ASHA American Speech-Language-Hearing Association

AJSLP

Research Article

Cross-Linguistic Insights From the Boston Naming Test: Structural Comparability and Performance Comparisons of English and Korean Speakers With Aphasia

Jee Eun Sung,[a,b] Junyoung Shin,[a] Adolfo M. García,[c,d,e,f] Michael Scimeca,[g] and Swathi Kiran[g]

[a] Department of Communication Disorders, Ewha Womans University, Seoul, Korea [b] Department of Brain and Cognitive Sciences, Ewha Brain Institute, Ewha Womans University, Seoul, Korea [c] Centro de Neurociencias Cognitivas, Universidad de San Andrés, Victoria, Provincia de Buenos Aires, Argentina [d] Global Brain Health Institute, University of California, San Francisco [e] Trinity College Dublin, Ireland [f] Departamento de Lingüística y Literatura, Facultad de Humanidades, Universidad de Santiago de Chile [g] Department of Speech, Language & Hearing Sciences, Center for Brain Recovery, Boston University, MA

ARTICLE INFO

Article History:
Received September 15, 2025
Revision received December 7, 2025
Accepted March 2, 2026

Editor-in-Chief: Catherine A. Off
Editor: Nichol Castro

https://doi.org/10.1044/2026_AJSLP-25-00417

ABSTRACT

Purpose: This study examined the structural comparability of the English and Korean versions of the Boston Naming Test (E-BNT and K-BNT) and evaluated whether structural similarity translates into comparable performance patterns in individuals with aphasia.
Method: In Study 1, we analyzed two psycholinguistic features—lexical frequency and syllable length—across the E-BNT and K-BNT. In Study 2, 38 monolingual individuals with aphasia (19 English- and 19 Korean-speaking), matched on demographic and severity variables, completed the test. Overall accuracy and cluster-based item performance were compared across groups.
Results: Study 1 revealed no significant differences in lexical frequency between versions, although the K-BNT contained longer syllabic forms. In both versions, lexical frequency was strongly associated with item order, supporting a preserved graded difficulty structure. Study 2 showed comparable overall naming accuracy across groups but a significant Language × Cluster interaction. In English, lower frequency clusters showed progressively stronger associations with aphasia severity. In Korean, severity associations were observed for mid- and high-difficulty clusters, including culturally salient items, but did not follow the same monotonic gradient.
Conclusions: The findings support broad structural comparability between the E-BNT and K-BNT while demonstrating that item-level contributions to severity may differ across languages. These results inform cautious cross-linguistic research comparisons and underscore the importance of considering cultural and linguistic factors, particularly in bilingual contexts.
Supplemental Material: https://doi.org/10.23641/asha.32351364

Accurate identification of linguistic symptoms and precise characterization of impairment severity following stroke fundamentally depend on standardized assessment batteries. To fulfill this shared clinical and research purpose, widely recognized instruments such as the Western Aphasia Battery (WAB; Kertesz, 1982) and the Boston Naming Test (BNT; Kaplan et al., 1983) were developed and have since become reference standards across diverse clinical settings. Many of the most commonly used aphasia measures worldwide are derived from these English-based instruments and have been translated or culturally adapted for use in other languages. As a result, the same test labels are frequently applied across linguistic contexts.

This global dissemination creates an implicit assumption that similarly named instruments measure comparable

Correspondence to Jee Eun Sung: jeesung@ewha.ac.kr. **Publisher Note:** This article is part of the Special Issue: Select Papers From the 54th Clinical Aphasiology Conference. ***Disclosure:*** *The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.*

constructs and levels of impairment across languages. For routine monolingual clinical diagnosis within a single language, cross-language structural comparability is not required for norm-referenced interpretation. However, the motivation of the present study extends beyond within-language diagnosis. In an increasingly globalized scientific and clinical environment, evidence from different linguistic communities is frequently synthesized, cited, and used to inform theoretical models, treatment efficacy claims, and clinical guidelines. When standardized tests share the same label across languages, it is often assumed—explicitly or implicitly—that they capture comparable aspects of impairment.

Such interpretive equivalence, however, cannot be assumed. If structural and measurement properties differ substantially across language versions, cross-linguistic interpretations of severity, treatment response, or impairment patterns may inadvertently reflect language-specific test characteristics rather than true neurocognitive differences. Accordingly, it becomes essential to determine whether instruments bearing the same name operate similarly at structural and behavioral levels when administered to monolingual speakers matched on demographic and clinical variables.

Among the core components of standardized aphasia batteries, confrontation naming tests occupy a central role in quantifying lexical–semantic impairment and operationalizing anomia severity. The BNT, in particular, is widely used as a benchmark measure of naming ability and has been translated or culturally adapted into numerous languages. To date, translated and/or adapted versions of the BNT have been developed in Spanish—across several regions including Spain (Peña-Casanova et al., 2009), Argentina (Allegri et al., 1997; Allegri et al., 2001), and Latin America more broadly (Aranciva et al., 2012; Olabarrieta-Landa et al., 2015)—as well as in Danish (Jørgensen et al., 2017), French–Canadian (Roberts & Doucet, 2011), Dutch (Marien et al., 1998), Brazilian Portuguese (Miotto et al., 2010), Greek (Patricacou et al., 2007), Chinese (Chen et al., 2014; Cheung et al., 2004; Li et al., 2022), Malay (Van Dort et al., 2007), Lebanese Arabic (Chedid & Stephan, 2024), Swedish (Tallberg, 2005), several Indian languages (Sahu et al., 2024), Italian (Vestito et al., 2023), and Korean (Kim & Na, 1999).

Given its global dissemination and central role in quantifying lexical retrieval deficits in aphasia, the structural and measurement comparability of adapted versions of the BNT warrants direct empirical examination. Nonetheless, despite the widespread reliance on these versions in cross-linguistic research and clinical discourse, no study to date has conducted a systematic, head-to-head comparison of monolingual aphasia groups to determine whether structural similarities and differences are reflected in performance outcomes. Although the BNT has been widely examined with respect to translation practices, cultural bias, and psychometric refinement (e.g., Cruice et al., 2000; Jahn et al., 2013; Martínez-Ferreiro et al., 2026; Shaikh et al., 2025; Spigarelli et al., 2024; Worrall et al., 1995), there remains no direct evaluation of structural comparability and cross-language performance alignment specifically in individuals with aphasia.

Prior work has highlighted the broader concern that comparable aphasia assessment tools across languages are largely lacking, noting that many translated batteries represent direct adaptations without systematic control for linguistic, typological, cultural, and psychometric equivalence, resulting in substantial cross-national variability despite shared test labels (e.g., Martínez-Ferreiro et al., 2026). Studies that have attempted to establish cross-linguistic comparability of the BNT have focused on nonaphasic populations. For example, Jahn et al. (2013) used item response theory (IRT) to construct a Spanish–English equivalent version of the BNT, but their sample consisted of cognitively normal adults and individuals with mild cognitive impairment or dementia rather than aphasia. Similarly, Sheppard et al. (2016) examined French–English performance differences of the BNT in neurologically healthy adults and demonstrated nonequivalent performance across languages, yet without reference to aphasic populations. Investigations of cultural bias within the English version of the BNT (E-BNT) have likewise centered on multicultural samples, including individuals with cognitive impairment, but did not compare structurally adapted language versions (Shaikh et al., 2025). Notably, efforts to address cross-linguistic comparability in aphasia have emerged in alternative instruments such as the Comprehensive Aphasia Test (Swinburn et al., 2005), which underwent systematic, parallel adaptation across multiple languages to establish comparable versions specifically for people with aphasia (Martínez-Ferreiro et al., 2026). In contrast, no comparable structural and performance-based cross-linguistic evaluation has been conducted for the BNT in individuals with aphasia. This absence of direct evidence leaves unresolved whether similarly labeled BNT versions yield structurally and functionally comparable measures of anomia severity across languages.

Efforts to translate and adapt the BNT into diverse languages have generally followed two broad approaches: direct translation or cultural–linguistic adaptation. Even within English-speaking contexts, however, lexical familiarity and cultural exposure vary geographically, influencing naming performance. For example, in New Zealand, items such as *pretzel* and *beaver* yielded approximately 60% more errors than North American norms (Barker-Collo, 2001), with substitutions reflecting locally salient animals (e.g., *platypus*, *possum*). Similar concerns were raised in Australia, where culturally appropriate replacements such as *platypus* for *beaver* were recommended (Cruice et al., 2000; Worrall et al., 1995). These findings illustrate that, even within

ostensibly homogeneous English-speaking populations, item validity is sensitive to cultural context.

Cross-cultural effects become more pronounced when the BNT is adapted across languages, where lexical–semantic representation, phonological complexity, and psycholinguistic properties such as frequency, familiarity, and typicality systematically differ (Kim & Na, 1999; Sung et al., 2024, 2025). For instance, *harp* is more culturally salient in Western contexts, whereas *abacus* is more familiar in East Asian contexts. Chen et al. (2014) reported over 90% accuracy for *abacus* among Chinese speakers, compared to markedly lower accuracy in English-speaking samples (55%–60.3% in the United States and Australia; Li et al., 2022). Such discrepancies highlight how cultural familiarity directly affects item difficulty.

