EXPLORING DIFFERENTIAL ITEM FUNCTIONING AMONG HAWAI‘I RESIDENTS ON THE BOSTON NAMING TEST

A THESIS SUBMITTED TO THE GRADUATE DIVISION OF THE UNIVERSITY OF HAWAI‘I AT MĀNOA IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF EDUCATION
IN
EDUCATIONAL PSYCHOLOGY

AUGUST 2012

By
Kendra S. W. Nip

Thesis Committee:
Seongah Im, Chairperson
Katherine Ratcliffe
Michael Salzman
Acknowledgements

I would like to express my deepest appreciation to the numerous individuals whose continued support made this thesis possible. First of all, I would like to thank my advisor, Dr. Seongah Im, for her guidance and encouragement throughout my graduate career. Also, I would like to extend my gratitude to my thesis committee members, Drs. Katherine Ratliffe and Michael Salzman for their valuable insight and advice. To my husband, parents, sisters, colleagues, family and friends, thank you for endless support and always believing that this lifelong dream would eventually come true.
# Table of Contents

**Acknowledgements** ......................................................... ii  
**List of Tables** ................................................................. v  
**List of Figures** ................................................................. vi  
**Introduction** ........................................................................ 1  
**Literature Review** ............................................................... 3  
  - The Boston Naming Test ...................................................... 3  
  - The Influence of Culture on Overall Test Performance ........ .......... 4  
  - The Influence of Culture on Individual Item Performance .............. 6  
  - Differential Item Functioning .................................................. 7  
  - A Study of Differential Item Functioning  
    - Adopting Item Response Theory ............................................. 11  
  - Importance of Exploring Differential Item Functioning among  
    - Hawai‘i Residents ............................................................... 12  
**Methods** ............................................................................... 14  
  - Participants ........................................................................... 14  
  - Instruments ........................................................................... 16  
  - Psychometric Analysis ............................................................ 18  
**Results** .................................................................................. 21  
  - Overall Group Differences ..................................................... 21  
  - Item Difficulty and Item Discrimination Using Classical  
    - Test Theory ........................................................................ 21  
  - Item Difficulty and Item Discrimination Using Item  
    - Response Theory ................................................................. 23  
  - Differential Item Functioning Using Item Response Theory .......... 24
List of Tables

Table 1: Demographic Characteristics ............................................ 16
Table 2: Percentage of Correct Responses for Selected Items ............... 22
Table 3: Item Parameters, Type of DIF, and Chi-Squared Statistics for DIF Items ................................................................. 25
List of Figures

Figure 1: ICC Examples of Uniform and Nonuniform DIF .................. 11
Figure 2: Item Characteristic Curves of the Five DIF Items .............. 26
Figure 3: Logistic Regression-based DIF Graphs ............................. 28
An important operation of human communication is the ability to name objects, locations, and people by retrieving the correct word from memory. Many individuals may encounter an occasional inability to recall or name a particular item, an event often referred to as the “tip-of-the-tongue” phenomenon. However, when the loss of this ability is severe, the region of the brain responsible for this operation may be affected by disease, injury or a developmental defect. Thus, the presence of word-finding difficulty can serve as an indicator for an underlying medical condition or specific language disability.

Tests to measure word-finding ability typically involve the presentation of pictures in which the examinee provides the name, action or description of the scene displayed before him or her. Although several picture naming tasks exist, of frequent use is the Boston Naming Test (BNT; Kaplan, Goodglass & Weintraub, 1983) which previous research has shown to reliably elicit word-finding difficulties among individuals with a particular language disorder, called aphasia (Lezak, 1995)

Although the BNT has been proven to be a valuable diagnostic instrument, its usefulness can be greatly reduced when the probability of test performance is unequal among individuals from, for example, different ethnic groups with comparable ability levels. In such cases, interpretation of test scores cannot be universally made since performance is influenced by factors other than language impairment. Without research that identifies these potential test biases, misclassification of an individual’s ability could have a substantial impact on clinical decisions.
Differential item functioning (DIF) refers to a psychometric approach to investigate performance discrepancies between comparable groups on a test. Ideally, examinees with equal trait, such as knowledge, should perform similarly on each item of a test regardless of ethnicity, gender or other characteristics. However, a test item displays DIF when examinees with equivalent trait from different groups have dissimilar probabilities of achieving a correct response to an item (Hambleton, Swamingthan, & Rogers, 1991). Researchers and practitioners have been concerned with the possibility that the BNT items may be biased either for or against a particular group of examinees, such as individuals from ethnic groups not represented in the original normative data set, and thus, have urged caution when interpreting their test scores (Strauss, Sherman, & Spreen, 2006).

The BNT has been used in many countries across different cultures, and it is important to examine DIF with all populations served. One of the populations that can be unique in culture and geographic climate are Hawai‘i residents, whose performance on the BNT has never been investigated before. Since it is unknown whether the BNT is a valid test of word-finding ability for Hawai‘i residents, it is possible that mistaken diagnostic impressions have occurred. To prevent further misdiagnosis, potential item biases against Hawai‘i residents should be identified and possibly evaluated for removal from an examinee’s overall test score. Therefore, this study attempted to minimize this gap in research and explored performance discrepancies between long-time Hawai‘i residents and individuals from the continental United States by examining differential item functioning on the BNT.
Literature Review

The Boston Naming Test

In 1983, the 60-item Boston Naming Test was published as an extension of the shorter six item naming subtest found within the Boston Diagnostic Aphasia Examination assessment battery (Goodglass & Kaplan, 1983). The BNT is also used as a stand-alone or independent test of word-finding ability (Mitrushina, Boone, Razani & D’Elia, 2005). This test consists of 60 black and white drawings of various objects, arranged in order of difficulty from commonly present items, such as a bed, to rarely encountered items, such as an abacus. Examinees are presented with the stimuli and asked to spontaneously name the object. If an examinee fails to respond correctly, a semantic then a phonemic cue is given. Only correct responses provided spontaneously or following a semantic cue are included in calculating the examinee’s total test score.

