Identifying Sources of Differential Item Functioning on an English Language Proficiency Assessment

By

Amy Clark

Submitted to the graduate degree program in Educational Psychology and Research and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Master of Science.

_____

Chairperson Neal Kingston

_____

Vicki Peyton

_____

Bruce Frey

Date Defended: December 8, 2011

The Thesis Committee for Amy Clark
certifies that this is the approved version of the following thesis:

Identifying Sources of Differential Item Functioning on an English Language Proficiency Assessment

_____

Chairperson Neal Kingston

Date Approved: December 9, 2011

Abstract

This paper examines the degree of DIF detected on an English language proficiency assessment and the extent to which item characteristics may be identified as sources of DIF for Vietnamese- and Spanish-speaking students. Logistic regression was used to determine the extent of DIF for each item. Of the 45 items individually analyzed, 27 items were flagged for evidence of containing uniform or nonuniform DIF, the majority of which favored Vietnamese-speaking students. Effect sizes reflecting the magnitude of DIF for each item were correlated with item characteristics to determine the extent that item characteristics could explain variability in proficiency. Both multisyllabic words and unique-to-English sounds were significantly correlated with effect size, indicating these variables may partially explain the difference in item proficiency between language groups.

Keywords: differential item functioning, English language proficiency

Introduction

In the United States, there is a growing population of English language learner (ELL) students. According to the National Clearinghouse for English Language Acquisition (2011), during the 2008-2009 school year there were approximately 5,346,673 ELL students enrolled in public schools. This number represents a 51% increase from ten years prior. As a result, ELL students now account for roughly 10% of the total public school population (Kopriva, 2008). This population is linguistically diverse, speaking over 400 different languages, with the two largest language groups being Spanish and Vietnamese native speakers (Kopriva, 2008).

When the Elementary and Secondary Education Act was reauthorized in 2001 (NCLB, 2001), Title III of the legislation included provisions to ensure the academic success of ELL students. Specifically, the legislation required the development and assessment of standards for ELL students in each of the fifty states (Kopriva, 2008). Scores obtained from these assessments are used to determine student placement and eventual exit from ELL programs, in addition to being used to monitor yearly progress for accountability purposes; therefore it is essential for the assessments to be valid and fair measures of what students know and are able to do.

One way to evaluate if test items accurately assess subgroups of the population is to examine performance on individual items. Items are said to have differential item functioning (DIF) if students of similar overall proficiency are more or less likely to correctly respond to an item based on their group membership

(Scheuneman & Grima, 1997; Angoff, 1993; Mazor, Kanjee, & Clauser, 1995). If items are found to function differentially based on language group, it may mean decisions regarding student proficiency have not been entirely accurate, thus impacting student placement decisions, exit status, and even school funding.

Literature Review

Differential Item Functioning

The practice of using statistics to identify items that function differently for subgroups of the population arose from issues of item bias in the 1960's, in which items assessed content that was not universally known by all cultural subgroups (Angoff, 1993; Cole, 1993). Contrasted with identifying bias, differential item functioning became a method that was used to explain a difference in item performance that was statistical, rather than social (Angoff, 1993; Sasaki, 1991; Dorans & Holland, 1993). Since that time, DIF has become a popular method for analyzing differences in the performance of ethnic, language, and gender groups.

Although there are various methods for analyzing DIF, each includes the same component parts. A variable that represents the construct of interest serves as a measure of student proficiency (Angoff, 1993). This is typically the total score of the items being examined for DIF; however, some studies have employed the use of an equated, external score that has previously been found to be free of DIF as the proficiency variable (Ferne & Rupp, 2007). A grouping variable is also included, which typically consists of two groups, a focal and reference group (Swanson, Clauser, Case, Nungester, & Featherman, 2002). Items found to assess the groups differently after controlling for proficiency are flagged as containing DIF. Follow up

analyses are then conducted to determine potential underlying causes for these differences (Scheuneman & Grima, 1997).

Two of the most common methods for detecting DIF are the logistic regression and Mantel-Haenszel methods (Swaminathan & Rogers, 1990; Rogers & Swaminathan, 1993; Mazor, Kanjee, & Clauser, 1995; Hidalgo & López-Pina, 2004; Dorans & Holland, 1993). Comparative studies have found both methods to be of equivalent power when detecting uniform DIF, which represents a level of DIF that is consistent across all levels of proficiency (Swaminathan & Rogers, 1990). However, unlike the Mantel-Haenszel method, logistic regression can accommodate the inclusion of an interaction term, which allows for detection of nonuniform DIF, in which items function differently as a result of the interaction between proficiency and group membership (Rogers & Swaminathan, 1993; Hidalgo & López-Pina, 2004). Thus, logistic regression is able to detect disordinal interactions, in which the group favored to correctly respond to an item changes depending on proficiency level. The Mantel-Haenszel method is insufficient for this purpose (Rogers & Swaminathan, 1993).

DIF Analyses for Language Assessments

A limited number of the technical manuals created for state accountability language assessments have included DIF studies comparing language groups. One explanation for the limited number of DIF studies examining proficiency assessments is the necessity for large sample sizes, with many statistical methods requiring at least 200 observations (Sasaki, 1991). For states with small ELL populations, the number of students assessed within each language group,

particularly for underrepresented language groups, is too small to be included in the analysis. In an attempt to address this issue, DIF analyses included in technical manuals have frequently compared the largest group, native Spanish speakers, with non-Spanish speakers, rather than with other specific language groups (MacGregor et al., 2010; Kopriva, 2008; Bowen, 2011; Lara et al., 2007; Martiniello, 2007; Mahoney, 2008). By including an amalgam of languages into a single category, interpretation of results for individual language groups can be difficult.

DIF studies have not been limited to examining assessments used for state accountability purposes. Additional language assessments have been examined for DIF to determine how items perform as a function of language. One such study compared the performance of Chinese- and Spanish-speaking test takers on the English as a Second Language Placement Examination (ATB Test Administration Manual, 2011). DIF was detected for only four items on the 150-item assessment, each favoring Spanish-speaking students. Following a visual inspection of the items, the DIF was attributed to the inclusion of Spanish-English cognates on all four items (Chen & Henning, 1985). A similar study analyzing items on the LanguEdge (LanguEdge Courseware, 2002) reading comprehension assessment detected DIF for 10 items using a confirmatory method. It was concluded that the items measured multiple constructs and thus functioned differently for Indo-European versus non-Indo-European test takers (Jang & Roussos, 2009).

Many DIF analyses of language assessments have extended beyond tests of proficiency in English to examine performance in other second languages (Ferne & Rupp, 2007). In one particular study, differences in performance were found

between Russian- and Arabic-speaking examinees on a Hebrew proficiency assessment. Upon further examination of the items, it was determined that the presence of DIF on vocabulary and grammar items was due to the structural similarities between Arabic and Hebrew as well as the inclusion of cognates (Allalouf & Abramzon, 2008). For a comprehensive review of the literature pertaining to DIF studies on language assessments, see Ferne & Rupp, 2007.

Identifying Sources of DIF

Perhaps one of the most challenging aspects of conducting DIF analyses is identifying the precise sources of DIF. One common method for determining sources of DIF on language assessments has been to have content experts examine items that have been flagged for DIF to identify shared features that may be the underlying cause for differences in student response patterns (Lara et al., 2007; Kopriva, 2008; MacGregor et al., 2010; Colorado Department of Education, 2010). However, this method has proven largely ineffective due to the difficulty of determining similarities between seemingly discreet items that may favor one group over the other (Linn, 1993; Clauser, Nungester, & Swaminathan, 1996; Allalouf & Abramzon, 2008; Abbott, 2007). If specific sources of DIF can be identified, the validity of language assessments will improve, thus allowing stakeholders to make more accurate decisions regarding student proficiency.

One way to more reliably identify sources of DIF is to use statistical methods to examine items for shared characteristics that may impact student performance. One such method makes use of effect size changes that result from the inclusion of the group and interaction term in a logistic regression. These effect sizes are used as

outcomes in a multiple regression, with item characteristics included as predictors to investigate sources of DIF (Zumbo, 1999; Swanson, et al., 2002).