To date, the BNT has been translated or culturally adapted into numerous languages as previously noted. Across these versions, adaptation strategies range along a continuum: (a) direct translation with minimal modification (e.g., Chinese version retaining 30 original items; Chen et al., 2014; Cheung et al., 2004), (b) preservation of original stimuli with modified scoring criteria (e.g., Dutch, Swedish; Mariën et al., 1998; Tallberg, 2005), and (c) substantial cultural–linguistic adaptation involving large-scale item replacement. At the far end of this spectrum lies the Korean version of the BNT (K-BNT; Kim & Na, 1999), one of the most extensively adapted forms. Of the 60 items, only 10 overlap with the original E-BNT; the remaining 50 were replaced with culturally familiar stimuli. The K-BNT was developed from a Korean corpus-derived pool of 175 candidate items, reduced through naming agreement testing and normed on 600 neurologically healthy adults. Although these modifications enhance cultural relevance and ecological validity, they raise a fundamental question: When extensive item replacement alters the psycholinguistic composition of a test, does the adapted version preserve the structural hierarchy and measurement properties of the original? Evidence suggests that modifications can affect diagnostic sensitivity. For example, the modified Spanish BNT demonstrated lower sensitivity for detecting dementia-related naming impairment compared to the Texas Naming Test, a culturally developed alternative (de la Plata et al., 2008; Olabarrieta-Landa et al., 2015). Such findings underscore that adaptation improves local validity but may simultaneously compromise cross-language comparability.

Taken together, BNT adaptations reflect a tension between cultural appropriateness and structural equivalence. Although many versions retain the same test label, their lexical composition, scoring rules, and item difficulty hierarchies may differ substantially. Nonetheless, despite the extensive dissemination of these adaptations, systematic head-to-head comparisons of English and non-English versions in people with aphasia remain scarce. This gap is particularly salient for Korean, given the magnitude of its modifications. Without direct structural and performance comparisons, it remains unclear whether similarly labeled BNT versions provide comparable measures of anomia severity across languages. Addressing this question is essential not only for evaluating the interpretive boundaries of adapted tests but also for advancing cross-linguistic aphasia research beyond Anglophone-centric frameworks.

To address this gap, the present study was designed to examine a central unresolved question: When two tests share the same label yet differ in their linguistic and cultural construction, to what extent can their scores be interpreted as representing comparable levels of naming impairment across languages? To answer this question, we investigated two distinct but complementary dimensions of comparability.

First, because the K-BNT has never been formally examined in relation to the original E-BNT within a cross-linguistic framework, Study 1 evaluated structural comparability between the two versions. Specifically, we assessed whether the culturally adapted K-BNT preserves the hierarchical organization and stimulus-level characteristics of the original test despite extensive item replacement. Here, structural comparability refers to the degree of similarity in item order and in core psycholinguistic properties, such as lexical frequency and syllabic length, which contribute to graded item difficulty. We hypothesized that, if the two versions are structurally comparable, they would exhibit similar hierarchical organization and parallel distributions of key psycholinguistic properties despite the substantial cultural adaptations in the K-BNT. Establishing structural comparability is necessary to determine whether the two instruments share a similar measurement architecture. However, structural similarity alone does not guarantee equivalent behavioral performance.

Second, Study 2 examined whether structural comparability is reflected in functional alignment of naming performance across English- and Korean-speaking individuals with aphasia. We hypothesized that, if structural comparability is preserved, severity-matched groups would demonstrate broadly comparable performance patterns across item clusters, supporting functional equivalence in behavioral outcomes. Conversely, if structural comparability is observed but performance patterns diverge, this would suggest that language-specific factors influence naming performance independent of test architecture. Alternatively, if structural differences are identified, any observed performance differences must be carefully scrutinized to determine whether they arise from test construction or from genuine language-specific variation in lexical–semantic processing. By applying clustering analyses to BNT items, we evaluated how stimulus-level discrepancies shape performance profiles in the two populations.

Together, these complementary analyses allow us to move beyond assumptions of equivalence based solely on shared labeling and to distinguish between structural similarity, functional alignment, and language-specific effects. This framework provides a principled basis for determining under what conditions the E-BNT and K-BNT may support cautious cross-linguistic comparisons in aphasia research.

Study 1: Structural Comparability

Method

BNT Materials

Two language-specific versions of the BNT were used: the English version (E-BNT), Second Edition (Kaplan et al., 2001) and the Korean adaptation (K-BNT; Kim & Na, 1997). Each version contained 60 picture items, of which 10 were shared across both versions (mushroom, camel, snail, globe, harmonica, acorn, cactus, escalator, stethoscope, and funnel; Kim & Na, 1999). The complete set of 60 items for each version, together with detailed lexical properties, is presented in Supplemental Material S1.

Data Analyses

Statistical analyses were conducted in R (Version 4.5.1; R Core Team, 2025) within a Google Colab environment. For structural comparability, we examined two psycholinguistic features—lexical frequency and syllabic length—for each item in the E-BNT and K-BNT. Lexical frequency values were log transformed (log frequency) using each language's respective corpus. For English, lexical frequency values corresponded to the log10 word frequency variable from the SUBTLEX-US corpus (Brysbaert & New, 2009). For Korean, raw frequency counts were obtained from the Survey of Modern Korean Language Usage Frequency 2 (Kim, 2005) and converted to a log10 scale using the Excel LOG10 function.

For the length-related measure, we selected syllabic length rather than phoneme count to enhance cross-linguistic comparability. Although Korean orthography is alphabetic, phonological encoding, speech production, and writing are organized primarily around syllable units, which function as core prosodic and structural units (e.g., consonant–vowel [CV], VC, CVC blocks). Given English's complex consonant clusters and Korean's more constrained, syllable-based phonotactics, syllable length offers a more functionally comparable metric across languages. It better reflects production-level segmentation relevant to naming performance in both systems.

Analyses were conducted separately for the full set of 60 items and the subset of 10 overlapping items. For the total 60 items, independent t tests were used to compare psycholinguistic properties across languages. In addition, to evaluate distributional differences beyond mean comparisons, two-sample Kolmogorov–Smirnov (KS) tests were conducted for both log frequency and syllabic length. Pearson correlations were computed to examine whether item order was systematically associated with log frequency or syllabic length. For the 10 overlapping items, Kruskal–Wallis tests were used due to the small sample size and nonnormal distributions. Spearman rank-order correlations were then conducted within each language to examine associations between item order and the two psycholinguistic variables.

Results

Log Frequency and Syllabic Length

Total 60 items. For log frequency, t tests revealed no significant differences between the E-BNT and K-BNT items, $t(118) = 0.017$, $p = .986$. Consistent with the t test results, the KS test did not reach statistical significance ($d = 0.23$, $p = .068$), although the K-BNT items showed slightly greater variability in log-frequency values compared to the E-BNT.

However, the K-BNT items contained a significantly greater syllabic length compared to those of the E-BNT, $t(118) = 3.296$, $p = .001$ (see Figure 1). The KS test revealed a significant distributional difference in syllabic length ($d = 0.27$, $p = .005$), with greater variability in Korean items.

Overlapping 10 items. For the overlapping 10 items, Kruskal–Wallis tests found no significant differences between the E-BNT and K-BNT items in their psycholinguistic properties (log frequency: $\chi^2[1] = 3.29$, $p = .070$; syllabic length: $\chi^2[1] = 2.17$, $p = .141$).

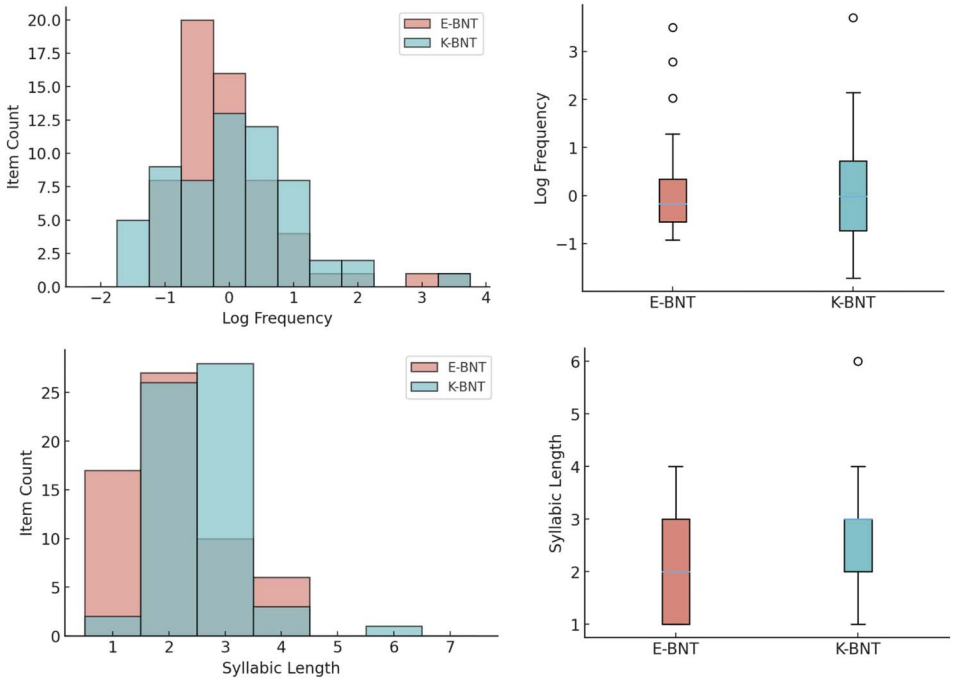
Item Order Correlations With Psycholinguistic Variables

Total 60 items. For the E-BNT, item order showed a significant negative correlation with log frequency ($r = -.570$, $p < .001$), suggesting that later presented items were less frequent in English. Similarly, no significant correlation emerged between item order and syllabic length ($r = .230$, $p = .076$).