For clinical judgment, total scores are used to classify word-finding ability into one of seven categories, ranging from very superior to impaired, through conversion to a Z-score value (Lezak, 1995; Mitrushina et al., 2005). In the standard classification system, each level of ability is defined by an upper and lower bound z-score. For example, borderline performance is considered having a z-score between -1.3 and -2.0.

Using the normative data set provided by the test authors, the score difference between an average and an impaired performance in individuals between 20 and 29 years old is merely 4 points, or 4 wrong answers. Put another way, if an individual, in his or her 20’s, incorrectly answers 10 test items, or achieves a total score of 50, that individual would be classified as having an impaired word-finding ability, possibly attributable to an
underlying neurological condition. In fact, the score difference between average and impaired categories for all age groups is between four and five points. Because only a small number of incorrect responses can determine the difference between levels of ability, it is important to consider the influence of other factors when interpreting test performance.

The Influence of Culture on Overall Test Performance

Previous research has already shown significant differences among ethnic groups on overall BNT performance (Boone, Victor, Razani, & Ponton, 2007; Pedraza et al., 2009; Whitfield, et al., 2000). Although overall test bias is not the focus of the present study, results from these studies provide justification for further investigation of items responses for these cultures.

Within the United States, research has shown that healthy Caucasian Americans significantly outperform healthy African Americans (Pedraza et al., 2009; Whitfield, et al., 2000). These differences also appeared in patient populations comparing Caucasians with African, Asian and Hispanic Americans, where types of health conditions were kept nearly equal among all ethnic groups (Boone et al., 2007). In this particular study, Boone and colleagues (2007) reported that African, Asian and Hispanic Americans performed, on average, 9 to 10 points lower than Caucasian Americans, after adjusting for age and educational levels. Recalling the classification method described previously, a score differential of 9 or 10 points could substantially impact an individual’s perceived level of ability. Another interesting result uncovered that for the Hispanic American group only, native-English speakers and English-as-a-second-language learners performed similarly
on the BNT, suggesting that regardless of when English was learned, differential performance exists between Hispanic and Caucasian Americans (Boone et al., 2007).

In other countries, Canadians of the same age performed comparably to Americans from the original test sample (Tombaugh & Hubley, 1997) or slightly better than other American groups (Strauss et al., 2006). However, Australians, younger New Zealanders and Dutch-speaking Belgians performed significantly worse than both Canadians and Americans (Worrall, Yiu, Hickson & Barnett, 1995 for Australians; Barker-Collo, 2001 for New Zealanders; Marien, Mampaey, Vervaet, Saerens & De Deyn, 1998 and Strauss et al., 2006 for Belgians). In fact, younger New Zealanders, on average, scored slightly more than one standard deviation below the North Americans (Barker-Collo, 2001).

Results from these studies within and outside of the United States suggest that overall performance on the BNT may be significantly influenced by cultural background rather than individual ability levels. Consequently, researchers have developed culturally specific norms, such as for African-Americans (see Strauss et al., 2006), Canadians (Tombaugh & Hubley, 1997), Australians (Worrall et al., 1995) and Dutch-Speaking Belgians (Marien et al., 1998). Or, instead of deriving cultural norms, many researchers have created a modified version of the BNT by deleting culturally inappropriate items (Cheung, Cheung & Chan, 2004; Kim & Na, 1999; Ponton et al., 1996). Both culturally specific normative data and test adaptations attempt to reduce the influence of culture on test performance so that fair judgments can be made about a person’s ability. However, some researchers argue in favor of further examination of test items to uncover other underlying factors contributing to group differences (Pedraza et al., 2009).
The Influence of Culture on Individual Item Performance

Along with overall performance, several studies have also reviewed test items and reported individual item difficulties. Item difficulties, in these studies, were expressed as the percentage of correct responses, as traditionally defined by classical test theory. However, using item difficulties in the form of percentages for comparison purposes can be problematic since percentages can be easily influenced by characteristics of the participant sample, such as age and educational achievement. Although results from these studies are limited by this measurement issue, these outcomes still highlight potentially biased items and provide useful insight into reasons for the cultural discrepancies.

An item that appeared to be especially difficult for Australians, younger New Zealanders and Dutch-speaking Belgians is the test item, pretzel, which produced less than 29% of correct responses in each of the three cultures (Barker-Collo, 2001; Marien et al., 1998; Worrall et al., 1995). This finding is a notable contrast to results from a Canadian study where 92% of participants correctly identified this item (Tombaugh & Hubley, 1997). Several authors have suggested that performance discrepancies may be due to infrequent word usage since pretzel is not a common food item in Australia and New Zealand (Barker-Collo, 2001; Worrall et al., 1995). To further expand on this notion, Barker-Collo (2001) reported that frequent responses to pretzel by New Zealanders were “knot” or “rope”, objects that are more widespread in New Zealand’s oceanic environment. Similarly, beaver, a difficult item for both Australians and young New Zealanders, elicited names of other physically comparable animals more commonly seen, such as “platypus” or “possum” (Barker-Collo, 2001). Thus, the author suggested
that any differential performance between New Zealanders and Canadians is simply a reflection of environmental experiences (Barker-Collo, 2001) and not a sign of pathology. Authors of culturally adapted versions of the BNT also agreed with pretzel’s cultural irrelevance and have excluded pretzel as a stimuli from the Korean-BNT (Kim & Na, 1999), the Ponton-Satz BNT for Hispanics (Ponton et al., 1996), and the adapted Chinese BNT (Cheung, Cheung & Chan, 2004).