One of the most frequently reported effect size measures for logistic regression is the Nagelkerke pseudo R^2 (Zumbo, 1999). The Nagelkerke pseudo R^2 statistic has a full range of possible values from zero to one, and as such is considered an improvement upon the Cox and Snell pseudo R^2 measure of effect size, which has a maximum value of less than one (Allen & Le, 2008). However, Peng, Lee, and Ingersoll (2002) warn that while the Nagelkerke pseudo R^2 can provide information about a model's goodness-of-fit, it can not be interpreted in the same way as the coefficient of determination in ordinary least squares regression, which is used to explain the amount of variance in the dependent variable that is accounted for by the independent variables. Thus, they argue that it should be interpreted with caution and only used in conjunction with other indices when determining a model's overall fit.

Item Characteristics by Language Group

In order to attribute DIF to a specific source, item characteristics relevant to the reference and focal groups must be identified. Transfer theory is one of the most widely accepted hypotheses regarding second language acquisition. According to this theory, linguistic skills obtained in the first language transfer when learning a second language (Gass, 1988). Thus, linguistic skills found in both the first and second language may be acquired more readily, while skills that exist only in the second language may take more time to master. Differences in language transfer

between Spanish and Vietnamese speakers may explain discrepancies in performance on English language proficiency assessments.

One major difference between the two languages is that Spanish and English share cognates, while Vietnamese and English do not. Cognates are words that are spelled similarly and have similar meaning across multiple languages (Nagy, Garcia, Durgunoglu, & Hancin-Bhatt, 1993). Spanish-speaking students may be more likely to correctly respond to an item due to the roughly 15,000 cognates shared between English and Spanish, many of which are spelled almost identically (Nagy et al., 1993; Cunningham & Graham, 2000). Studies have found Spanish-speaking students are able to use cognates to process information in English and improve their performance on vocabulary and comprehension assessments (Nagy et al., 1993; Cunningham & Graham, 2000; Chen, Ramirez, Luo, Geva, & Ku, 2011; Jang & Roussos, 2009; Chen & Henning, 1985). However, much of the research conducted on the use of cognates has examined students in fourth grade and above since examinees must have significant vocabulary knowledge in both the first and second language in order to recognize the words as cognates.

Another major difference between the two languages is in how Spanish and Vietnamese are represented phonetically and orthographically, despite both having a basis in Latin. The Spanish alphabet consists of thirty letters, including the twenty-six letters of the English alphabet and four unique letters. Spanish only contains ten vowel sounds, as opposed to the twenty vowel sounds found in the English language (Barlow, 2005; Hegde & Pomaville, 2008; Fashola, Drum, Mayer & Kang, 1996; Coe, 2001).

In contrast, the Vietnamese alphabet consists of twenty-nine letters: twenty-two letters similar to those found in the English alphabet, excluding f, j, w, and z, and seven unique letters. Vietnamese has a total of thirty-five different vowel sounds. In addition, Vietnamese is considered a monosyllabic tone language; thus the majority of words are single syllable. Furthermore, the syllabic structure of words governs which consonants may occur at the beginning or end of a syllable (Sato, 1984). For example, in a closed syllable, only the consonant sounds /p, t, k, m, n, and ng/ are found in the final position (Tang, 2006). Approximately eighty percent of Vietnamese words end with a consonant (Sato, 1984).

According to language transfer theory (Gass, 1988), letters, sounds, and their subsequent placement within words are more readily acquired when they are shared between the native and second language. Therefore, unique letters, sounds, and their placement within words that are only found in English may prove more challenging for ELL students to master. Depending on the frequency that these unique letters, sounds, and placements are included on an assessment, students in different language groups may perform differently on items. This is particularly true if shared aspects of one language are more thoroughly represented than the other.

A final difference between the two languages is in how words are altered to convey meaning. In English and Spanish, suffixes are added to a root word to modify meaning. Many of these suffixes are shared between the two languages (Chen et al., 2011). The addition of suffixes to words in Spanish also signifies that multisyllabic words are not an unusual occurrence. As was previously stated, Vietnamese is a monosyllabic tone language, and as such word meaning is altered by six tones,

rather than suffixes (Sato, 1984; Tang, 2006). As a result, Spanish-speaking students may be more likely than Vietnamese-speaking students to correctly respond to items that include suffixes, as well as multisyllabic words.

Method

Test Material

Each year, the Kansas English Language Proficiency Assessment (KELPA) is administered to approximately 33,000 kindergarten through twelfth-grade ELL students in the state of Kansas who have been identified as having limited English proficiency. The two largest language groups assessed are native speakers of Spanish and Vietnamese. When the KELPA was administered in 2009, approximately 27,000 students (81%) were native Spanish-speakers and roughly 1,100 (3%) were native Vietnamese-speakers (Peyton, Kingston, Skorupski, Glasnapp, & Poggio, 2009).

The KELPA is administered in grade level band assessments. Test forms are consistent within each grade level band. The bands span from kindergarten to first grade, second to third grade, fourth to fifth grade, sixth to eighth grade, and ninth to twelfth grade. As grade level increases, there are fewer students assessed, and also fewer students are classified in the Beginner performance category. In an effort to maximize sample size while obtaining a wide spread of test scores, the second to third grade level band assessment was selected for this study.

All grade level band test forms consist of four subsections that are combined to create an overall composite score: reading, writing, listening, and speaking. The reading subsection of the assessment contains the most items and is self-

administered. As a result, this study examines items included in the reading subsection of form A, which was administered in 2007, 2009, and 2011 and form B, which was administered in 2008, and 2010. This approach of combining two equated forms in a single study allows for a larger item pool and subsequently increases the power of the analysis. Coefficient alpha was calculated for form A following the 2007 administration and for form B following the 2008 administration. Table 1 includes the reliability estimates for the reading subsection of the assessment (Peyton et al., 2009).

Table 1
Reliability Estimates for Reading Subtest (23 items)

	Form	Reliability (α)
2nd grade	A	0.81
2nd grade	B	0.79
3rd grade	A	0.84
3rd grade	B	0.83

Participants

The participants included in this study were native Spanish- and Vietnamese-speaking ELL students. The students all took the second and third grade band assessment between 2007 and 2011. To be eligible to take the KELPA, the students had to be classified as limited English proficient at the time of administration. Table 2 includes a summary of descriptive statistics across the five administrations. The total reflects the number of administrations, rather than unique students, as it is possible some students may have taken the assessment twice: once as a second grade student on one form, and then again as a third grade student on the second form.

Table 2
Number of Administrations 2007 - 2011

	2nd grade	3rd grade	Total
Spanish	16,205	15,445	31,660
Vietnamese	791	695	1,486
Total	16,205	15,445	33,146

Procedure

Forms A and B were compared across five administrations to ensure the items remained consistent. One item included on form A was eliminated from the analysis due to the replacement of one of the distracters prior to the 2011 administration. Additionally, the wording for two items was altered on form A prior to the 2011 administration; the stem of each item was changed from "in the story" to "in the passage." Additionally, the words "or passage" were added to one item included on form B prior to the 2008 administration. Since neither of these changes were large revisions nor did they impact any of the answer choices, the three items were included in the DIF analysis. In total, forty-five items were examined: twenty-two items from form A and twenty-three items from form B.

The researcher examined each item for characteristics that may impact either language group in a positive or negative manner. Item characteristics included a count of the number of Spanish-English cognates, multisyllabic words, and suffixes. In total, thirty-four of the forty-five items, or 76%, contained at least one Spanish-English cognate, forty-two items (93%) contained at least one multisyllabic word, and forty-four items (98%) contained at least one suffix. Additionally, each item was analyzed for phonetic and orthographic representations

unique to English from Vietnamese or Spanish (see Appendix A). The number of English vowel and consonant sounds not found in Vietnamese or Spanish were counted for each item to create the variables Vietnamese_sounds and Spanish_sounds respectively. Table 3 includes descriptive information for each item characteristic variable.

Table 3
Item Characteristic Descriptives

	Minimum	Maximum	Mean	Standard deviation
Cognates	0	4	1.11	1.05
Multisyllabic	0	20	7.56	4.10
Suffixes	0	14	4.44	2.89
Vietnamese_sounds	8	134	40.80	25.26
Spanish_sounds	4	115	28.38	24.12

Analyses

Preliminary analyses were conducted to examine overall performance on the reading subtest prior to examining the individual items for DIF. Table 4 includes means and standard deviations for total reading score by language group for forms A and B. An independent samples *t* test was conducted to determine if language groups had significantly different total reading scores. Equal variance was not assumed, $t(1422.005) = 14.900$, $p < .001$, indicating Vietnamese-speaking students had significantly higher total scores than Spanish-speaking students on the reading subtest.