For the K-BNT, item order was also significantly and negatively correlated with log frequency ($r = -.617$, $p < .001$), indicating that items presented later in the test tended to be less frequent words. However, no significant correlation was found between item order and syllabic length ($r = .209$, $p = .110$).

Overlapping 10 items. For the 10 overlapping items, the E-BNT showed a significant negative correlation between item order and log frequency ($\rho = -.745$, $p = .013$), suggesting that later presented items were less frequent. No significant association emerged between item order and syllabic length ($\rho = .465$, $p = .176$).

Figure 1. Distributions and mean comparisons of (A) log-transformed lexical frequency (log frequency) and (B) syllabic length for Boston Naming Test items. E-BNT = English version of the Boston Naming Test; K-BNT = Korean version of the Boston Naming Test.



In the K-BNT, item order was not significantly correlated with log frequency ($\rho = -.515$, $p = .128$). However, item order showed a significant positive association with syllabic length ($\rho = .679$, $p = .031$), indicating that the overlapping items appearing later in the test tended to have more syllables.

To summarize, analyses from Study 1 indicate that the two versions of the BNT are structurally comparable in terms of lexical frequency and item order. Building on this structural comparability, Study 2 examined whether cross-linguistic and cross-cultural effects would emerge by applying cluster analyses to subsets of items, thereby scrutinizing potential item-level influences on performance.

Study 2: Cross-Linguistic Comparisons in Aphasia

Method

Participants

Thirty-eight monolingual PWA participated in Study 2: 19 were native Korean speakers and 19 were native English speakers. All Korean-speaking participants were native Koreans recruited in Seoul, the capital city of South Korea, where the standard Korean dialect is spoken. All English-speaking participants were recruited in Boston and self-identified as native monolingual speakers of English. All individuals reported non-Hispanic ethnicity. Regarding race, 18 of the 19 English-speaking participants identified as Caucasian/White, and one participant identified as Black or African American.

Inclusion criteria were (a) diagnosis of stroke aphasia based on the WAB–Revised (WAB–R; Kertesz, 2006) for English speakers or the Korean version (Paradise–Korean Version of the Western Aphasia Battery–Revised; Kim & Na, 2012) for Korean speakers, (b) history of left-hemisphere or subcortical stroke (ischemic or hemorrhagic), (c) right-handed before the stroke, (d) being in the chronic stage (> 6 months postonset [MPO]), (e) no other neurological disease (e.g., dementia), (f) no history of developmental disorders or premorbid intellectual disability, (g) English or Korean as the first language, and (h) normal or corrected-to-normal vision.

The two language groups were matched on Aphasia Quotient (AQ; a 0–100 index of overall aphasia severity, with higher scores indicating milder impairment), age, years of education, and MPO (see Table 1). Specifically, the Korean-speaking cohort consisted of 19 participants. English-speaking participants were selected from a larger available pool, with AQ prioritized to ensure comparability in overall aphasia severity across language groups. Age, education, and MPO were subsequently considered to minimize between-group differences at the group level while maintaining independent samples. As shown in Table 1, no significant between-group differences were observed for

Table 1. Demographic information of the mean (standard deviation) for language groups.

Variable	Korean PWA (n = 19)	English PWA (n = 19)	t test results
Age (years)	48.74 (15.53)	52.47 (13.28)	$t(36) = -0.797$, $p = .430$
Sex (female:male)	5:14	6:13	NA
Education (years)	13.42 (2.56)	14.95 (2.29)	$t(36) = -1.931$, $p = .061$
WAB-AQ (total = 100)	67.61 (14.44)	70.61 (11.65)	$t(36) = 1.815$, $p = .486$
MPO	79.21 (64.45)	47.84 (38.97)	$t(36) = -0.703$, $p = .077$

Note. PWA = persons with aphasia; NA=not available; WAB-AQ = Western Aphasia Battery–Aphasia Quotient; MPO = months postonset.

AQ, age, education, or MPO (all ps > .05). Individual-level demographic and clinical information is provided in Supplemental Material S2. Ethical approval was granted by the institutional review board of Ewha Womans University (2022-0140) and by the institutional review board of the Charles River Campus at Boston University (Reference 3309E). All participants gave written informed consent before completing study procedures.

Both language groups included a mix of aphasia subtypes. In the Korean sample, anomic (36.8%), conduction (31.6%), Broca's (10.5%), Wernicke's (10.5%), transcortical motor (5.3%), and global (5.3%) aphasia were observed; in the English sample, anomic (47.4%), conduction (26.3%), Broca's (21.1%), and transcortical motor (5.3%) aphasia were observed. Collapsing syndromes by fluency yielded comparable proportions across languages—fluent (anomic, conduction, Wernicke's): Korean 15/19 (78.9%) versus English 14/19 (73.7%); nonfluent (Broca's, transcortical motor, global): Korean 4/19 (21.1%) versus English 5/19 (26.3%). Group differences were not statistically significant, $\chi^2(5) = 4.01$, $p = .55$ (fluent vs. nonfluent: Fisher's exact, $p = 1.00$). Individual-level diagnoses are provided in Supplemental Material S2.

Procedure

Participants were administered either the E-BNT (Kaplan et al., 2001) or the K-BNT (Kim & Na, 1997), according to their language group. Administration of both versions followed the standardized procedures outlined in their respective manuals, ensuring consistency in item presentation and scoring.

Data Analyses

We first compared overall accuracy for the full set of 60 items using a generalized linear mixed-effects model (GLMM), with language group (reference: English) as a fixed effect. Log frequency and syllabic length were entered as item-level covariates to evaluate whether any language-group effect remained beyond these structural characteristics. We additionally tested Pearson coefficient correlations between item order and accuracy in both versions of the BNT. For the 10 overlapping items, group differences in accuracy were evaluated using the Mann–Whitney U test due to the small sample size and nonnormal distributions, and associations between item order and accuracy were examined using Spearman rank-order correlations.

For the cluster analyses, we applied k-means clustering separately within each language group to the item-by-participant binary response matrix (Items × Participants). Each BNT item was represented as a binary response vector across participants within that language group (1 = *correct*, 0 = *incorrect*), such that each item corresponded to a point in a multidimensional space with dimensionality equal to the number of participants. k means then grouped items by iteratively minimizing within-cluster variance based on Euclidean distance between these response vectors. Clustering was not determined solely by the total number of correct responses per item but rather by which specific participants answered each item correctly. In other words, two items could have the same overall accuracy yet differ in how responses were distributed across individuals. For example, consider two items each answered correctly by 15 participants. Item A might have the response pattern: (1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0), whereas Item B might show: (1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, 1, 0). Although both items have 15 correct responses, the identity of the participants who responded correctly differs. This produces nonidentical response vectors and a nonzero Euclidean distance between items, which can lead to different cluster assignments. Within each language group, we further examined whether significant differences existed across cluster levels with respect to four dependent variables: overall accuracy, log frequency, syllabic length, and item order. Overall accuracy was analyzed using a GLMM, including cluster level as a fixed factor, covariates of log frequency and syllabic length, and random intercepts for subjects and items. The other dependent variables (log frequency, syllabic length, and item order) were analyzed with separate linear mixed-effects models that included cluster level as a fixed factor and random intercepts for items.

For the between-group comparisons in accuracy across cluster levels, we fit a GLMM with a binomial link function. The dependent variable was trial-level accuracy (0 = *incorrect*, 1 = *correct*). The fixed effects were all sum coded and included language group (English = +1, Korean = −1), cluster (Cluster 1 = −1, Cluster 2 = 0,

Cluster 3 = +1), and the interaction between language and cluster, with log frequency and syllabic length entered as covariates. Random intercepts for both subject and item were specified to account for within-subject and within-item variability. We further explored how overall and cluster-level BNT accuracy were associated with aphasia severity (AQ scores).

All analyses were conducted in R (Version 4.5.1; R Core Team, 2025) within a Google Colab environment. For statistical modeling, we used lme4 (Bates et al., 2015), lmerTest (Kuznetsova et al., 2017), and emmeans (Lenth, 2021). Independent-samples t tests, Pearson correlation coefficients, and k-means clustering were implemented or calculated via the stats package (R Core Team, 2025). Least absolute shrinkage and selection operator logistic regressions were conducted using the glmnet package (Friedman et al., 2021).

Results

Overall Performance Comparisons Between the Language Groups

Total 60 items. We first conducted a between-language group comparison of overall accuracy using a Welch independent-samples t test. There was no significant between-group difference (English group: $M = 34.42$, $SD = 13.01$, Korean group: $M = 31.95$, $SD = 11.30$), $t(35.31) = -0.63$, $p = .536$.

