unconstrained	1.3	3	2.21	-1.25	0.53	1.08	-2.16	0.49
06 - scissors	Unconstrained	0.0	3	0.18	-99.04	0.50	0.18	-99.04	0.50
07 - comb	Unconstrained	0.0	3	0.18	-99.04	0.50	0.18	-99.04	0.50
08 - flower	Unconstrained	0.0	3	0.18	-99.04	0.50	0.18	-99.04	0.50
09 - saw	Unconstrained	0.0	3	0.18	-99.04	0.50	0.18	-99.04	0.50
10 - toothbrush	Unconstrained	0.0	3	0.18	-99.04	0.50	0.18	-99.04	0.50
11 - helicopter	Unconstrained	0.9	3	1.05	-2.32	0.50	1.39	-1.56	0.48
12 - broom	Unconstrained	0.0	3	0.18	-99.04	0.50	0.18	-99.04	0.50
13 - octopus	Unconstrained	0.0	3	2.52	-0.26	0.30	2.10	-0.10	0.36
14 - mushroom	Unconstrained	2.6	3	2.77	-1.10	0.54	2.27	-0.71	0.49
15 - hanger	Unconstrained	0.7	3	-0.03	72.33	0.50	0.74	-3.51	0.50
16 - wheelchair	Unconstrained	1.0	3	52.58	-1.88	0.53	60.84	-1.58	0.48
17 - camel	Unconstrained	6.9 [*]	3	1.24	-2.56	0.50	3.35	-0.93	0.48
	1	0.1	1	1.21	-2.59	0.50	3.43	-0.90	0.50
	2	0.4	1	3.05	-1.92	0.51	3.05	-0.93	0.51
	3	6.4 [*]	1	1.57	-1.68	0.50	1.57	-1.68	0.50
18 - mask	Unconstrained	2.8	3	0.55	-3.04	0.50	1.55	-1.14	0.48
19 - pretzel	Unconstrained	2.2	3	1.56	0.28	0.32	1.57	-0.13	0.40
20 - bench	Unconstrained	1.0	3	1.00	-2.83	0.50	0.89	-8.04	0.50
21 - racquet	Unconstrained	5.4 [*]	3	72.22	-0.59	0.29	61.86	-0.63	0.55
	1	5.4 [*]	1	61.86	-0.62	0.34	61.86	-0.62	0.34
	2	0.0	1	61.86	-0.62	0.34	61.86	-0.62	0.34
	3	0.0	1	61.86	-0.62	0.34	61.86	-0.62	0.34
22 - snail	Unconstrained	23.2 [*]	3	47.94	-0.79	0.44	2.52	0.03	0.39
	1	0.2	1	47.81	-0.80	0.42	2.58	0.06	0.42
	2	1.8	1	2.86	-0.95	0.39	2.86	0.07	0.39

Item	Hypothesis	G^2	d.f.	Reference Group			Focal Group		
				a	b	g	a	B	g
	3	21.2 [*]	1	1.75	-0.25	0.44	1.75	-0.25	0.44
23 - volcano	Unconstrained	7.2 [*]	3	16.48	-0.56	0.37	2.57	-0.17	0.41
	1	0.0	1	8.25	-0.64	0.32	2.40	-0.25	0.33
	2	2.0	1	2.66	-0.68	0.30	2.66	-0.23	0.30
	3	7.1 [*]	1	2.19	-0.42	0.33	2.19	-0.42	0.33
24 - seahorse	Unconstrained	1.2	3	2.77	-0.28	0.28	1.92	0.11	0.32
25 - dart	Unconstrained	17.4 [*]	3	13.04	0.18	0.24	1.98	-0.36	0.40
	1	0.0	1	9.16	0.15	0.20	1.82	-0.54	0.20
	2	4.0 [*]	1	2.24	0.16	0.19	2.24	-0.42	0.19
	3	17.0 [*]	1	2.77	-0.10	0.26	2.77	-0.10	0.26
26 - canoe	Unconstrained	2.8	3	0.94	-0.96	0.49	1.73	-0.39	0.45
27 - globe	Unconstrained	25.1 [*]	3	15.08	-1.03	0.45	2.23	0.14	0.44
	1	0.0	1	29.91	-1.03	0.40	2.16	0.10	0.40
	2	1.4	1	2.40	-1.12	0.42	2.40	0.15	0.42
	3	24.2 [*]	1	1.44	-0.16	0.51	1.44	-0.16	0.51
28 - wreath	Unconstrained	6.7 [*]	3	0.37	-1.67	0.50	2.60	-0.24	0.48
	1	0.2	1	0.37	-1.61	0.51	2.66	-0.21	0.51
	2	6.5 [*]	1	2.30	-0.16	0.62	2.30	-0.16	0.62
	3	0.0	1	2.30	-0.16	0.62	2.30	-0.16	0.62
29 - beaver	Unconstrained	9.6 [*]	3	2.47	0.82	0.18	1.25	0.56	0.35
	1	0.0	1	2.74	0.71	0.16	1.12	0.29	0.16
	2	0.7	1	1.17	1.29	0.15	1.17	0.30	0.15
	3	13.0 [*]	1	1.45	0.51	0.20	1.45	0.51	0.20
30 - harmonica	Unconstrained	26.1 [*]	3	597.85	0.25	0.42	48.64	-0.00	0.73
	1	21.5 [*]	1	49.46	0.06	0.49	47.13	-0.01	0.49
	2	0.0	1	47.82	0.08	0.49	47.82	-0.01	0.49
	3	4.6 [*]	1	48.19	0.03	0.58	48.19	0.03	0.58
31 - rhinoceros	Unconstrained	0.0	3	2.93	-0.00	0.32	2.27	0.20	0.38
32 - acorn	Unconstrained	8.5 [*]	3	21.90	0.01	0.25	2.25	0.38	0.42
	1	0.0	1	14.17	-0.00	0.20	1.72	0.12	0.20
	2	8.9 [*]	1	1.99	0.16	0.25	1.99	0.19	0.25
	3	1.0	1	2.02	0.21	0.27	2.02	0.21	0.27
33 - igloo	Unconstrained	6.5 [*]	3	41.34	0.48	0.19	6.08	0.17	0.31
	1	0.0	1	20.42	0.46	0.15	4.69	0.08	0.15
	2	0.5	1	5.21	0.43	0.15	5.21	0.10	0.15