Table 4
Raw Score Means and Standard Deviations by Group (N=23)

	Spanish	Vietnamese
Form A total score		
Mean	13.57	15.51
(SD)	(5.11)	(5.03)
Form B total score		
Mean	13.47	15.66
(SD)	(4.85)	(4.59)
Combined form total scores		
Mean	13.52	15.59
(SD)	(4.98)	(4.81)

Logistic regression was used to analyze each of the forty-five items to determine the extent of differential performance due to language group, after accounting for proficiency. The logistic regression equation for each item included a proficiency variable comprised of an individual's total score on the reading subtest of the KELPA,, and a variable that indicated group membership, coded zero for Vietnamese-speaking students and one for Spanish-speaking students. An interaction term was included to determine the extent of nonuniform DIF that existed for each item. Each term was entered hierarchically into the equation, and the subsequent change in the Nagelkerke pseudo R^2 measure of effect size was recorded. Following the logistic regression analysis, effect sizes obtained from adding in the group and interaction terms were combined to form a total effect size variable. Item characteristics, found in Table 3, were then correlated with effect size to determine if a relationship existed. Significantly correlated item characteristics were included in a stepwise multiple regression predicting effect size.

Results

Results from the logistic regression analyses indicated that for 23 of the 45 items the language variable had a statistically significant impact after controlling for proficiency. These findings indicate that just over half the items on the reading subtest function differently based on language group (see Appendix B and C). Eleven of the items flagged for DIF favored Spanish-speaking students, while twelve items favored Vietnamese-speaking students. In addition, eleven of the forty-five items contained a significant interaction term after controlling for proficiency and language group, indicating nonuniform DIF was present. The combined increase in effect size after controlling for proficiency spanned from no change to a .003 increase. However, these marginal increases did not always accurately reflect which items contained significant language and interaction terms.

To further examine the nature of the significant interactions, a graph was constructed for each using Microsoft Excel (see Appendix D). Of the eleven total items containing interactions between proficiency and language group, ten were ordinal interactions, indicating the group predicted most likely to respond to the item correctly remained constant across all levels of proficiency. For ten of the interactions, Vietnamese-speaking students were predicted to be more likely to respond correctly to the item. One item, Item 5 on Form A, contained a disordinal interaction, in which the group predicted to most likely respond to the item correctly changed from Spanish-speaking at the low end of the proficiency scale to Vietnamese-speaking as the level of proficiency increased. However, the difference between the groups at the low end was miniscule.

The interaction graphs revealed that for ten of the eleven significant interactions, even high performing Spanish-speaking students were not very likely to correctly respond to the item. For six of these items, the probability was .5 or less. Overall, these graphs indicate that Vietnamese students become increasingly more likely to correctly respond to an item as proficiency level increases. At low proficiency levels, language group does not appear to have a very large impact, presumably because students are not very likely to correctly respond to the item regardless of their native-language group.

To determine if item characteristics contributed to the underlying DIF, zero-order correlations were obtained between effect size and each item characteristic. The results are displayed in Table 5. Of the five item characteristics examined, only three variables, Spanish_sounds, Vietnamese_sounds, and multisyllabic, were significantly correlated with effect size. These three variables were also highly correlated with each other. As a result, when placed in a stepwise multiple regression, only Vietnamese_sounds was included in the model due to it being the variable most highly correlated with effect size, $R^2 = .160$, $F(1, 43) = 8.162$, $p = .007$.

Table 5
Zero Order Correlations Between Item Characteristics (N= 45)

	Effect_size	Multisyllabic	Spanish_sounds	Vietnamese_sounds	Cognates
Multisyllabic	-.35*				
Spanish_sounds	-.34*	.72**			
Vietnamese_sounds	-.40**	.74**	.95**		
Cognates	-.27	.44**	.23	.24	
Suffixes	-.27	.85**	.66**	.66**	.38**

* $p < .05$. ** $p < .01$.

Upon revisiting the actual items flagged for DIF, it was discovered that all three rhyming items included on the reading subtest (items 6 & 7 on form A, and item 6 on form B) favored Vietnamese-speaking students. This may be due to the words being single-syllable words, or Vietnamese-speaking students may have been able to detect the discrepancy between the vowel sounds better than Spanish-speaking students due to the differences in their native language. However, this was the only similarity evident when examining items flagged for DIF. Additionally, distracter analysis was conducted comparing percent selected by language group. A pattern was not evident that would explain why one language group was more likely to correctly respond to an item.

Since only one variable was included in the stepwise multiple regression, a post hoc analysis was conducted to determine if the same item characteristic variables would predict the occurrence of uniform DIF, rather than the magnitude, using a logistic regression method. Items containing a significant language term were coded as one, while items that did not contain a significant language term were coded as zero. The variables Vietnamese_sounds, Spanish_sounds, and Multisyllabic were entered in a logistic regression to determine how well item characteristics predicted the occurrence of DIF on an item. The model was statistically significant, χ^2 (1, N = 45) = 15.011, p = .002, the Nagelkerke pseudo R^2 = .378. When examining the classification rates, the Block 0 classification table placed all cases in the no DIF category, achieving an overall accuracy of 51.1%. This indicates the model with no predictors was accurate roughly half the time. The inclusion of the item characteristic variables resulted in a more accurate classification of DIF and no DIF,

with an overall percentage of 77.8. Only Vietnamese_sounds was a significant predictor of DIF status (p = .009, Exp(B) = .802). Interestingly, a subsequent model that included only Vietnamese_sounds and Spanish_sounds was able to correctly predict the occurrence of uniform DIF 82.2% of the time, with both Spanish_sounds (p = .047, Exp(B) = 1.169) and Vietnamese_sounds (p = .009, Exp(B) = .825) significantly predicting the occurrence of uniform DIF. A similar analysis predicting the occurrence of nonuniform DIF revealed that none of the three item characteristics was a significant predictor. These findings may indicate that there are separate sources that contribute to the occurrence of uniform and nonuniform DIF on English language proficiency items.

Discussion

The present study advances the literature in that five years of data across two equated test forms were aggregated in order to assess the hypothesis that items on an English language proficiency assessment perform differently based on language group. The results of the study indicate that 51% of the items were found to function differently when comparing native Spanish- and Vietnamese-speaking students. These findings reiterate the importance of conducting DIF studies for English language proficiency assessments, particularly when the results have a profound impact on students, teachers, schools, and even test developers.

One especially interesting finding from this study was the presence of interaction effects on 24% of the items. One possible conclusion is these flagged items include construct irrelevant variance. This may be particularly true for Spanish-speaking students, since even the high performing students were unlikely

to answer many of these items correctly. The inclusion of these ten items that were found to increasingly favor Vietnamese students makes it all the more likely for high performing Vietnamese-speaking students to be correctly labeled as "proficient" while Spanish-speaking students of equivalent proficiency are less likely to be classified as such.

This study also evaluated the hypothesis that the extent of DIF detected could be attributed to characteristics of individual items. The results from this portion of the study revealed several unexpected outcomes. First, only three of the five characteristics were significantly correlated with DIF effect size. Additionally, each of the three correlations was negative. However, due to the small magnitude of DIF detected, additional research is necessary in this area to determine the precise source of DIF for these language groups.

Another important finding of the present study concerns the two variables that were not significantly correlated with effect size: cognates and suffixes. Previous research had found each of these variables to significantly impact language groups; however the current study indicates when sporadically included on a second and third grade band assessment, cognates and suffixes do not seem to contribute to the magnitude of DIF for these language groups. This is an especially important finding in that each of these variables was anticipated to benefit the Spanish-speaking group. The data did not support this hypothesis, especially when considering only one item with an interaction between proficiency and language group was found to favor Spanish-speaking students. Future research may be

needed to examine additional grade bands for similar findings, especially given that the magnitude of the effect size was so small.