We then performed the adjusted analysis using a GLMM, including log frequency and syllabic length as covariates. For overall accuracy in the full set of 60 items, no significant difference was found between the language groups ($\beta = -0.224$, $p = .611$), with the English group a mean of 34.42 ($SD = 13.01$) and the Korean group scoring a mean of 31.95 ($SD = 11.30$). We further examined whether item-level accuracy was influenced by item order by analyzing Pearson correlations between item order and the number of participants who responded correctly. Strong negative associations were found for both language groups—E-BNT: $r = -.85$, $p < .001$; K-BNT: $r = -.81$, $p < .001$—indicating that the PWA in both groups had greater difficulty with items appearing later in the test (see Figure 2).

Overlapping 10 items. For the 10 overlapping items, overall accuracy did not differ significantly between the language groups. The English-PWA group ($M = 0.60$, $SD = 0.25$) and the Korean-PWA group ($M = 0.58$, $SD = 0.21$) performed comparably, with no significant difference observed (Mann–Whitney $U = 189.5$, $p = .80$). Spearman correlations revealed significant negative correlations in both languages—English: $\rho = -.733$, $p = .016$; Korean: $\rho = -.636$, $p = .048$—indicating that, as in the full 60-item set, later positioned items were more difficult for both language groups.

Cluster Analyses

k-means clustering identified three clusters within each language group (see Figures 3A and 3B for E-BNT and K-BNT, respectively). With respect to the number of clusters, although the silhouette coefficient reached its maximum at $k = 2$, the three-cluster was selected because it provided a more interpretable graded structure while still maintaining a relatively high silhouette value. Notably, the silhouette coefficient declined sharply at $k = 4$ and continued to decrease thereafter for both the K-BNT and E-BNT (Kaufman & Rousseeuw, 2009; see Supplemental Material S3). The cluster assignments of all 60 BNT items across both language groups are provided in Supplemental Material S1.

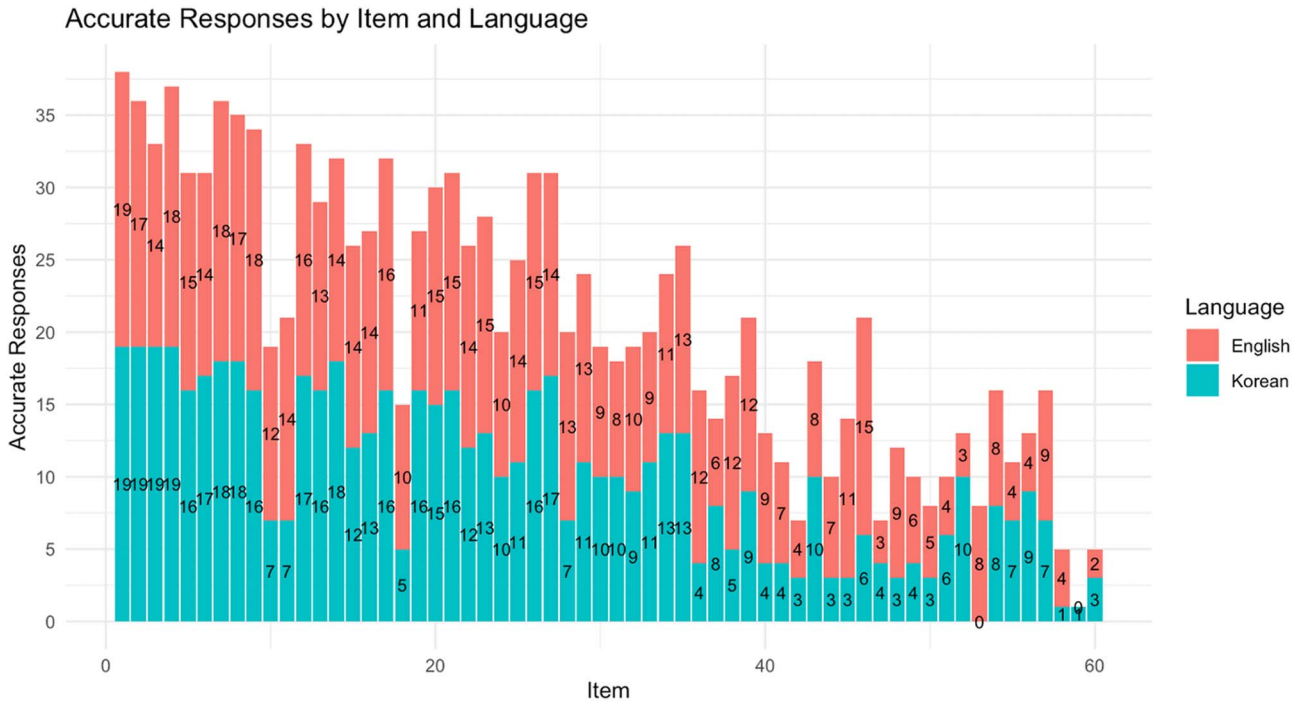
To visualize the spatial arrangement of item clusters, we applied principal component analysis (PCA) to the same item-by-participant response matrix and projected items onto the first two principal components. PCA was used solely for visualization and did not influence the clustering solution or statistical inference.

From the visual inspection of the two-dimensional (2D) plot on the K-BNT (see Figure 3A), it is interesting to note that Cluster 3 (colored in yellow) includes culturally salient items, such as the Korean national flag (Taegeukgi), the traditional drum (Janggu), the historical observatory (Cheomseongdae), and the Joseon-era warship (Geobukseon). To respect copyright restrictions on the original K-BNT artwork, the 2D plot displays concept-matched, black-and-white line drawings licensed as copyright-free from Shutterstock (www.shutterstock.com). These substitutes do not reproduce the proprietary K-BNT images. To further explore the characteristics of the clusters, we conducted follow-up analyses within each group as well as between-group comparisons.

Cluster comparisons for each language group. Within each language group, we examined whether significant differences existed among the three clusters (1–3; reference = Cluster 1) across four outcome variables: accuracy, log frequency, syllabic length, and item order. Accuracy was analyzed using a GLMM with log frequency and syllabic length entered as covariates, while the other outcomes were tested with LLMs. Descriptive statistics for each cluster are presented in Table 2 and Figure 4.

For the E-BNT, significant effects of cluster were observed across all four variables: overall accuracy, log frequency, item order, and syllabic length (see Table 2). Overall accuracy systematically decreased from Cluster 1 to Cluster 2 ($\beta = -1.32$, $p < .001$), from Cluster 1 to Cluster 3 ($\beta = -3.15$, $p < .001$), and from Cluster 2 to Cluster 3 ($\beta = 1.828$, $z = 7.382$, $p < .001$). Log frequency decreased from Cluster 1 to Cluster 2 ($\beta = -0.81$, $p = .001$) and from Cluster 1 to Cluster 3 ($\beta = -1.20$, $p < .001$), whereas the

Figure 2. Distribution of accurate responses by item order for each language group.



difference between Clusters 2 and 3 was not significant (β = 0.388, t = 1.707, p = .211). For the syllabic length, there was a significant increase from Cluster 1 to Cluster 3 (β = 0.77, p = .013) and between Cluster 2 and Cluster 3 (β = 0.699, t = 2.623, p = .030) but not from Cluster 1 to Cluster 2 (β = 0.07, p = .811). Item order increased from Cluster 1 to Cluster 2 (β = 15.46, p < .001), from Cluster 1 to Cluster 3 (β = 36.11, p < .001), and from Cluster 2 to Cluster 3 (β = 20.7, t = 6.397, p < .001).

In the K-BNT, significant effects of cluster were observed across accuracy, log frequency, and item order, but not syllabic length (see Table 2). Overall accuracy decreased progressively from Cluster 1 to Cluster 2

Figure 3. *k*-means clustering of Boston Naming Test (BNT) items for each language group: (A) English version of the BNT and (B) Korean version of the BNT. Each panel displays the distribution of BNT items across three clusters.

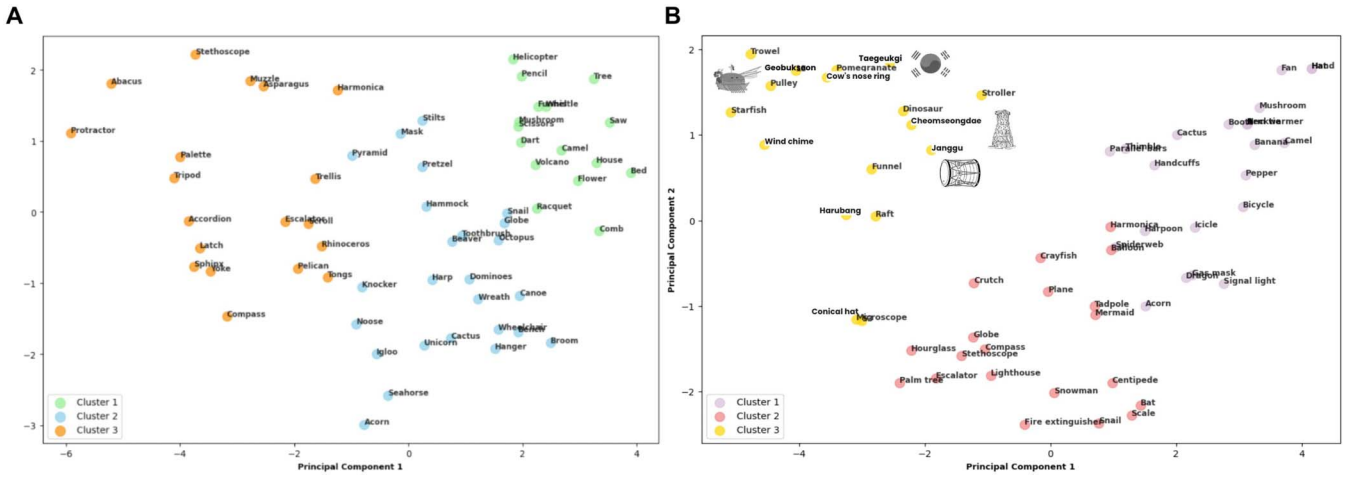


Table 2. Descriptive statistics by cluster for item order, log frequency, and syllabic length in the Korean version of the Boston Naming Test (K-BNT) and English version of the Boston Naming Test (E-BNT).