However, it should be noted that the samples of Australians, younger New Zealanders, Dutch-speaking Belgians and Canadians also differed on average age and educational achievement- factors known to be significantly related to BNT performance (Mitrushina et al., 2005; Strauss et al., 2006). Therefore, it is difficult to accurately determine if the observed between-group differences are attributable to age, education, or cultural factors.

Overall, the findings from these studies suggest that it is important to recognize items that appear to be uncharacteristically challenging for examinees from a particular culture because such items embody sources of difficulty that are not relevant to the test construct and are invalid or less valid for individuals from that culture.

**Differential Item Functioning**

Since researchers have long been concerned with the possibility that test items may be potentially biased against particular groups, psychometric methods have been developed to identify these items. In the past, typically, researchers have evaluated test items under the Classical Test Theory (CTT) framework by comparing item difficulty estimates expressed as the proportion of the examinees who correctly answered the item.
However, item difficulty estimated this way is not invariant across different samples, but is largely dependent on the participants tested, making it difficult to accurately compare performances of different groups. Other psychometric approaches have been developed to identify measurement biases such as the Mantel-Haenszel method (Mantel & Haenszel, 1959), logistic regression (Zumbo, 2007) and latent trait modeling (Embertson & Reise, 2000). In latent trait modeling, test performance is conceptualized as the manifestation of an examinee’s latent trait and the resulting latent trait scores can be compared between groups (Embertson & Reise, 2000). Because latent trait modeling overcomes sampling issues under classical approaches, it has become an attractive framework to examine differential item functioning.

One method under the latent trait modeling framework is Item Response Theory (IRT) which can be used to analyze both dichotomous and polytomous items (Embertson & Reise, 2000). In IRT, a mathematical model uses item responses to determine probabilities of providing a correct answer or endorsing a particular response category at each level of the latent trait, denoted as $\theta$. For example, an equation for a two parameter logistic (2PL) model, incorporating item difficulty and item discrimination, is presented below.

$$P_i(\theta) = \frac{e^{1.7a_i(\theta-b_i)}}{1 + e^{1.7a_i(\theta-b_i)}}$$

where $a_i$ is the item discrimination parameter of an item $i$, $b_i$ is the item difficulty parameter of an item $i$, $\theta_j$ is the latent trait level of a subject $j$, and $P_i(\theta_j)$ is the probability of a correct response to an item $i$ of a subject $j$. 
A unique aspect of IRT is the ability to plot response probabilities against a spectrum of the latent ability which is represented on a standardized normal distribution scale, typically ranging from -3.0 to + 3.0 (Embertson & Reise, 2000). Furthermore, the relationship between response probabilities and corresponding trait levels can be illustrated through an item characteristic curve (ICC). In a 2-parameter logistic model, where both item difficulty and item discrimination are estimated, the ICC provides a graphical representation of both the degree of discriminability and level of difficulty for each item. Under IRT, item difficulty (b-parameter) is defined as the trait level associated with a 50% possibility of a correct response and plotted at that particular location along the curve (Embertson & Reise, 2000). More difficult items will have b-parameter estimates higher along the latent trait scale. For example, an item with $b = +1.0$ is more difficult than an item with $b = -1.0$ since the trait level associated with a 50% probability of a correct response is estimated at 1 standard deviation (SD) above the mean for the first item and at 1 SD below the mean for the second item. Conceptualized in another way, for the first item, only 16% of examinees have greater than a 50% chance of success, whereas, 84% of examinees have greater than a 50% response probability for the second item. Item discrimination (a-parameter) is represented as the slope at the location of the estimated difficulty parameter (Embertson & Reise, 2000). Steeper slopes indicate a greater degree of discrimination, meaning that the item is better able to differentiate between low and high ability examinees.

One of the well-known theoretical features of IRT is parameter invariance, which implies that ability and item parameter estimates, such as item difficulty and item discrimination, are relatively stable across different samples and therefore, are not
influenced by sample characteristics. This differs from CTT where calculation of item parameters can be changed by types of samples collected by researchers. Thus, IRT techniques offer an improvement over traditional CTT approaches by overcoming sampling limitations.

An exception to parameter invariance is when differential item functioning is detected. Ideally, examinees of equal ability should perform similarly on a test item and should not have a different probability of a correct response due to their cultural background, gender or another characteristic. However, items considered to demonstrate DIF will produce significantly different item parameters between groups. This group difference can manifest in two ways- uniform and nonuniform DIF (Pedraza et al., 2009).