A final important implication of this study concerns the use of equivalent forms. Despite being equated, sixteen of the twenty-three items on form B contained evidence of uniform or nonuniform DIF as compared to only eleven items on form A. Most items on form B contained either uniform or nonuniform DIF, but not both, as opposed to items on form A that were found to contain significant terms for nonuniform DIF after accounting for uniform DIF. This evidence reiterates the importance for examining items for DIF when using multiple test forms purporting to measure a single construct.

There were several limitations to this study. First, after controlling for proficiency, effect sizes did not dramatically increase as a result of including the group variable or the interaction term. As a result, the follow-up correlations and multiple regression analysis contained a variable that lacked variability. In addition, item characteristics were all positively skewed. Furthermore, the number of items included in the analysis was small. Future research in this area would benefit from including a larger item pool to increase the power of follow-up analyses when determining the sources of DIF.

There are a number of ways to expand on the current study in addition to increasing the item pool and examining additional grade bands. Items analysis could extend beyond the reading domain. Impact on other language groups could be examined to determine the implications of analyzing DIF for multiple subgroups. Future analyses could further examine the impact of item characteristics, including

the degree of vocabulary difficulty or the cognitive complexity of the items. An external proficiency measure or scale purification method could be used to determine if DIF detection would remain the same. In addition, the present literature on DIF studies contains a significant number of articles comparing logistic regression with the Mantel-Haenszel method; however, few articles compare logistic regression with the 3-parameter item response theory model. Since the 3-parameter model includes both a lower asymptote and discrimination factor, these variables may explain some of the DIF that was detected in the current study using logistic regression.

The ultimate goal of this study was to determine the extent that an English language proficiency assessment currently in use validly assessed students from two language groups of the testing population. Now that items have been determined to contain DIF, proper steps may be taken to address the issue. Ultimately, these results stand to improve the assessment to better allow stakeholders to make accurate decisions regarding what students know and can do.

References

Abbott, M. L. (2007). A confirmatory approach to differential item functioning on an ESL reading assessment. *Language Testing*, *24*(1), 7-36.

Allalouf, A., & Abramzon, A. (2008). Constructing better second language assessments based on differential item functioning analysis. *Language Assessment Quarterly*, *5*(2), 120-141.

Allen, J., & Le, H. (2008). An additional measure of overall effect size for logistic regression models. *Journal of Educational and Behavioral Statistics, 33*(4), 416-441.

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. Holland and H. Wainer (Eds.), *Differential item functioning*, (pp. 3-24). Hillsdale, NJ: Lawrence Erlbaum Associates.

ATB Test Administration Manual for Ability-to-Benefit Testing with COMPASS/ESL Internet Version. (2011). Iowa City, IA: ACT, Inc.

Barlow, J. A. (2005). Phonological change and the representation of consonant clusters in Spanish: A case study. *Clinical Linguistics and Phonetics 19*(8), 659-679.

Bowen, D., & Joldersma, K. (2011). The effects of controlling for distributional differences on the Mantel-Haenszel statistic. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Chen, Z., & Henning, G. (1985). Linguistic and cultural bias in language proficiency tests. *Language Testing, 2*(2): 155-163.

Chen, X., Ramirez, G., Luo, Y. C., Geva, E., & Ku, Y. M. (2011). Comparing vocabulary development in Spanish- and Chinese-speaking ELLs: The effects of metalinguistic and sociocultural factors. *Reading and Writing*, *24,* 1-30.

Clauser, B. E., Nungester, R. J., & Swaminathan, H. (1996). Improving the matching for DIF analysis by conditioning on both test score and an educational background variable. *Journal of Educational Measurement*, *33*(4), 453-464.

Coe, N. (2001). Speakers of Spanish and Catalan. In M. Swan & B. Smith (Eds.), *Learner English: A teacher's guide to interference and other problems.* (pp. 90-112). Ernst Klett Sprachen.

Cole, N. S. (1993). History and development of DIF. In P. Holland and H. Wainer (Eds.), *Differential item functioning*, (pp. 25-30). Hillsdale, NJ: Lawrence Erlbaum Associates.

Colorado Department of Education. (2010). *Colorado English language acquisition assessment program technical report.* Monterey, CA: CTB/McGraw-Hill, LLC.

Cunningham, T. H., & Graham, C. (2000). Increasing native English vocabulary recognition through Spanish immersion: Cognate transfer from foreign to first language. *Journal of Educational Psychology*, *92*(1), 37-49.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. Holland and H. Wainer (Eds.), *Differential item functioning*, (pp. 35-66). Hillsdale, NJ: Lawrence Erlbaum Associates.

Fashola, O. S., Drum, P. A., Mayer, R. E., & Kang, S. J. (1996). A cognitive theory of orthographic transitioning: Predictable errors in how Spanish-speaking

children spell English words. *American Educational Research Journal*, *33*(4), 825-843.

Ferne, T. & Rupp, A. A. (2007). A synthesis of 15 years of research on DIF in language testing: Methodological advances, challenges, and recommendations. *Language Assessment Quarterly, 42*(2) 113-148.

Gass, S. M. (1988). Second language acquisition and linguistic theory: The role of language transfer. In S. Flynn & W. O'Neil (Eds.), *Linguistic theory in second language acquisition* (pp. 384-403). Dordrecht: Kluwer Academic Publishers.

Hidalgo, M. D., & López-Pina, J. É. A. (2004). Differential item functioning detection and effect size: A comparison between logistic regression and Mantel-Haenszel procedures. *Educational and Psychological Measurement*, *64*(6), 903-915.

Hegde, M. N., & Pomaville, F. (2008). *Assessment of communication disorders in children: Resources and protocols*. San Diego, CA: Plural Publishing Inc.

Jang, E. E., & Roussos, L. (2009). Integrative analytic approach to detecting and interpreting L2 vocabulary DIF. *International Journal of Testing, 9,* 238-259.

Kopriva, R. J. (2008). *Improving testing for English language learners*. New York, NY: Routledge.

LanguEdge Courseware Score Interpretation Guide. (2002). Princeton, NJ: Educational Testing Service.

Lara, J., Ferrara, S., Calliope, M., Sewell, D., Winter, P., Kopriva, R., Bunch, M., et al. (2007). The English language development assessment. In J. Abedi (Ed.),

English language proficiency assessment in the nation: Current trends and future practice (pp. 47-62). Davis: University of California.

Linn, R. L. (1993). The use of differential item functioning statistics: A discussion of current practice and future implications. In P. Holland and H. Wainer (Eds.), *Differential item functioning*, (pp. 349-364). Hillsdale, NJ: Lawrence Erlbaum Associates.

MacGregor, D., Louguit, M., Yanosky, T., Grim Fidelman, C., Pan, M., Huang, X., & Kenyon, D. M. (2010). *Annual Technical Report for ACCESS for ELLs English Language Proficiency Test* (Technical Report No. 5). Washington, DC: Center for Applied Linguistics.

Mahoney, K. (2008). Linguistic influences on differential item functioning for second language learners on the national assessment of educational progress. *International Journal of Testing, 8*(1), 14.

Martiniello, M. (2007). Linguistic complexity and differential item functioning (DIF) for English language learners (ELL) in math word problems. (Doctoral dissertation, Harvard University). Retrieved from http://gradworks.umi.com

Mazor, K. M., Kanjee, A., & Clauser, B. E. (1995). Using logistic regression and the Mantel-Haenszel with multiple ability estimates to detect differential item functioning. *Journal of Educational Measurement*, *32*(2), 131-144.

Nagy, W. E., Garcia, G. E., Durgunoglu, A. Y., & Hancin-Bhatt, B. (1993). Spanish-English bilingual students' use of cognates in English reading. *Journal of Literacy Research*, *25*(3), 241-259.

National Clearinghouse for English Language Acquisition. (2011). *The growing number English learner students 1998-99 – 2008-09*. Washington, DC.

No Child Left Behind Act of 2001. Public Law. 107-110, 115 Stat. 1425.

Peng, C., J., Lee, K. L., Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *Journal of Educational Research, 96*(1), 3-14.

Peyton, V., Kingston, N.M, Skorupski, W., Glasnapp, D., & Poggio, J. (2009). *Kansas English language proficiency assessment technical manual.* Center for Educational Testing and Evaluation, University of Kansas.