Item	Hypothesis	G^2	d.f.	Reference Group			Focal Group			
				a	b	g	a	B	g	
34 - stilts		3	11.7 [*]	1	6.61	0.21	0.20	6.61	0.21	0.20
	Unconstrained	15.5 [*]	3	1.33	−0.10	0.43	4.19	0.41	0.30	
		1	4.8 [*]	1	1.29	−0.15	0.40	4.57	0.47	0.40
		2	10.8 [*]	1	3.76	0.42	0.43	3.76	0.44	0.43
35 - dominoes		3	0.0	1	3.65	0.42	0.41	3.65	0.42	0.41
	Unconstrained	1.6	3	0.36	−3.14	0.50	0.80	−2.33		0.50
	36 - cactus	Unconstrained	1.3	3	26.13	−0.29	0.44	2.44	−0.45	0.45
37 - escalator	Unconstrained	12.3 [*]	3	6.96	−0.45	0.38	1.92	0.19	0.49	
38 - harp		1	0.6	1	6.53	−0.47	0.34	1.64	−0.02	0.34
		2	4.3 [*]	1	1.81	−0.52	0.38	1.81	0.06	0.38
		3	7.4 [*]	1	1.48	−0.13	0.40	1.48	−0.13	0.40
	Unconstrained	11.9 [*]	3	57.93	−0.01	0.70	61.79	0.01	0.43	
		1	11.2 [*]	1	59.24	0.02	0.64	59.60	0.02	0.64
39 - hammock		2	0.0	1	59.59	0.02	0.64	59.59	0.02	0.64
		3	0.7	1	59.19	0.02	0.61	59.19	0.02	0.61
	Unconstrained	45.5 [*]	3	4.22	−0.58	0.49	2.41	0.56	0.27	
		1	12.9 [*]	1	4.25	−0.58	0.49	3.14	0.76	0.49
		2	30.6 [*]	1	2.62	0.51	0.55	2.62	0.76	0.55
40 - knocker		3	2.0	1	2.16	0.62	0.46	2.16	0.62	0.46
	Unconstrained	4.8 [*]	3	4.22	0.48	0.35	1.73	0.15	0.42	
		1	0.0	1	3.79	0.46	0.33	1.59	0.02	0.33
		2	1.0	1	1.74	0.53	0.31	1.74	0.05	0.31
41 - pelican		3	5.0 [*]	1	2.06	0.22	0.35	2.06	0.22	0.35
	Unconstrained	0.1	3	3.66	0.19	0.21	1.57	0.56	0.28	
	42 - stethoscope	Unconstrained	12.8 [*]	3	2.36	1.07	0.17	2.81	0.39	0.33
43 - pyramid		1	0.0	1	2.41	0.93	0.12	2.39	0.24	0.12
		2	0.1	1	2.40	0.94	0.12	2.40	0.24	0.12
		3	17.7 [*]	1	2.94	0.40	0.18	2.94	0.40	0.18
	Unconstrained	17.3 [*]	3	22.81	−0.27	0.27	2.42	0.31	0.28	
44 - muzzle		1	0.0	1	23.77	−0.27	0.21	2.32	0.26	0.21
		2	4.8 [*]	1	2.60	−0.34	0.21	2.60	0.28	0.21
		3	17.0 [*]	1	1.91	0.11	0.23	1.91	0.11	0.23
	Unconstrained	15.0 [*]	3	1.26	−0.07	0.42	2.84	0.69	0.32	
		1	2.3	1	1.22	−0.13	0.39	3.06	0.75	0.39
		2	12.5 [*]	1	3.00	0.70	0.46	3.00	0.78	0.46