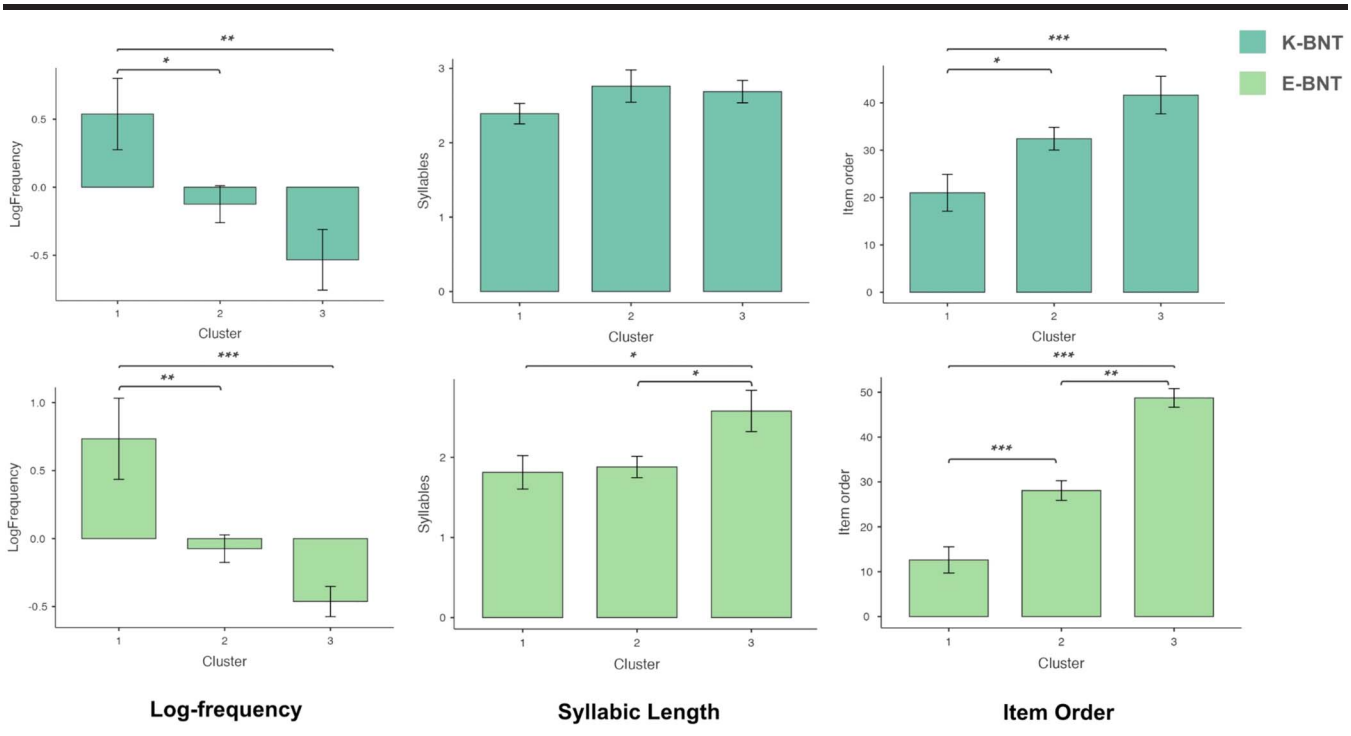
Cluster	Variable	K-BNT		E-BNT		mean_diff [95% CI]	*p*
		Item *n*	*M* (*SD*)	Item *n*	*M* (*SD*)		
1	Log frequency	23	0.538 (1.257)	16	0.734 (1.193)	–0.196 [–1.003, 0.611]	.625
1	Syllables	23	2.391 (0.656)	16	1.812 (0.834)	0.579 [0.067, 1.090]	.084
1	Item	23	21.000 (18.638)	16	12.625 (11.695)	8.375 [–1.482, 18.232]	.186
2	Log frequency	21	–0.124 (0.623)	25	–0.074 (0.508)	–0.050 [–0.394, 0.293]	.769
2	Syllables	21	2.762 (0.995)	25	1.880 (0.666)	0.882 [0.364, 1.400]	.003
2	Item	21	32.429 (11.016)	25	28.080 (10.973)	4.349 [–2.218, 10.915]	.378
3	Log frequency	16	–0.533 (0.891)	19	–0.463 (0.486)	–0.070 [–0.587, 0.446]	.780
3	Syllables	16	2.688 (0.602)	19	2.579 (1.121)	0.109 [–0.502, 0.719]	.718
3	Item	16	41.625 (15.899)	19	48.737 (9.054)	–7.112 [–16.392, 2.168]	.127

Note. *p* values were based on Welch's *t* test–applied Holm–Bonferroni correction. CI = confidence interval.

(β = −1.44, $p < .001$) and from Cluster 1 to Cluster 3 (β = −2.78, $p < .001$). A pairwise comparison using emmeans further identified a significant difference between Cluster 2 and Cluster 3 (β = 1.630, $z = 4.57$, $p < .001$). Log frequency also decreased significantly from Cluster 1 to Cluster 2 (β = −0.66, $p = .028$) and from Cluster 1 to Cluster 3 (β = −1.071, $p = .001$), whereas the difference between Clusters 2 and 3 was not significant (β = 0.408, $t = 1.259$, $p = .424$). Item order systematically increased with cluster level, from Cluster 1 to Cluster 2 (β = 11.43, $p = .018$) and from Cluster 1 to Cluster 3 (β = 20.63, $p < .001$); however, the Clusters 2 and 3 contrast was not significant (β = −9.2, $t = 1.777$, $p = .186$). No significant differences were found in the syllabic length across clusters (all *p*s > .05).

Cluster comparisons between language groups. Prior to comparing naming accuracy between language groups, we examined whether the K-BNT and E-BNT differed in structural characteristics within each cluster using Welch's *t* tests. To control for multiple comparisons, Holm–Bonferroni

Figure 4. Psycholinguistic properties per each cluster in Boston Naming Tests: log frequency, syllabic length, and item order. E-BNT = English version of the Boston Naming Test; K-BNT = Korean version of the Boston Naming Test.



corrections were applied within each cluster, with three comparisons conducted per cluster. Results are presented in Table 2. For Cluster 1, no significant between-language differences were observed in log frequency, syllabic length, or item order after correction (all ps > .05). For Cluster 2, syllabic length remained significantly greater in K-BNT than in E-BNT after correction (adjusted p = .003), whereas log frequency and item order did not differ between languages (ps > .05). For Cluster 3, no significant between-language differences were observed for any variable (all ps > .05).

We next examined overall accuracy across clusters between language groups using a GLMM, including log frequency and syllabic length as covariates. The results are summarized in Table 3. There was a significant main effect of cluster level (β = −1.321, p < .001), indicating that accuracy decreased as cluster number increased. Post hoc pairwise comparisons further showed that accuracy was significantly lower from Cluster 1 to Cluster 2 (β = −1.260, z = −5.715, p < .001), from Cluster 1 to Cluster 3 (β = −1.290, z = −6.455, p < .001), and from Cluster 2 to Cluster 3 (β = −2.55, z = −9.859, p < .001). The main effect of language group was not significant (β = 0.296, p = .218), indicating no overall difference between English and Korean speakers.

We observed a significant interaction between language group and cluster level (β = 0.231, p = .019; see Figure 5). Post hoc tests using interaction contrasts in emmeans revealed that the between-language group difference (Korean vs. English) was significantly larger at Cluster 2 than at Cluster 1 (β = 0.935, z = 2.747, p = .006) and at Cluster 3 than at Cluster 1 (β = 1.102, z = 2.736, p = .006). In contrast, the Korean–English difference between Cluster 2 and Cluster 3 was not significant (β = 0.167, z = 0.493, p = .621).

Correlations with overall aphasia severity. In the English PWA, AQ was significantly correlated with overall BNT accuracy (r = .759, p < .001), as well as with performance in all clusters—Cluster 1 (r = .566, p = .012), Cluster 2 (r = .664, p = .007), and Cluster 3 (r = .796, p < .001). In the Korean PWA, AQ was significantly associated with overall K-BNT accuracy (r = .581, p = .009). At the cluster level, AQ was significantly correlated with performance in Cluster 2 (r = .599, p = .007) and Cluster 3 (r = .538, p = .018), but not in Cluster 1 (r = .334, p = .163).