When DIF is not present, the ICCs for both groups look similar and are virtually overlapping. In uniform DIF, groups differ in the location of the estimated item difficulty parameter. The ICCs for each group will maintain the same slope; however, their location on the graph will be different and visibly spaced apart (see Figure 1 below). In this case, one group is favored over another across the entire spectrum of ability levels. For nonuniform DIF, groups will have significantly different parameter estimates for both item difficulty and discrimination. The ICCs represented for each group would look clearly dissimilar and would intersect (see Figure 1 below). In this case, at a certain range of ability, one group is favored, and then this relationship is reversed at a different range of ability.
A Study of Differential Item Functioning Adopting Item Response Theory

A review of the relevant BNT literature only resulted in one article that utilized IRT methods for DIF detection. In this study, Pedraza et al. (2009) examined item responses between Caucasian and African Americans through both IRT and hierarchical logistic regression procedures to determine if IRT is a comparable method for identifying DIF items. Results show that both methods discovered the same six items with moderate DIF. These six items included dominoes, escalator, tripod, palette, muzzle and latch, where Caucasian Americans were significantly favored on the first four items and muzzle and latch were significantly easier for African Americans (Pedraza et al., 2009). These six items were not previously identified as problematic with examinees from other cultures. However, in general, it is difficult to compare results from Pedraza et al. (2009) to outcomes from other cultural studies since IRT methods were not used.

Although it is clear from this study by Pedraza et al. (2009) that a number of BNT items display DIF, it is not clear whether these differences are attributable to ethnicity.
alone since all of the African American participants were recruited from Florida and the Caucasian American participants from Minnesota. Therefore, the non-equivalent probabilities observed may be the result of different regional experiences.

**Importance of Exploring Differential Item Functioning among Hawai‘i Residents**

As described above, Pedraza et al. (2009) have expertly illustrated the attractive quality of using IRT for detecting items which may be unfairly estimating performance. To build upon this study, IRT can be further demonstrated as an effective tool for DIF research with a population not yet studied, Hawai‘i residents.

Hawai‘i residents are citizens of the United States and as a result may experience similar general education practices, television programs and participate in popular American past time activities as their North American counterparts. However, Hawai‘i possesses a considerably disparate climatic and cultural environment from North America. Since results from the literature suggest that cultural experiences within a particular region may be the source of differential performance on BNT items, it can be hypothesized that although Hawai‘i residents are a part of the American culture, due to its unique climatic and cultural conditions, items on the BNT may not be appropriate to use with this population. Therefore, the present study explored whether differences exist between Hawai‘i residents and North Americans on BNT performance utilizing IRT methodology for DIF detection.

Even in Hawai‘i, the BNT is employed to detect word-finding difficulties as an indicator of underlying disorders. Therefore, it is important to establish whether word-finding ability can be estimated in the same way as the original test sample or if not, how
we can distinguish between poor performance due to culturally irrelevant items and performance due to pathology.
Methods

Participants

Both male and female undergraduate and graduate students were recruited through flyers, class announcements, and by word-of-mouth on the University of Hawai‘i at Mānoa campus in Honolulu, Hawai‘i. The Mānoa campus was selected as the location for recruitment due to the number of available Hawai‘i resident and non-resident students. For the fall 2010 semester, the university’s Institutional Research Office reported that 13,306 Hawai‘i residents and 4,961 non-resident or Western exchange program students were enrolled at the Mānoa campus (University of Hawai‘i, 2010). Therefore, both Hawai‘i and North American residents could be selected from the same locale, reducing the potential effect of test score differences due to college campus. Students were informed their participation would help gain a better understanding about the effect of culture on object recognition. All participants received a $2.00 gift card for their time and some students received extra credit from their course instructor.

Between February and April 2012, a total of 64 students were tested on the University of Hawai‘i Mānoa campus. Due to the sensitive nature of releasing medical information, participants were not required to divulge their disability status or reveal any prior neurological diagnoses. However, observations of individual test performances and behaviors did not lead the researcher to suspect any participant of pre-existing language impairments. Therefore, all examinees were considered healthy, normal participants and no exclusions were made based on a medical condition. However, nine participants who reported that they were born and raised in a country outside of North America were
excluded from the final participant sample, since previous research advises against using the BNT with foreign-born or bilingual speakers (Strauss et al., 2006). After exclusions, the final participant sample totaled 55 examinees.

Participants were classified into two study groups, North American residents and Hawai‘i residents, based on the self-reported location of where they were raised and the length of time they reportedly resided there. All North American participants reported living in one of the 49 continental United States or Canada for more than 10 years. Similarly, Hawai‘i participants also lived in the State of Hawai‘i for 10 or more years. The one student who had lived 10 years in Hawai‘i and in North America was asked which state she was most influenced by and then subsequently assigned to that particular group.

Table 1 below displays demographic, as well as other descriptive information about the participants categorized by study group. An independent samples t-test determined that the groups were not significantly different on age ($t(53) = 1.540, p = 0.131$), years reported living in the location they were raised ($t(53) = -0.674, p = 0.503$), and years educated in the United States ($t(53) = 0.416, p = 0.679$). However, the groups were significantly different on the number of years they reported living in Hawai‘i ($t(53) = -16.153, p = 0.000$). On average, Hawai‘i participants resided in the State of Hawai‘i for 18.067 years longer than the North American participants. The North American group was composed of a mixture of ethnicities- 37% Caucasian American, 19% Asian American, 11% African American, 8% Hispanic American, 7% Hawaiian or Pacific Islander, and 19% other ethnicities or a combination of multiple ethnicities. The Hawai‘i
group was less diversified with 68% Asian American, 18% Caucasian American, 11% Hawaiian or Pacific Islander, and 4% African American.