Item	Hypothesis	G^2	d.f.	Reference Group			Focal Group		
				a	b	g	a	B	g
45 - unicorn	Unconstrained	3	1	2.81	0.74	0.43	2.81	0.74	0.43
		8.3 [*]	3	19.63	0.07	0.27	1.80	0.63	0.39
		1	1	18.33	0.05	0.23	1.44	0.40	0.23
46 - funnel	Unconstrained	2	1	1.57	0.28	0.24	1.57	0.42	0.24
		9.5 [*]	1	1.56	0.43	0.26	1.56	0.43	0.26
		3	1	0.43	-1.60	0.51	1.96	-0.11	0.49
47 - accordion	Unconstrained	31.1 [*]	3	-175.57	1.52	0.53	2.83	1.22	0.65
		1	1	38.40	1.47	0.76	38.40	1.47	0.76
		31.1 [*]	1	38.40	1.47	0.76	38.40	1.47	0.76
48 - noose	Unconstrained	2	1	38.40	1.47	0.76	38.40	1.47	0.76
		0.0	1	38.40	1.47	0.76	38.40	1.47	0.76
		3	1	38.40	1.47	0.76	38.40	1.47	0.76
49 asparagus	Unconstrained	41.6 [*]	3	0.79	-1.68	0.50	2.45	0.43	0.31
		1	1	2.42	0.56	0.63	2.68	0.66	0.63
		40.3 [*]	1	2.66	0.55	0.63	2.66	0.66	0.63
50 - compass	Unconstrained	2	1	2.34	0.56	0.56	2.34	0.56	0.56
		0.0	1	1.02	0.74	0.39	1.27	-0.48	0.50
		15.2 [*]	3	1.09	0.53	0.37	1.18	-0.72	0.37
51 - latch	Unconstrained	1	1	1.18	0.54	0.38	1.18	-0.71	0.38
		2	1	1.69	-0.11	0.42	1.69	-0.11	0.42
		15.3 [*]	1	0.50	0.60	0.50	2.26	1.38	0.28
52 - tripod	Unconstrained	23.8 [*]	3	2.81	1.48	0.41	3.11	1.51	0.41
		1	1	3.07	1.48	0.40	3.07	1.50	0.40
		23.9 [*]	1	3.02	1.50	0.40	3.02	1.50	0.40
53 - scroll	Unconstrained	0.0	1	0.41	0.16	0.50	0.71	0.28	0.50
		0.3	3	67.80	2.22	0.21	1.66	0.87	0.20
		0.0	3	89.84	0.73	0.28	2.55	0.48	0.27
54 - tongs	Unconstrained	33.7 [*]	3	0.45	-1.38	0.50	3.06	0.99	0.32
		1	1	2.75	0.94	0.56	2.99	0.98	0.56
		33.2 [*]	1	2.99	0.94	0.55	2.87	0.96	0.55
55 - sphinx	Unconstrained	2	1	2.87	0.96	0.54	1.94	-0.02	0.54
		0.2	1	13.38	2.16	0.15	2.15	1.48	0.13
		0.2	1	12.18	2.20	0.32	1.93	0.91	0.27
56 - yoke	Unconstrained	0.0	3	12.95	0.76	0.39	1.74	0.71	0.28
		0.8	3	307.48	0.99	0.18	1.67	1.19	0.19
		0.0	3	18.13	0.83	0.22	1.26	1.41	0.23
57 - trellis	Unconstrained	0.0	3	-1951.24	-4.50	0.25	-21.37	-4.54	0.26
		0.6	3						
58 - palette	Unconstrained								
59 - protractor	Unconstrained								
60 - abacus	Unconstrained								

Note. "Unconstrained" = No parameters constrained.

"1" = asymptote (g) parameter constrained equal for reference and focal groups.

"2" = both asymptote and slope parameters constrained to be equal.

“3” = asymptote, slope, and threshold parameters constrained to be equal.

* ² significant, $p < .05$.