Discussion

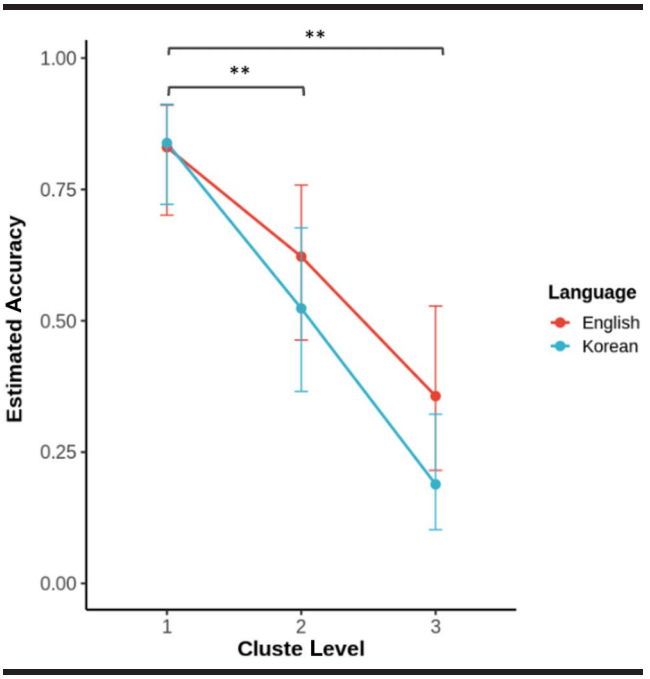
The current study demonstrated overall structural comparability between the E-BNT and its Korean adaptation (K-BNT). However, cluster analyses of the BNT items revealed item-specific effects, particularly within the cluster comprising culturally salient items for Korean-speaking PWA.

Study 1 showed that the two versions of the BNT were broadly comparable in terms of item difficulty structure, as reflected in lexical frequency. Correlational analyses indicated that, in both versions, lexical frequency was strongly associated with item order, consistent with the intended difficulty gradient of the original BNT design. These findings suggest that, despite substantial item replacement, the K-BNT preserves the structural principle of graded difficulty. Similarly, several other adapted versions of the BNT in diverse languages have demonstrated efforts to maintain this gradient—often by rearranging item order to align with lexical frequency, familiarity, and item performance (e.g., Lebanese: Chedid & Stephan, 2024; Swedish: Tallberg, 2005). In contrast, the syllabic length did not show a significant correlation with item order in either version, suggesting that syllabic length did not contribute to the hierarchical difficulty structure of the test. This result likely reflects the fact that the systematic increase in item-level difficulty was primarily intended to be controlled by lexical factors rather than phonological components. Some adaptations have attempted to match word length and syllable count

Table 3. Generalized linear mixed-effects model with language group, cluster level, and covariates as fixed effects and accuracy as the dependent variable (pairwise contrasts are indented).

Parameter	Estimate	*SE*	*z*	*p*
Language group	0.296	0.240	1.232	.218
Cluster level	−1.321	0.116	−11.342	< .001
Cluster 2–Cluster 1	−1.260	0.220	−5.715	< .001
Cluster 3–Cluster 2	−2.550	0.259	−9.859	< .001
Cluster 3–Cluster 1	−1.290	0.201	−6.455	< .001
Log frequency	0.309	0.096	3.214	.001
Syllabic length	−0.175	0.072	−2.420	.016
Language Group × Cluster Level	0.231	0.099	2.329	.019
Cluster 2 (Korean vs. English)–Cluster 1 (Korean vs. English)	0.935	0.341	2.747	.006
Cluster 3 (Korean vs. English)–Cluster 2 (Korean vs. English)	0.167	0.338	0.493	.621
Cluster 3 (Korean vs. English)–Cluster 1 (Korean vs. English)	1.102	0.403	2.736	.006

Figure 5. Language-group differences in estimated accuracy across cluster levels in Korean and English versions of Boston Naming Tests.



when replacing items that deviated from the original version (e.g., Chedid & Stephan, 2024), suggesting that, while syllable length was considered a feature to align, it was not the primary principle guiding item ordering.

Although neither BNT version showed a relationship between item order and the systematic increase in syllable length, the K-BNT contained significantly more syllables per word than the E-BNT. This suggests that the Korean adaptation may involve a greater number of syllables per lexical item, largely driven by language-specific structural features. Korean, as an agglutinative language, is characterized by rich morphophonological processes that yield longer syllabic and morphological units, encoded within syllabic Hangul blocks (Sohn, 2001). By contrast, English employs a linear alphabetic system, typically producing shorter and less syllabically dense lexical forms. These cross-linguistic differences in phonological systems provide important context for understanding why the Korean adaptation naturally results in longer word forms, even when items showed comparable lexical frequency between the two languages.

Despite the structural difference in syllable length between the E-BNT and K-BNT, no significant between-group difference in overall naming accuracy was observed. This pattern was consistent in both the unadjusted between-language comparison and the covariate-adjusted GLMM, conducted in samples matched on demographic (age, education) and clinical (aphasia severity, MPO) characteristics. This suggests that the greater syllabic length of Korean lexical items, though reflecting intrinsic properties of the language, does not jeopardize cross-group comparability in overall accuracy. Instead, it underscores the importance of matching participants on relevant variables to ensure fair cross-linguistic comparisons. In other words, while Korean items may impose inherently higher phonological load due to syllable structure, individuals with aphasia were not disproportionately disadvantaged when contrasted with their English-speaking counterparts under well-controlled conditions. However, it should be noted that, while overall accuracy was unaffected, the influence of phonological structure may emerge in outcome measures that are time sensitive. Paradigms involving temporal constraints, such as verbal fluency tasks or response latency measures, are more likely to reflect differential phonological demands across languages. In such contexts, the greater syllabic length and morphophonological complexity of Korean could plausibly contribute to slower retrieval or production, thereby producing performance differences that remain obscured when only accuracy-based outcomes are considered. A series of cross-linguistic studies employing verbal fluency tasks in Korean and English speakers, both with and without brain damage, have shown that overall accuracy levels (e.g., total animal fluency scores) were comparable between English- and Korean-speaking individuals with aphasia. However, qualitative differences emerged in the types of items generated, particularly when culturally specific criteria were applied—for instance, the inclusion of zodiac-related animals (Sung et al., 2025; see Shin et al., 2025, for evidence in Alzheimer's disease). These findings underscore the necessity of accounting for linguistic and cultural differences when developing or adapting neuropsychological assessments for diverse language populations, as such features may differentially influence task demands. Nevertheless, with appropriately designed methodological controls, valid and equitable cross-linguistic comparisons can be achieved.

Although the overall structural properties of the K-BNT and E-BNT were broadly comparable, the cluster analyses in Study 2 revealed meaningful language-specific differences. Within each version, psycholinguistic properties were primarily driven by lexical frequency, and naming performance decreased systematically across clusters, consistent with the intended hierarchical gradation of item difficulty. However, a significant interaction between language group and cluster type emerged. Specifically, the K-BNT elicited lower accuracy than the E-BNT in Clusters 2 and 3, whereas no between-language difference was observed in Cluster 1, the easiest condition characterized by higher lexical frequency and shorter syllabic length.

This raises an important question: Why were cross-language differences observed only in the more demanding clusters for Korean speakers? Between-group comparisons indicated no significant differences in lexical frequency

across clusters, suggesting that frequency alone does not account for the observed performance divergence. However, Cluster 2 items in the K-BNT had significantly greater syllabic length than those in the E-BNT, and Cluster 3 consisted largely of culturally salient proper nouns (e.g., historical or geographical landmarks) that lack direct English equivalents.

These characteristics provide plausible explanations for the interaction effect. Greater syllabic length in Cluster 2 may impose increased phonological encoding and articulatory demands in Korean, particularly under conditions of lexical retrieval difficulty. Meanwhile, Cluster 3 items, although low in corpus-based frequency, include culturally embedded proper names whose familiarity may vary across individuals and whose retrieval processes may differ from those involved in common object naming. Thus, the clusters that showed cross-language performance divergence were precisely those that incorporated either increased phonological load (Cluster 2) or culturally specific lexical content (Cluster 3).

Importantly, overall naming accuracy did not differ significantly between language groups. The observed differences emerged only at the cluster level, suggesting that language-specific phonological and cultural factors may subtly modulate performance under higher task demands without altering global accuracy scores. These findings indicate that structural comparability at the test level does not necessarily ensure uniform functional equivalence across all item subsets. To further clarify the item-level mechanisms underlying these cluster-specific effects, we are extending this line of research using larger pooled data sets across both language groups. We are developing an IRT-based framework to directly estimate item difficulty and discrimination parameters, enabling more rigorous testing of measurement invariance and differential item functioning between the E-BNT and K-BNT. The present study provides the foundational empirical rationale for these next-step item-level analyses by identifying cluster-based patterns that cannot be fully explained at the aggregate test level.

In line with the cluster-based accuracy findings discussed previously, the correlations between BNT performance and AQ further revealed language-specific patterns. In the English-speaking group, all three clusters were significantly associated with AQ, and the strength of the correlations increased systematically as lexical frequency decreased (Cluster 1: $r = .566$; Cluster 2: $r = .664$; Cluster 3: $r = .796$). This graded pattern is consistent with the intended hierarchical structure of the BNT and suggests that lower frequency items provide greater sensitivity to overall aphasia severity in the E-BNT.