Table 1. Demographic Characteristics

<table>
<thead>
<tr>
<th>Group</th>
<th>n</th>
<th>Gender</th>
<th>Age</th>
<th>Years in Raised State</th>
<th>Years Educated in United States</th>
<th>Years living in Hawai‘i</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Male</td>
<td>M</td>
<td>SD</td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>North American</td>
<td>27</td>
<td>41%</td>
<td>25.41</td>
<td>8.61</td>
<td>19.57</td>
<td>7.35</td>
</tr>
<tr>
<td>Hawai‘i</td>
<td>28</td>
<td>32%</td>
<td>22.46</td>
<td>5.04</td>
<td>20.77</td>
<td>5.71</td>
</tr>
</tbody>
</table>

**Instruments**

*Demographic Survey.* Prior to testing, participants were asked to complete a demographic survey created for this study (see Appendix). Participants self-reported the location (city, state, country) where they were raised, the number of years they lived at this location, and the number of years they had lived in Hawai‘i. Participants were also asked to provide general demographic details such as gender, age, and ethnicity.

*Boston Naming Test.* The test authors have published a second edition of the Boston Naming Test. However, since the current edition retains all of the same 60 items as the previous edition but only adds a short form and recognition version of the test (Mitrushina et al., 2005), the first edition was used for this study.

Regarding reliability, reports of internal consistency (coefficient alpha) for the BNT range from .78 to .96 (Strauss et al., 2006). For test-retest reliability, following a short one to two week delay, Strauss et al. (2006) reviewed a study that reported a one point gain in score (standard error of measurement equal to 1.02) after the second
administration; thus indicating scores are relatively stable over short gaps. For longer intervals, retest reliability correlations range from .62 to .89 after three administrations with healthy Caucasian Americans older than 57 years of age (Strauss et al., 2006).

Test Administration and Scoring Procedures. For adults, standard administration of the BNT instructs the examiner to begin with item 30 instead of with the first item in the booklet. If an incorrect response is provided between items 30 and 38, the examiner will then turn to item 29 and continue backwards until a correct response is spontaneously received. At any time, six consecutive incorrect answers will result in the discontinuation of the test. However, because of the exploratory purpose of this study, I adopted a modified set of testing procedures, similar to those used in Barker-Collo (2001), Marien et al. (1998), Tombaugh & Hubley (1997) and Worrall et al. (1995). Testing began with the first item and continued until the last. All participants completed all 60 test items without initiating the discontinuation rule.

Returning to standard test procedures, participants were allowed 20 seconds to spontaneously provide a response. Any self-corrections within these 20 seconds were given full credit. If an incorrect answer was given or the participant was unable to provide a response, a stimulus cue, also known as a semantic cue, was provided. After 20 seconds, if a correct response still could not be reached, a phonemic cue was given. All incorrect responses were recorded, exactly as stated by the participant, on the scoring sheet. As recommended by the test authors, one point was awarded for every correct spontaneous response or correct responses following the stimulus cue.
Psychometric Analysis

Classical Test Theory. Item difficulties, expressed as the percentage of correct responses, and item discrimination, expressed as a correlational value, were calculated separately for each group. Item discrimination was estimated by the degree of correlation between dichotomous responses to each item and the total test score. A higher positive magnitude of correlation indicates a greater degree of item discrimination, meaning a better ability to distinguish between low and high ability examinees. Since point biserial correlations permits the use of a dichotomous and continuous variable, this method was used to estimate item discrimination values for all test items with less than 100% of correct responses. Therefore, item discrimination was estimated for 27 items for the Hawai‘i group and 22 items for the North American group.

Item Response Theory. An assumption of IRT is that a person’s test responses stem from a single latent trait. In the literature, this is referred to as the unidimensionality of a test. Mplus statistical program (Muthen & Muthen, 2010) was utilized to examine the test dimensionality using a nonlinear factor analysis through a tetrachoric correlation matrix since the item responses were dichotomous.

To identify DIF items, IRTLRDIF version 2b (Thissen, 2001), a psychometric program using log-likelihood values of IRT models, was utilized. This program was selected because of its ability to detect both uniform and nonuniform DIF under a two parameter logistic model. In the first stage of the analysis, one by one, each item is removed from the whole item set to create a less constrained model. The size of the difference between the log-likelihood values of the less constrained model and the full
model will determine if this particular item should be considered as a candidate for DIF. Differences between models are represented by $G^2$ statistic on a chi-squared ($\chi^2$) distribution with one degree of freedom for each free parameter allowed in the less constrained model. A $G^2$ statistic greater than the critical value with $p < .05$ signifies the presence of DIF. Once candidate items are identified, the same procedure is used to individually test each parameter to determine where the DIF occurs. After this first run, candidate items are separated from items considered to be “DIF-free”, called anchor items. The process is repeated on the anchor item set until all items demonstrating DIF are identified.

**Logistic Regression.** As an alternative method to confirm DIF items detected through IRT methods, a logistic regression analysis was also conducted on the test items. Since logistic regression methods do not estimate ability levels in the terms of a standardized scale, total test scores were used as a measure of an examinee’s estimated ability level. To determine the presence of DIF, model comparisons were made using the difference in log-likelihood values. The first model in the procedure considers only the total scores and then group assignment is added as a categorical predictor. If the difference between these models is greater than the critical value with $p < .05$ on a chi-squared distribution, this difference would indicate that there is a significant effect for group, controlling for ability level. Thus, this item would be considered as demonstrating uniform DIF. Significant differences between a regression model including total score and group and a model with total score, group and an interaction term would indicate that the effect of group functions differently across the range of total scores. In this case, nonuniform DIF would be suspected. Probabilities were plotted against total scores to
illustrate the probabilistic differences between groups for each total score value. In addition, the size of the DIF was estimated by computing the change in Nagelkerke $R^2$ scores between models. By convention, differences in Nagelkerke $R^2$ values less than 0.13 are considered to portray a trivial magnitude of DIF. Scores between 0.13 and 0.26 possess moderate DIF and scores greater than 0.26 are deemed as large DIF.
Results