Rogers, H. J., & Swaminathan, H. (1993). A comparison of logistic regression and Mantel-Haenszel procedures for detecting differential item functioning. *Applied Psychological Measurement*, *17*(2), 105-116.

Sasaki, M. (1991). A comparison of two methods for detecting differential item functioning in an ESL placement test. *Language Testing 8*(2), 95-111.

Sato, C. J. (1984). Phonological processes in second language acquisition: Another look at interlanguage syllable structure. *Language Learning*, *34*(4), 43-58.

Scheuneman, J. D., & Grima, A. (1997). Characteristics of quantitative word items associated with differential performance for female and black examinees. *Applied Measurement in Education, 10*(4), 299-319.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement*, *27*(4), 361-370.

Swanson, D. B., Clauser, B. E., Case, S. M., Nungester, R. J., & Featherman, C. (2002). Analysis of differential item functioning (DIF) using hierarchical logistic regression models. *Journal of Educational and Behavioral Statistics*, *27*(1), 53.

Tang, G. (2006). Cross-linguistic analysis of Vietnamese and English with implications for Vietnamese language acquisition and maintenance in the United States. *Journal of Southeast Asian-American Education & Advancement*, *2*, 1-33.

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores*. Ottawa, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Appendix A

English Only Orthographic and Phonetic Representations Included in the Reading Section of the KELPA

Not found in Spanish		Not found in Vietnamese	
Vowel sounds:	a (about)*	Vowel sounds:	i (did)
	a (cap)		u (under)
	a (take)		er (her)
	i (did)		o (go)
	oo (book)		ow (how)
	u (under)		i (like)
	ir (birth)		oi (moist)
	er (her)		a (take)
	e (give)		a (about)
Consonant sounds:	v (view)	Syllable initial sounds:	g (good)
	y (year)		t (tiny)
	z (zoo)		th (thunder)
	sh (shin)		th (the)
	th (the)		sh (shin)
	th (thunder)		ch (check)
			j (join)
Initial sounds:	h (have)		w (word)
	j (join)		sn (snow)
	sn (snow)		st (store)
	st (store)		br (brought)
	sw (switch)		dr (dream)
	qu (quiet)		cr (cream)
			gr (grind)
Final sounds:	g (dog)		cl (cloud)
	k (suck)		fl (flies)
	m (them)		sw (switch)
	p (wrap)		qu (quiet)
	t (not)		
	kt (walked)	Syllable final sounds:	d (odd)
	ks (trucks)		g (dog)
	ft (swift)		th (with)
	ld (build)		f (off)
	mp (jump)		v (have)
	nt (sent)		s (case)

Final sounds cont:	nd (find)	Syllable final sounds cont:	z (is)
	nz (means)		ch (which)
	nk (think)		j (huge)
	skt (asked)		l (will)
	lpt (helped)		kt (walked)
	mpt (jumped)		sh (fresh)
			ks (trucks)
* unemphasized			ft (swift)
			ld (build)
			mp (jump)
			nt (sent)
			nd (find)
			nz (means)
			nk (think)
			skt (asked)
			lpt (helped)
			mpt (jumped)

Appendix B

Form A Logistic Regression Results

	Effect Size			e^B		
Item	Proficiency	Language	Interaction	Proficiency	Language	Interaction
12	.4195***	.0035***	.0000	1.3948	2.3525	.9695
7	.3400***	.0016***	.0012***	1.4316	.6697	.8948
5	.2077***	.0017***	.0010***	1.1326	1.5827	1.0742
9	.4006***	.0024***	.0001	1.3853	.4608	.9633
6	.3349***	.0010***	.0003**	1.3557	.6771	.9417
8	.3463***	.0007***	.0003*	1.3582	1.5768	.9511
23	.1615***	.000	.0008***	1.2611	1.1939	.9322
16	.3375***	.0007***	.0001	1.3092	1.4188	.9903
1	.2445***	.0007**	.0000	1.2474	.7098	1.0290
20	.2454***	.0005**	.0001	1.2521	.7615	.9746
10	.4107***	.0004**	.0001	1.4023	.7019	.9654
4	.3762***	.0000	.0002	1.2867	.9447	1.0353
13	.2167***	.0000	.0002	1.2458	1.0875	.9694
3	.4748***	.0001	.0001	1.4763	1.1715	.9850
18	.3082***	.0001	.0001	1.3072	1.1236	.9634
19	.2630***	.0001	.0001	1.2598	1.1936	.9771
2	.3108***	.0001	.0000	1.2643	.8854	1.0180
11	.4342***	.0000	.0001	1.4180	1.0322	.9661
17	.4028***	.0000	.0001	1.4139	.8983	.9593
21	.2725***	.0000	.0001	1.2597	1.0608	.9821
22	.2982***	.0001	.0000	1.2793	.9317	.9799
15	.3727***	.0000	.0000	1.4432	.9728	.9991

* $p < .05$. ** $p < .01$. ***$p < .001$.

0 = Vietnamese
1 = Spanish

Appendix C

Form B Logistic Regression Results

	Effect Size			e^B		
Item	Proficiency	Language	Interaction	Proficiency	Language	Interaction
11	.2908***	.0011*	.0016***	1.4461	.7398	.8703
7	.1810***	.0020***	.0000	1.2179	.5593	.9828
8	.3547***	.0016***	.0002	1.3958	.5847	.9409
16	.3462***	.0016***	.0000	1.3684	1.7138	.9879
12	.2514***	.0000**	.0016***	1.4351	1.5459	.8758
6	.2589***	.0008**	.0005*	1.3286	.7214	.9341
13	.3451***	.0008**	.0002	1.3678	1.5435	.9574
20	.2730***	.0008**	.0001	1.2169	1.3147	1.0334
14	.1379***	.0008**	.0000	1.1890	.7000	.9874
19	.2955***	.0008**	.0000	1.2662	1.3717	1.0031
10	.3132***	.0000	.0006*	1.3809	1.0319	.9259
1	.2439***	.0006*	.0000	1.3272	.6360	1.0007
21	.3021***	.0002*	.0004*	1.3555	1.3332	.9387
18	.2869***	.0005*	.0000	1.2318	1.2190	1.0263
15	.2739***	.0005*	.0001	1.2310	1.3101	1.0309
9	.2494***	.0001	.0004*	1.3041	1.2138	.9449
2	.3197***	.0003	.0000	1.3400	.7471	.9985
22	.2513***	.0000	.0003	1.2920	.9912	.9544
5	.3193***	.0002	.0000	1.3161	.8148	.9870
23	.1607***	.0002	.0000	1.1763	1.1485	.9971
3	.4497***	.0001	.0000	1.4427	.8353	.9889
4	.3487***	.0001	.0000	1.3343	1.0656	1.0006
17	.1865***	.0000	.0001	1.2140	1.0238	.9861

* $p < .05$. ** $p < .01$. ***$p < .001$.

0 = Vietnamese

1 = Spanish

Appendix D

Interaction Graphs

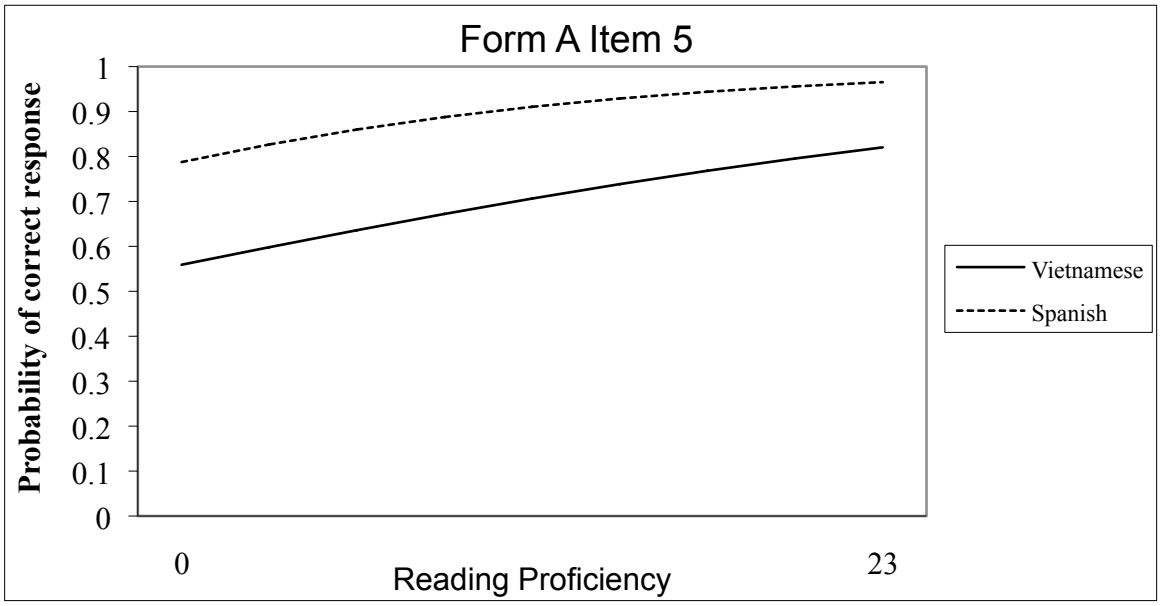

Figure 1: Graph of interaction between language and proficiency for form A item 5.