In contrast, the Korean-speaking group demonstrated a less strictly graded relationship. Only Clusters 2 and 3 were significantly correlated with AQ (Cluster 2: $r = .599$; Cluster 3: $r = .538$), whereas Cluster 1 (highest-frequency items) was not ($r = .334$). Moreover, the strongest association emerged for Cluster 2 rather than Cluster 3, diverging from the monotonic frequency–severity gradient observed in the English group. Together with the previously noted cluster-level interaction, these findings suggest that, while lexical frequency remains an important organizing principle in both versions, the way item subsets map onto aphasia severity may differ across languages, potentially reflecting the influence of language-specific phonological and cultural features embedded within the adapted test.

Taken together, the correlation patterns reinforce the cluster-level findings and provide further nuance to their interpretation. The absence of a strictly graded frequency–severity relationship in the Korean group suggests that culturally embedded and item-specific features may shape how naming performance maps onto overall aphasia severity. Importantly, multiple clusters in the K-BNT were significantly associated with AQ, and we do not contend that culturally salient items are inherently superior predictors of impairment. Rather, their relative contribution within the hierarchical structure appears to differ across language versions, indicating that comparable test architecture does not guarantee identical functional weighting of item subsets. To more precisely disentangle item-level difficulty and its relation to latent severity across languages, future research should adopt item IRT approaches, as illustrated by Jahn et al. (2013). Such modeling would allow a more rigorous examination of measurement invariance and differential item functioning between the E-BNT and K-BNT, clarifying whether observed cross-language differences reflect test construction or genuine language-specific processing factors.

In clinical settings where Korean–English bilingual individuals with aphasia are evaluated, these findings warrant careful and measured interpretation. Culturally embedded items should be interpreted with caution, irrespective of conclusions regarding structural comparability between test versions. The K-BNT was normed for monolingual Korean speakers (Kim & Na, 1999), and performance may vary as a function of language dominance, exposure, and sociocultural experience. In heavily adapted versions such as the K-BNT, culturally weighted items may introduce variability unrelated to aphasia severity in bilingual individuals. This issue reflects differences in familiarity and proficiency rather than structural inequivalence per se. Accordingly, while the present study supports overall structural comparability for cross-linguistic research in monolingual populations, assessment in bilingual contexts should explicitly account for cultural exposure and language history when interpreting naming performance.

Although the current study provides novel insights into structural comparability and item clustering across the E-BNT and K-BNT, several limitations warrant consideration. First,

our analyses were restricted to two language groups with a relatively limited sample size; incorporating additional language versions (e.g., Spanish, Chinese, French) would provide a broader cross-linguistic perspective. Second, the psycholinguistic properties examined in this study were limited to a small set of variables (e.g., lexical frequency, syllabic length). This limitation reflects the current lack of large-scale Korean databases for other psycholinguistic factors, such as age of acquisition, familiarity, and imageability, which prevents their inclusion in cross-linguistic analyses. Third, while we conducted separate analyses of healthy control groups to validate the hypotheses in Study 1, we did not include healthy participants directly matched to the PWA groups in each language. Future research should incorporate larger samples encompassing a wider range of aphasia severities and subtypes, as well as matched healthy control groups, to enable more robust cross-group comparisons. In addition, further work is needed to identify which psycholinguistic variables most strongly influence cross-linguistic similarities and differences in word-retrieval performance by examining a broader set of linguistic properties. Finally, although our clustering approach identified meaningful patterns of item difficulty, future studies should extend this framework to bilingual populations, as cross-language interactions in naming performance may clarify the extent to which BNT adaptations capture universal versus language-specific aspects of lexical retrieval deficits in aphasia.

Conclusions

The present study provides evidence regarding the structural comparability and cross-language performance alignment of the E-BNT and K-BNT in monolingual individuals with aphasia. At the structural level, the two versions were broadly comparable in lexical frequency distribution and graded item ordering, despite inherent cross-linguistic differences in phonological structure, including the greater syllabic length observed in Korean. Importantly, overall naming accuracy did not differ significantly between language groups when demographic and clinical variables were matched, supporting cautious cross-linguistic comparisons at the global test level in research contexts.

At the same time, cluster-level analyses revealed a significant Language × Cluster interaction, indicating that structural alignment at the test level does not guarantee uniform functional equivalence across all item subsets. In English speakers, lower frequency clusters showed progressively stronger associations with aphasia severity, consistent with the intended difficulty gradient of the original BNT. In Korean speakers, however, the severity relationship was less strictly graded, and clusters containing greater syllabic length or culturally salient proper nouns showed differential contributions. These findings suggest that language-specific phonological and cultural features may subtly modulate performance under higher task demands, even when overall accuracy remains comparable.

Accordingly, the clinical use of each version within its respective monolingual context remains supported. However, the results underscore the importance of empirically examining structural and functional comparability when similarly labeled instruments are used to inform cross-linguistic research, synthesize global evidence, or interpret findings across language communities. In bilingual contexts, culturally embedded items should be interpreted with caution, not because structural equivalence fails per se but because familiarity and language experience may independently influence performance. Future research should extend this framework by incorporating additional language versions, larger samples, and item-level modeling approaches (e.g., IRT) to evaluate measurement invariance and differential item functioning more precisely. Such efforts will be essential for strengthening the global interpretability of standardized aphasia assessments while maintaining sensitivity to linguistic and cultural diversity.

Author Contributions

Jee Eun Sung: Conceptualization, Supervision, Writing – original draft. **Junyoung Shin:** Conceptualization, Data analysis, Writing – original draft. **Adolfo M. García:** Conceptualization, Writing – review & editing. **Michael Scimeca:** Data analysis, Writing – review & editing. **Swathi Kiran:** Writing – review & editing.

Data Availability Statement

The data sets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Acknowledgments

Jee Eun Sung is partially supported by the National Research Foundation of Korea grants funded by the Ministry of Science and ICT (RS-2022-NR070151, RS-2024-00461617) and by the Ewha Global Excellence Program. Adolfo M. García is partially supported by the National Institute on Aging (Grants R01AG075775, R01AG083799, and 2P01AG019724); ANID (FONDECYT Regular 1250317 and 1250091); Agencia Nacional de Promoción Científica y Tecnológica (Grant 01-PICTE-2022-05-00103); Programa Interdisciplinario de Investigación Experimental

en Comunicación y Cognición, Facultad de Humanidades, USACH; and the Multi-partner Consortium to Expand Dementia Research in Latin America (ReDLat), which is supported by the Fogarty International Center and the National Institutes of Health, the National Institute on Aging (Grants R01AG057234, R01AG075775, R01AG21051, and CARDS−NIH), Alzheimer's Association (Grant SG-20-725707), Rainwater Charitable Foundation's Tau Consortium, the Bluefield Project to Cure Frontotemporal Dementia, and the Global Brain Health Institute. Michael Scimeca is partially supported by the National Institute on Deafness and Other Communication Disorders (Grant F31DC021628). We express our gratitude for the thought-provoking discussions around this article's topic with members of the International Network for Cross-Linguistic Research on Brain Health (Include) and to the participants and their families who contributed to these data.

References

Allegri, R. F., Butman, J., Drake, M., Harris, P., Nagle, C., Ranalli, C., & Serrano, C. (2001). Versión abreviada 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 Neurologia, 33*(7), 624–627. https://doi.org/10.33588/rn.3307.2001238

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

Aranciva, F., Casals-Coll, M., Sánchez-Benavides, G., Quintana, M., Manero, R. M., Rognoni, T., Calvo, L., Palomo, R., Tamayo, F., & Peña-Casanova, J. (2012). Spanish normative studies in a young adult population (NEURONORMA young adults project): Norms for the Boston Naming Test and the Token Test. *Neurología, 27*(7), 394–399. https://doi.org/10.1016/j.nrleng.2011.12.010

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

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Grothendieck, G., Green, P., & Bolker, M. B. (2015). Package "lme4." *Convergence, 12*(1), Article 2.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Chedid, G., & Stephan, M. (2024). Adaptation and norm determination of the Boston Naming Test for healthy Lebanese adults aged between 50 and 88 years. *Language Testing in Asia, 14*(1), Article 25. https://doi.org/10.1186/s40468-024-00294-0

Chen, T.-B., Lin, C.-Y., Lin, K.-N., Yeh, Y.-C., Chen, W.-T., Wang, K.-S., & Wang, P.-N. (2014). Culture qualitatively but not quantitatively influences performance in the Boston Naming Test in a Chinese-speaking population. *Dementia and Geriatric Cognitive Disorders Extra, 4*(1), 86–94. https://doi.org/10.1159/000360695

Cheung, R. W., Cheung, M.-C., & Chan, A. S. (2004). Confrontation naming in Chinese patients with left, right or bilateral brain damage. *Journal of the International Neuropsychological Society, 10*(1), 46–53. https://doi.org/10.1017/S1355617704101069

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*(2), 143–155. https://doi.org/10.1080/026870300401522

De La Plata, C. M., Vicioso, B., Hynan, L., Evans, H. M., Diaz-Arrastia, R., Lacritz, L., & Munro Cullum, C. (2008). Development of the Texas Spanish Naming Test: A test for Spanish speakers. *The Clinical Neuropsychologist, 22*(2), 288–304. https://doi.org/10.1080/13854040701250470

Friedman, J., Hastie, T., Tibshirani, R., Narasimhan, B., Tay, K., Simon, N., & Qian, J. (2021). *glmnet: Lasso and elastic-net regularized generalized linear models*. CRAN R Repository. https://cran.r-project.org/web/packages/glmnet/index.html

Jahn, D. R., Mauer, C. B., Menon, C. V., Edwards, M. L., Dressel, J. A., & O'Bryant, S. E. (2013). A brief Spanish–English equivalent version of the Boston Naming Test: A project FRONTIER study. *Journal of Clinical and Experimental Neuropsychology, 35*(8), 835–845. https://doi.org/10.1080/13803395.2013.825234

Jørgensen, K., Johannsen, P., & Vogel, A. (2017). A Danish adaptation of the Boston Naming Test: Preliminary norms for older adults and validity in mild Alzheimer's disease. *The Clinical Neuropsychologist, 31*(Suppl. 1), 72–87. https://doi.org/10.1080/13854046.2017.1371337

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

Kaplan, E., Goodglass, H., & Weintraub, S. (2001). *Boston Naming Test–Second Edition*. The Clinical Neuropsychologist.