Overall Group Differences

On average, the North American participants scored 2.403 points higher (SD = 0.821) than Hawai‘i residents. A one-way ANOVA indicated that this mean difference was statistically significant ($F(1, 53) = 8.445$, $p = 0.005$). Although the Hawai‘i participant group was predominantly composed of Asian Americans, minorities also contributed more than half of the participants to the North American group. Furthermore, age and years of education in the United States did not statistically differ between groups. Therefore, the reason for the observed overall performance difference cannot be attributable to ethnicity, age or years of American education.

Item Difficulty and Item Discrimination Using Classical Test Theory

North American examinees performed better than Hawai‘i residents on 20 out of the 60 test items. Items with the greatest difference in proportion of examinees who responded correctly were items 47 (accordion), 59 (protractor) and 56 (yoke). In fact, 85% of the North American participants correctly recognized accordion in comparison to 43% of the Hawai‘i group participants. Hawai‘i examinees outperformed their North American counterparts on seven items, including items 50 (compass), 42 (stethoscope), and 49 (asparagus). However, the difference in the proportion of examinees with correct responses between the Hawai‘i and North American groups was small for all seven items, ranging from 1.9% to 4.6%. For the items where there was a difference in the rate of recognition between groups, Table 2 below displays the percentage of correct responses for each participant group.
Under the CTT framework, the most difficult item for Hawai’i examinees was item 56, a picture of the yoke. Only 10.7% of the Hawai’i participants could correctly recognize this stimulus even after the semantic cue was provided. For the North American examinees, the most difficult item was item 57, trellis, where 25.9% of the participants provided the correct answer following the first cue.

Item discrimination under CTT is expressed as the degree of correlation between each test item and the test total scores, often using point biserial correlation denoted as \( r_{pb} \). A greater positive correlation indicates that the test item has higher ability to distinguish different levels of examinees. For the Hawai’i adults, item 35, dominoes, was the most discriminating item with a high positive correlation \( (r_{pb} = 0.775) \). For this group, the least discriminating item was item 46, funnel \( (r_{pb} = -0.08) \). A negative correlation between the funnel stimulus and total test scores indicates that examinees that
generally scored very well on the test, performed poorly on this particular item. For North American adults, items 29 (beaver) and 52 (tripod) were equally the most discriminating test items \( r_{pb} = 0.859 \). Item 4, house, was the least discriminating item for this group \( r_{pb} = -0.505 \).

**Item Difficulty and Item Discrimination Using Item Response Theory**

First, the result of a nonlinear factor analysis with the tetrachoric correlation matrix confirmed the psychometric unidimensionality of the test. The first eigenvalue, 11.677, accounted for 38.9% of the total variance of the item response variables. The second eigenvalue, 4.311, accounted for 14.4% of the total variance. The difference between the first and second eigenvalues was large, 7.366, ensuring these results were appropriate for IRT analysis.

Next, using a two parameter logistic model (2PL), both item difficulty and discrimination parameters were estimated. The most difficult item for North American examinees was item 57, trellis, where the difficulty \( b \) parameter was 1.040, indicating that North American participants with a word-finding ability greater than 1 standard deviation (SD) above the mean have a 50% chance of correctly recognizing this item. For Hawai‘i participants, the most difficult item was item 56, yoke, where the difficulty parameter was 2.048, indicating that an ability greater than 2 SD above the mean is associated with a 50% probability of success. The same results were achieved under the CTT approach. The easiest test items are the items in which 100% of the examinees correctly recognized the items; therefore, the probability of a correct response is 1.00 across the ability spectrum. For the Hawai‘i participant group, 33 items were correctly
answered by all group members and 38 items were correctly answered by the North American group.

Regarding item discrimination, item 51 (latch), was the most discriminating item for the Hawai‘i adults ($a=1.458$) and item 52 (tripod) was the least ($a=0.745$). For the North Americans, item 56 (yoke) was the most discriminating ($a=1.569$) and item 59 (protractor) was the least ($a=0.685$). These results differ from those estimated using a point biserial correlation from the CTT analysis.

**Differential Item Functioning Using Item Response Theory**

The initial run using the IRTLRDIF program (Thissen, 2001) identified five DIF candidate items (pelican, accordion, noose, tripod and protractor) where either the difficulty ($b$) or discrimination ($a$) parameters significantly differed between groups. All of the other items were considered to be “DIF-free” and were selected as anchor items for the second iteration. The second operation did not result in any additional candidate items and furthermore, candidate items discovered in the initial run continued to demonstrate DIF. Item parameters, type of DIF detected and significance test results are presented in Table 3 for the five DIF items.
Table 3. Item parameters, type of DIF and chi-squared statistics for DIF items