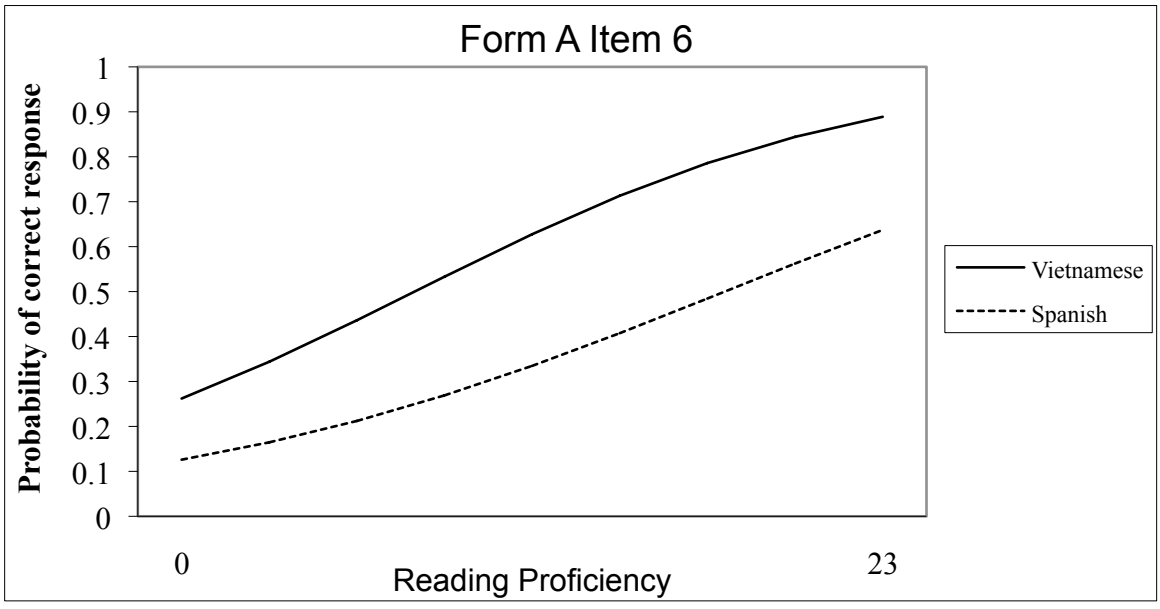

Figure 2: Graph of interaction between language and proficiency for form A item 6.

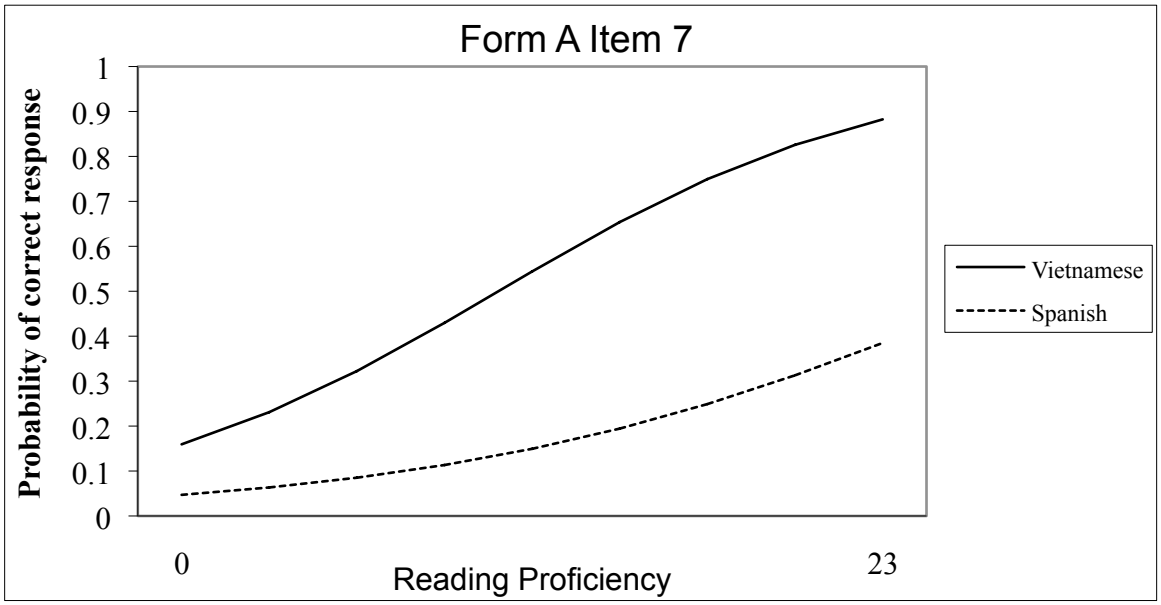

Figure 3: Graph of interaction between language and proficiency for form A item 7.

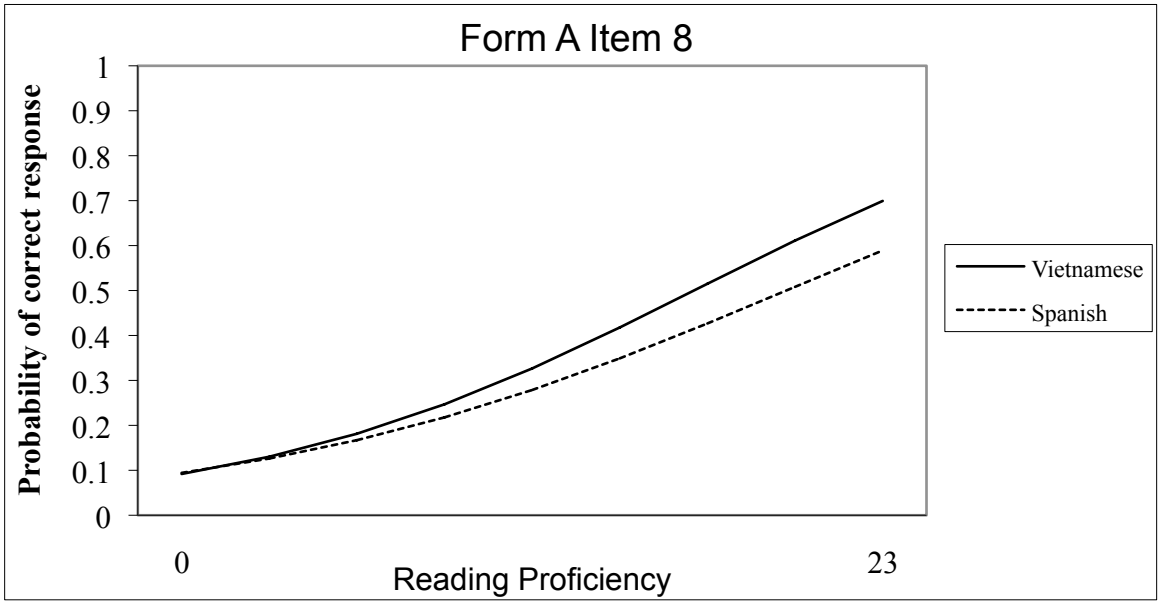

Figure 4: Graph of interaction between language and proficiency for form A item 8.