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: An introduction to cluster analysis*. Wiley. https://onlinelibrary.wiley.com/doi/book/10.1002/9780470316801

Kertesz, A. (1982). *Western Aphasia Battery*. Grune & Stratton.

Kertesz, A. (2006). *Western Aphasia Battery–Revised (WAB-R)* [Database record]. APA PsycTests. https://doi.org/10.1037/t15168-000

Kim, H. (2005). *Hyeondae Gugeo Sayong Bindo Josa 2* [A frequency survey of contemporary Korean usage II]. National Institute of the Korean Language.

Kim, H., & Na, D. L. (1997). *Korean Version–Boston Naming Test*. Hakji-sa.

Kim, H., & Na, D. L. (1999). Normative data on the Korean version of the Boston Naming Test. *Journal of Clinical and Experimental Neuropsychology, 21*(1), 127–133. https://doi.org/10.1076/jcen.21.1.127.942

Kim, H., & Na, D. L. (2012). *Paradise·Korean Version of the Western Aphasia Battery–Revised (PK-WAB-R)*. Paradise Welfare Foundation.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lenth, R. V. (2021). *emmeans: Estimated marginal means, aka least-squares means* (R package Version 4.5.1). https://cran.r-project.org/web/packages/emmeans/index.html

Li, Y., Qiao, Y., Wang, F., Wei, C., Wang, R., Jin, H., Xie, B., You, J., Jia, J., & Zhou, A. (2022). Culture effects on the Chinese version Boston Naming Test performance and the normative data in the native Chinese-speaking elders in mainland China. *Frontiers in Neurology, 13,* Article 866261. https://doi.org/10.3389/fneur.2022.866261

Marien, P., Mampaey, E., Vervaet, A., Saerens, J., & De Deyn, P. P. (1998). Normative data for the Boston Naming Test in native Dutch-speaking Belgian elderly. *Brain and Language, 65*(3), 447–467. https://doi.org/10.1006/brln.1998.2000

Martínez-Ferreiro, S., Arslan, S., Fyndanis, V., Howard, D., Kraljević, J. K., Škorić, A. M., Munarriz-Ibarrola, A., Norvik, M., Peñaloza, C., Pourquié, M., Simonsen, H. G., Swinburn, K., Varlokosta, S., & Soroli, E. (2026). Guidelines and recommendations for cross-linguistic aphasia assessment: A review of 10 years of comprehensive aphasia test adaptations. *Aphasiology, 40*(2), 215–239. https://doi.org/10.1080/02687038.2024.2343456

Miotto, E. C., Sato, J., Lucia, M. C. S., Camargo, C. H. P., & Scaff, M. (2010). Development of an adapted version of the Boston Naming Test for Portuguese speakers. *Brazilian Journal of Psychiatry, 32*(3), 279–282. https://doi.org/10.1590/s1516-44462010005000006

Olabarrieta-Landa, L., Rivera, D., Morlett-Paredes, A., Jaimes-Bautista, A., Garza, M. T., Galarza-del-Angel, J., Rodríguez, W., Rábago, B., Schebela, S., Perrin, P. B., Luna, M., Longoni, M., Ocampo-Barba, N., Aliaga, A., Saracho, C. P., Bringas, M. L., Esenarro, L., García-Egan, P., & Arango-Lasprilla, J. C. (2015). Standard form of the Boston Naming Test: Normative data for the Latin American Spanish speaking adult population. *NeuroRehabilitation, 37*(4), 501–513. https://doi.org/10.3233/NRE-151278

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*(12), 1157–1170. https://doi.org/10.1080/02687030600670643

Peña-Casanova, J., Quiñones-Úbeda, S., Gramunt-Fombuena, N., Aguilar, M., Casas, L., Molinuevo, J. L., Robles, A., Rodríguez, D., Barquero, M. S., Antúnez, C., Martínez-Parra, C., Frank-García, A., Fernández, M., Molano, A., Alfonso, V., Sol, J. M., & Blesa, R., for the NEURONORMA Study Team. (2009). Spanish multicenter normative studies (NEURONORMA Project): Norms for Boston Naming Test and Token Test. *Archives of Clinical Neuropsychology, 24*(4), 343–354. https://doi.org/10.1093/arclin/acp039

R Core Team. (2025). *R: A language and environment for statistical computing* (Version 4.5.1) [Computer software]. R Foundation for Statistical Computing. https://www.r-project.org/

Roberts, P. M., & Doucet, N. (2011). Performance of French-speaking Quebec adults on the Boston Naming Test. *Canadian Journal of Speech-Language Pathology & Audiology, 35*(3), 184–197. https://cjslpa.ca/files/2011_CJSLPA_Vol_35/No_03_214-277/Roberts_Doucet_CJSLPA_2011.pdf [PDF]

Sahu, A., Rajeshree, S., Kalika, M., Ravat, S., & Shah, U. (2024). Naming assessment in bilinguals for epilepsy surgery—Adaptation and standardization of Boston Naming Test in India. *Applied Neuropsychology: Adult, 33*(1), 201–208. https://doi.org/10.1080/23279095.2024.2343009

Shaikh, K. T., Zaidi, K. B., Wong Gonzalez, D., Dimech, C., Gilson, Z. M., Stokes, K. A., & Paterson, T. S. (2025). Cultural bias in the assessment of language: A closer look at the Boston Naming Test among multicultural Canadian older adults. *Applied Neuropsychology: Adult*. Advance online publication. https://doi.org/10.1080/23279095.2024.2449172

Sheppard, C., Kousaie, S., Monetta, L., & Taler, V. (2016). Performance on the Boston Naming Test in bilinguals. *Journal of the International Neuropsychological Society, 22*(3), 350–363. https://doi.org/10.1017/S135561771500123X

Shin, J., García, A. M., Scimeca, M., Kiran, S., & Sung, J. E. (2025, May 28). *Cross-cultural differences in animal fluency between Korean- and English-speaking individuals with Alzheimer's disease* [Poster presentation]. 54th Clinical Aphasiology Conference 2025, Albuquerque, NM, United States.

Sohn, H. M. (2001). *The Korean language*. Cambridge University Press.

Spigarelli, M., Gotti, D., & Lalancette, A. (2024). Unmasking the psychometric challenges of the Boston Naming Test in the North American multicultural context. *Canadian Journal of Speech-Language Pathology & Audiology, 48*(2), 157–170. https://cjslpa.ca/files/2024_CJSLPA_Vol_48/No_2/CJSLPA_Vol_48_No_2_2024_1317.pdf [PDF]

Sung, J. E., Scimeca, M., Li, R., & Kiran, S. (2024). Cross-linguistic and multicultural considerations in evaluating bilingual adults with aphasia. *American Journal of Speech-Language Pathology, 33*(6), 2716–2731. https://doi.org/10.1044/2024_AJSLP-23-00496

Sung, J. E., Shin, J., Scimeca, M., Li, R., & Kiran, S. (2025). Cross-linguistic and multicultural effects on animal fluency performance in persons with aphasia. *American Journal of Speech-Language Pathology, 34*(6S), 3611–3621. https://doi.org/10.1044/2025_AJSLP-24-00398

Swinburn, K., Porter, G., & Howard, D. (2005). *The Comprehensive Aphasia Test*. Psychology Press.

Tallberg, I. M. (2005). The Boston Naming Test in Swedish: Normative data. *Brain and Language, 94*(1), 19–31. https://doi.org/10.1016/j.bandl.2004.11.004

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. *Jurnal Sains Kesihatan Malaysia, 5*(1), 27–36.

Vestito, L., Mori, L., Trompetto, C., Tomatis, M., Alessandria, G., De Carli, F., Cocito, L., & Bandini, F. (2023). The 15-item version of the Boston Naming Test in Italian: Normative data for adults. *Aphasiology, 37*(1), 83–98. https://doi.org/10.1080/02687038.2021.1988506

Worrall, L. E., Yiu, E. M.-L., Hickson, L. M. H., & Barnett, H. M. (1995). Normative data for the Boston Naming Test for Australian elderly. *Aphasiology, 9*(6), 541–551. https://doi.org/10.1080/02687039508248713