<table>
<thead>
<tr>
<th>Item</th>
<th>Stimulus</th>
<th>North American</th>
<th>Hawai'i</th>
<th>α DIF test G² statistic</th>
<th>b DIF test G² statistic</th>
<th>Type of DIF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>(a)</td>
<td>(b)</td>
<td>(a)</td>
<td>(b)</td>
<td>(G²)</td>
</tr>
<tr>
<td>41</td>
<td>Pelican</td>
<td>-3.68</td>
<td>2.22</td>
<td>40.54</td>
<td>.075</td>
<td>14.3**</td>
</tr>
<tr>
<td>47</td>
<td>Accordion</td>
<td>1.75</td>
<td>.016</td>
<td>4.79</td>
<td>1.75</td>
<td>1.20</td>
</tr>
<tr>
<td>48</td>
<td>Noose</td>
<td>4.79</td>
<td>-1.91</td>
<td>4.79</td>
<td>0.60</td>
<td>1.75</td>
</tr>
<tr>
<td>52</td>
<td>Tripod</td>
<td>0.57</td>
<td>-4.44</td>
<td>0.57</td>
<td>-0.88</td>
<td>1.8</td>
</tr>
<tr>
<td>50</td>
<td>Protractor</td>
<td>0.23</td>
<td>-4.09</td>
<td>0.23</td>
<td>1.04</td>
<td>0.7</td>
</tr>
</tbody>
</table>

Note. \(a\) = discrimination parameter; \(b\) = difficulty parameter; \(\chi²\) = chi-squared test with 1 degree of freedom; NU = Nonuniform DIF, U = Uniform DIF; * \(p < 0.05\); ** \(p < 0.01\)

Uniform DIF was detected for accordion, noose, tripod and protractor in favor of the North American participants. In fact, the difference between groups on the difficulty parameter ranged between 1 SD and 5 SD in ability levels. For example, for the protractor stimulus, North American examinees with ability levels 4 SD below the mean had at least a 50% chance of correct recognition. In contrast, the ability level associated with the same probability of a correct response was 1 SD above the mean for Hawai'i participants. Presented in another way, for this item, North American adults with average ability, i.e., \(\theta = 0\) had an 83% chance of correct recognition. For Hawai'i adults with exactly the same estimated ability level, their likelihood of a correct response was lower at 40%.

A peculiar item was item 41 (pelican) in which nonuniform DIF was detected. For this item, the probability of correct recognition was nearly 0 for Hawai'i participants until ability levels reached 0.6 SD above the mean \((\theta = 0.6)\) when a steep increase in probability estimates was observed. In contrast, the probability of correct response was 1.0 for North American examinees until ability levels neared 2 SD above the mean \((\theta = \)
2.0) when the probability estimates severely declined. It is unusual that the probability of correct recognition would be negatively associated with higher ability levels.

Figure 2. Item characteristic curves of the five DIF items

**Item 41 - Pelican**

**Item 46 - Accordion**
Item 48 - Noose

Item 52 - Tripod

Item 59 - Protractor
Differential Item Functioning Using Logistic Regression Analysis

Results from the logistic regression analysis indicated that for two items- item 47 (accordion) and item 52 (tripod), the change in -2Loglikelihood (-2LL) values between regression models reached significance on a chi-squared test. For item 47 (accordion), a significant effect for group alone was also revealed ($b = 1.598$, $p = 0.023$). This result suggests the presence of uniform DIF where North Americans had a greater average probability of a correct response when compared to Hawai‘i participants who achieved the same total test score. IRT-based analysis also reached the same conclusion.

Furthermore, change in Nagelkerke $R^2$ values indicated that the DIF effect size is small ($\Delta R^2 = 0.121$). However, for item 52 (tripod), logistic regression methods reached a different conclusion than the IRT-based approach. Instead, nonuniform DIF was detected with a medium DIF effect size ($\Delta R^2 = 0.172$). However, the interaction between group and total score was not found to be a significant predictor ($b = 0.31$, $p = 0.158$). The lack of significance on this predictor could be the result of a small sample size. Figure 3 below plots predicted probabilities against total score for each group to showcase the differential likelihoods in a correct response.

Figure 3. Logistic Regression-based DIF graphs
Discussion

The current study examined potential item biases of the BNT using IRT and logistic regression. The results indicated that five items of the BNT presented significantly different probabilities of correct responding between long time Hawai‘i and North American residents with equivalent overall word-finding ability. Four of these items, (accordion, noose, tripod and protractor) displayed uniform DIFs which signified the item difficulty parameters would be significantly different between groups. In fact, for these four items, the conditional probability of success was consistently lower for Hawai‘i residents across the entire spectrum of the latent ability. Furthermore, one item, pelican, exhibited nonuniform DIF where the favored group changed across the ability spectrum resulting in significant differences on both item difficulty and discrimination parameters.

Further inspection of the incorrect responses to the pelican item suggested that Hawai‘i examinees mistakenly perceived this stimulus as a seagull. In fact, seagull was the most common response to the pelican drawing, accounting for nearly 86% of the incorrect responses. In comparison, none of the North American examinees mistakenly recognized the pelican as a seagull. A checklist of endemic bird species found in the state of Hawai‘i indicated that several different species of gulls are prevalent in the islands but birds from the pelican family are not (Pyle, 2002). As Barker-Collo (2001) suggested, the reason for differential performance between cultures may reflect the lack of an object’s presence in one’s immediate environment. Therefore, examinees raised in Hawai‘i may have performed poorly on this item due to the lack of pelicans found in Hawai‘i and is not reflective of a word-finding deficit. However, IRT results for this
item also indicate that Hawai‘i examinees estimated with above average ability will have near a 100% chance of correct recognition. Perhaps this outcome is the result of availability and exposure to other media sources such as television, film, and the internet, which featured bird species outside of Hawai‘i or opportunities to travel to states where pelicans are prevalent. Future studies may want to investigate the influence of media on mitigating differential item performance.