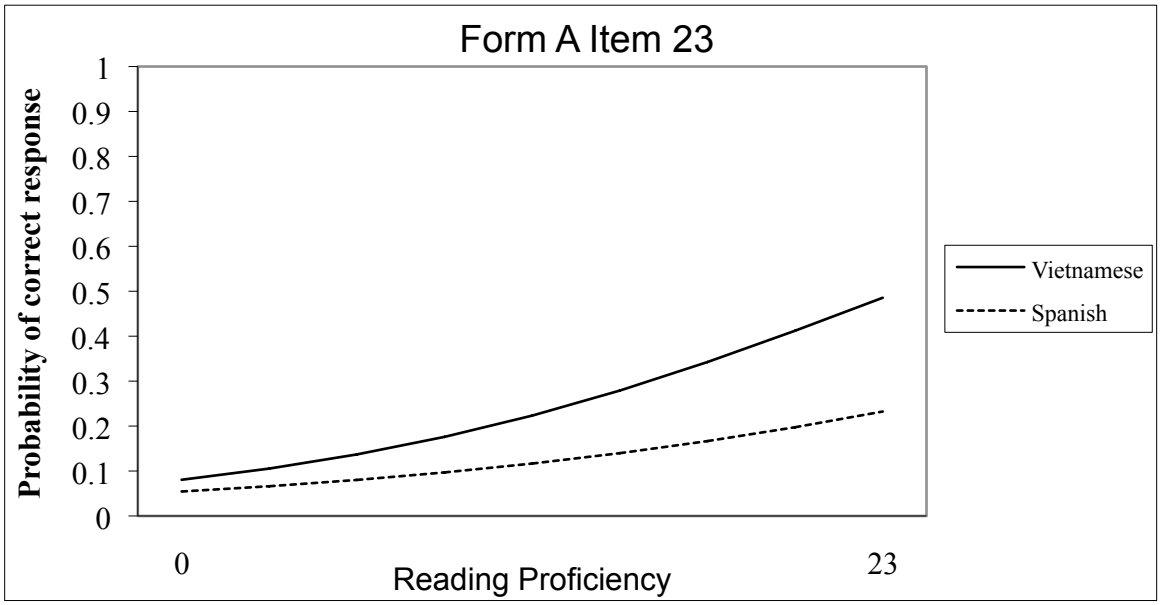

Figure 5: Graph of interaction between language and proficiency for form A item 23.

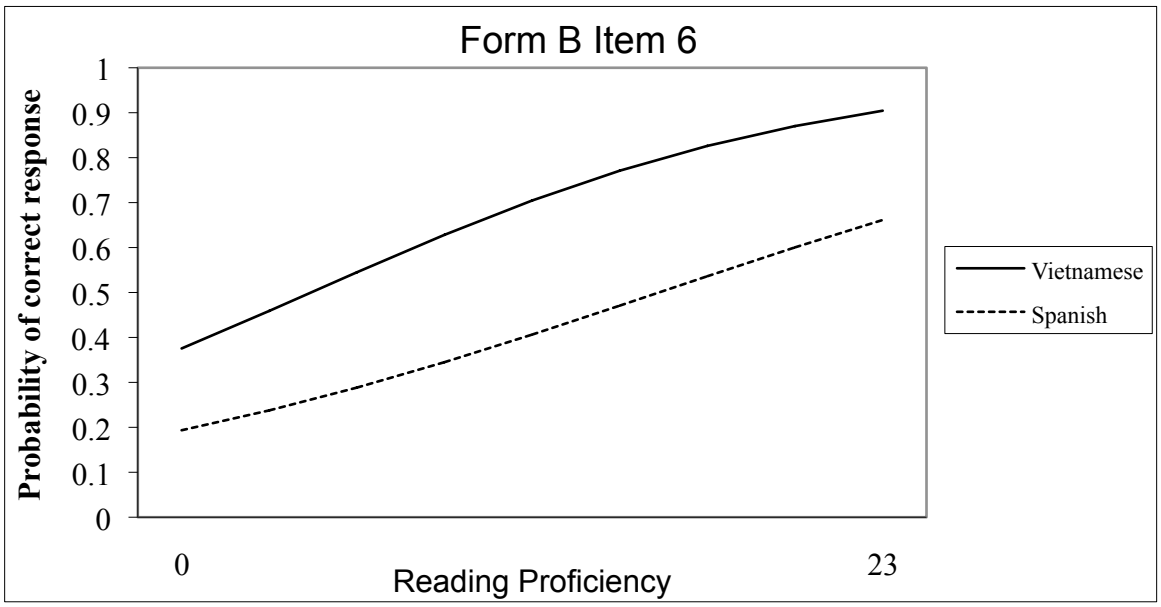

Figure 6: Graph of interaction between language and proficiency for form B item 6.

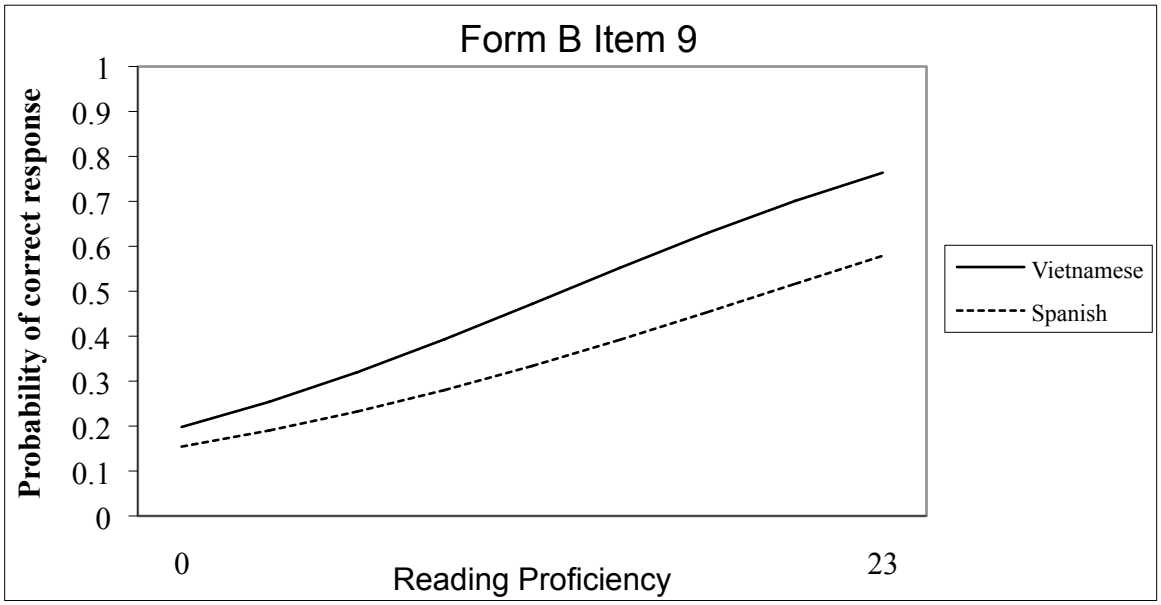

Figure 7: Graph of interaction between language and proficiency for form B item 6.

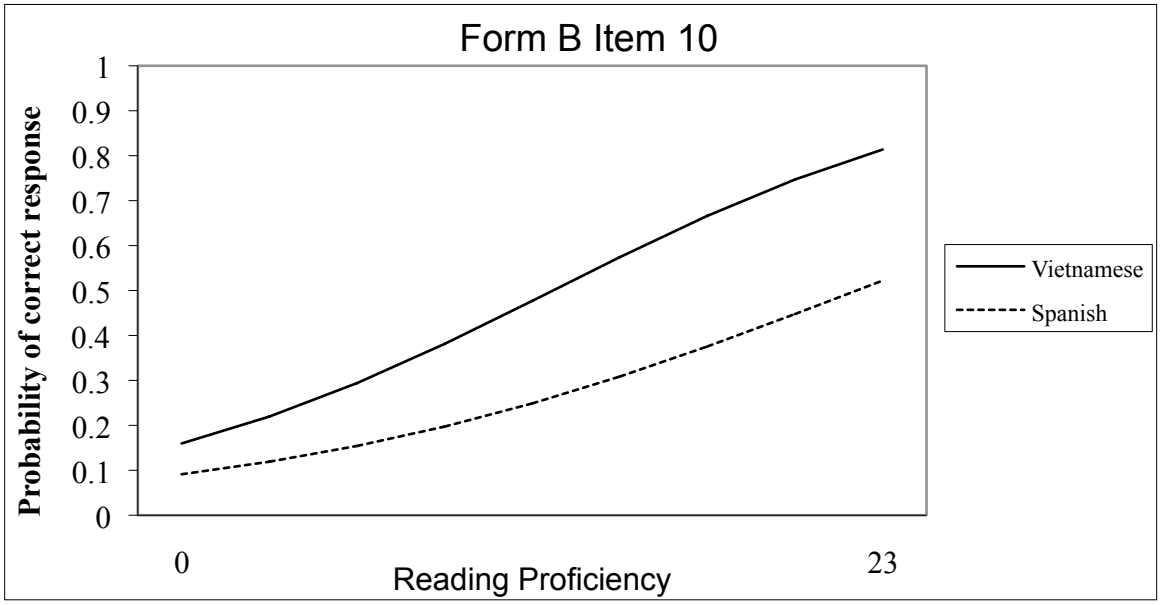

Figure 8: Graph of interaction between language and proficiency for form B item 10.