For accordion, noose, tripod and protractor, the overwhelming response was “I don’t know” among Hawai‘i participants which indicated a clear lack of object recognition. Even after stimulus and phonemic cues were given, the Hawai‘i participants either provided the correct answer or were still unable to recognize the object. Although the information collected for this study does not reveal a clear reason for the presence of DIF on these four items, it can be speculated that this outcome occurred because accordions are not a common musical instrument in Hawaiian culture and that Hawaiian history might lack the use of a noose or public hangings as a form of capital punishment. However, tripods and protractors are commonly used items in Hawai‘i. Barker-Collo (2001) suggested that tripods are no longer common classroom items; however in this study, Hawai‘i and North American participants did not differ on age therefore suggesting that the reason for the presence of DIF for the tripod item is most likely due to another factor.

To avoid spurious findings due to the use of a single methodology, item responses were also analyzed using a statistical method. Results from logistic regression methods detected the presence of DIF among two items- accordion and tripod. Together with
IRT-based DIF outcomes, it can be concluded that at least these two items exhibit significantly different probabilities of success as a result of group membership.

Pedraza and colleagues (2009) also examined BNT items between Caucasian and African Americans with IRT and logistic regression. Altogether, they identified 12 DIF items under an IRT approach and 6 items were confirmed to possess a moderate amount of DIF under a logistic regression approach (Pedraza et al, 2009). Using IRT methods, noose, tripod and protractor were identified as DIF items by both Pedraza and colleagues (2009) and in the present study. Furthermore, in both studies, DIF was confirmed through logistic regression in one item, tripod. However, unlike the present study, one cultural group did not show an advantage over all recognized DIF items. Instead, African Americans were significantly favored on noose and protractor while Caucasian Americans found tripod easier (Pedraza et al, 2009). For the item, pelican, similar to Hawai‘i residents, Pedraza and colleagues (2009) reported that Caucasians displayed an advantage over African Americans on this item detected only by a logistic regression approach. However, for the item, accordion, Caucasians and African Americans had non-significant difficulty parameters (Pedraza et al, 2009). Therefore, the differential performance observed on the accordion item may be unique to Hawai‘i examinees. Overall, results from Pedraza and colleagues (2009) and the current study provide a strong argument for the presence of DIF among three BNT items (noose, tripod and protractor) using IRT methods which affect examinees from three different populations.

Because of the widespread use of cognitive testing, it is important to uncover construct irrelevant items that threaten the overall validity of the test. When items produce non-equivalent parameter estimates for different population groups, further
examination of these items is warranted by an expert panel to conclude whether a bias exists. In addition, authors of the BNT originally designed the items of this test to appear in increasing order of difficulty (Goodglass & Kaplan, 1983). However, results from this study show that for both Hawai‘i and North American participant groups, items are lacking any type of orderly pattern and instead fluctuate between easy and difficult. In fact, for the North American group, one of the easiest items where 100% of these participants provided a correct response occurred near the end of the test, on item 48 out of 60. Other studies conducted elsewhere have also reached the same conclusion and researchers have suggested revising the order of appearance using item difficulty parameter estimates collected from the body of literature (Barker-Collo, 2007; Pedraza et al, 2009; Tombaugh & Hubley, 1997).

This study was designed to be exploratory in nature since BNT performance has not been studied with long-time Hawai‘i residents. The generalization of the results can be limited in that the number of the participants was small, and all participants were undergraduate or graduate students from the same university campus. Future studies should focus on expanding the participant sample to be more representative of the general Hawai‘i population characteristics. Also, larger sample sizes will allow for tests on parameter invariance within and between cultural groups that would bolster the findings for the presence of DIF items.

Given that the BNT is commonly utilized in Hawai‘i, the detection of DIF among the test items are concerning in that the probability of success is contingent on the examinee’s geographical region and not the manifestation of a word-finding deficiency. Furthermore, since only a small number of incorrect responses can impact the final
classification of the examinee, it is imperative to address these concerns to avoid misperceptions of performance that can mistakenly impact diagnostic conclusions.
Appendix: Demographic Survey

Project Number __________
______________

Date
______________

Gender: Male or Female

Age: __________

Ethnicity:

*If you have more than one ethnicity, please check the ethnic background that you most identify with.*

___ White (not Hispanic)   ___ Filipino  ___ Hawaiian
___ African American  ___ Vietnamese  ___ Guamanian/Chamorro
___ Native American Indian  ___ Korean  ___ Micronesian
___ Hispanic  ___ Thai  ___ Samoan
___ Japanese  ___ Laotian  ___ Tongan
___ Okinawan  ___ Indian  ___ Maori
___ Chinese
___ Taiwanese

Other (describe):
__________________________________________

Primary Language spoken ____________

Second Language spoken ____________

Where you were raised? __________________

________________________

______________  ____________  ____________

City  State  Country

(If you were raised in the United States, please list the city and state where you were raised. If you were raised outside of the United States, please list the country where you were raised.)
Number of years living in the state where you were raised: 
___________________

Number of years living in Hawai‘i: ________________

Number of years educated in the United States: ____________________
References


Ponton, M.O., Satz, P., Herrera, L, Ortiz, F., Urrutia, C. P., Young, R., … Namerow, N. (1996). Normative data stratified by age and education for the
Neuropsychological Screening Battery for Hispanics (NeSBHIS): Initial report.

*Journal of the International Neuropsychological Society, 2, 96-104.*


Zumbo, B. (2007). Three generations of DIF analyses: Considering where it has been, where it is now and where it is going. *Language Assessment Quarterly, 4*(6), 223-233