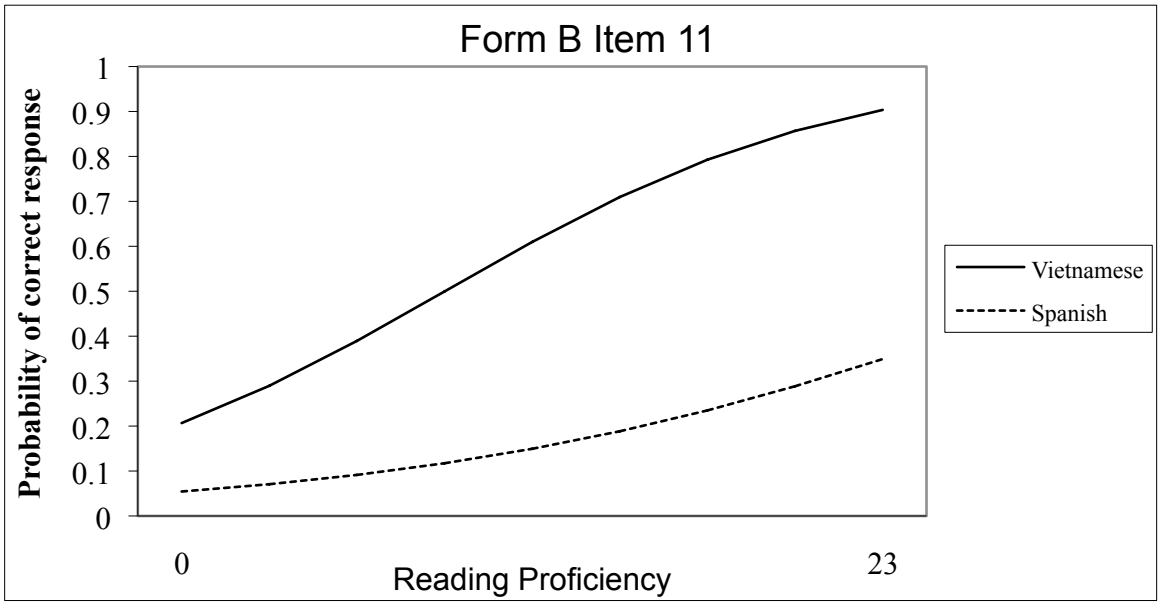

Figure 9: Graph of interaction between language and proficiency for form B item 11.

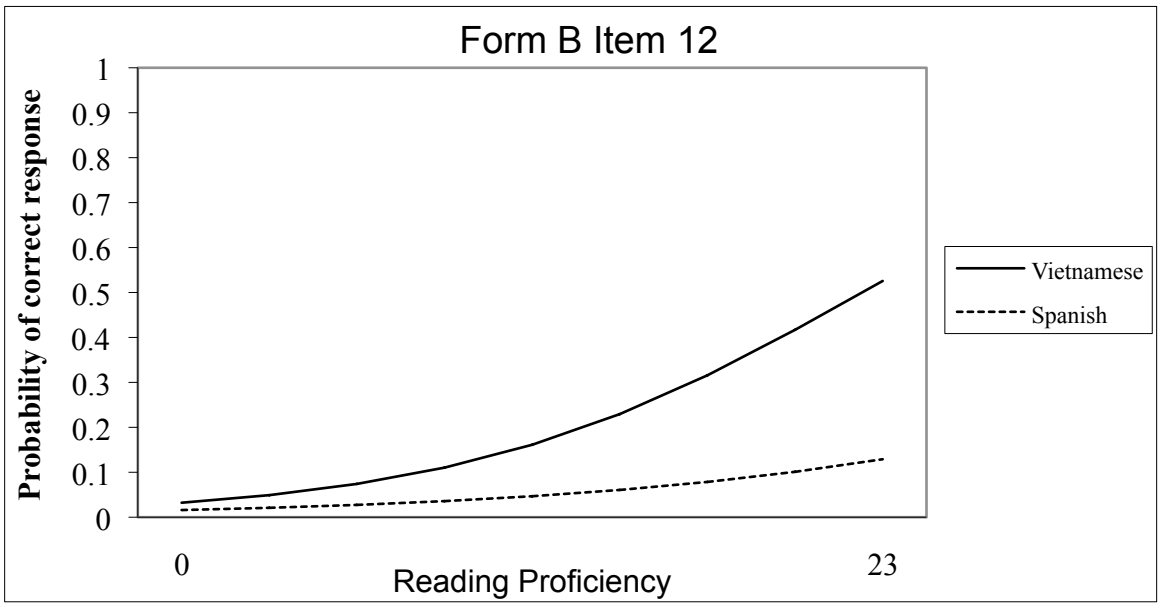

Figure 10: Graph of interaction between language and proficiency for form B item 12.

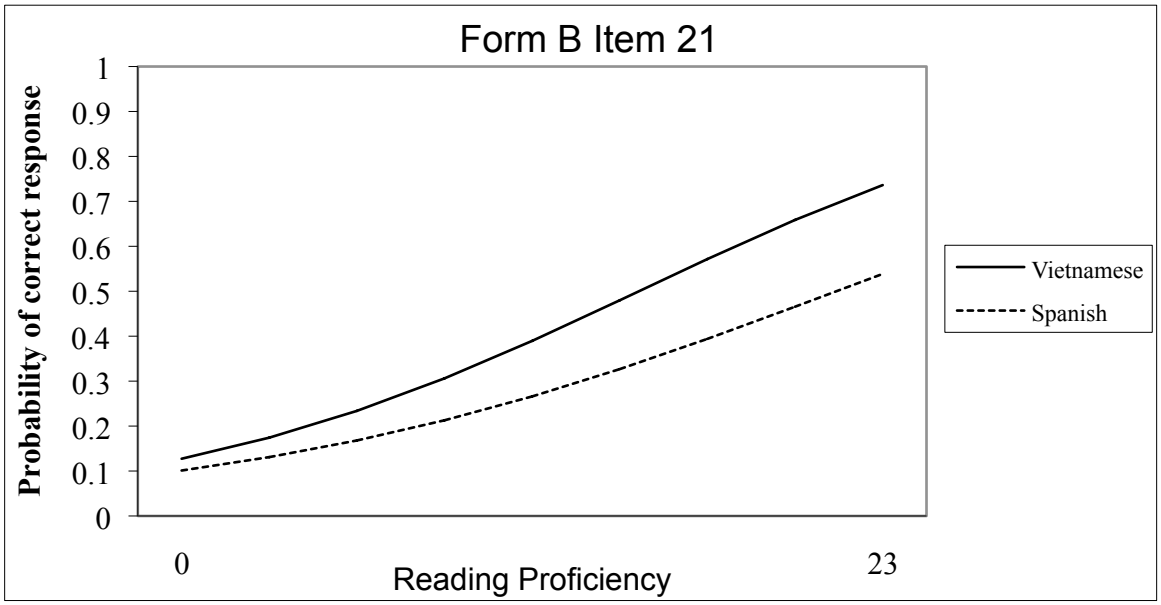

Figure 11: Graph of interaction between language and proficiency for form B item 21.