

Poncheri, R. M., Meade, A. W. & Surface, E. A. (2007, April). *Differential item functioning and personality: Comparing native and non-native speakers*. Paper presented at the 22st annual conference of the Society for Industrial and Organizational Psychology, New York, NY.

Differential Item Functioning between Native and Non-Native English Speakers on the International Personality Item Pool

Reanna M. Poncheri
North Carolina State University
Surface, Ward, & Associates

Adam W. Meade
North Carolina State University

Eric A. Surface
Surface, Ward, & Associates
North Carolina State University



APRIL 2007

APPROVED FOR PUBLIC RELEASE;
DISTRIBUTION UNLIMITED

Copyright Notice

This document and its content is copyright ©1997-2010 of SWA Consulting Inc. All rights reserved.

Any redistribution or reproduction of part, or the entire document in any form is prohibited except for: (1) you may print or download to a local hard disk extracts for your personal and non-commercial use only, and (2) you may copy the content to individual third parties for their personal use, but only if you acknowledge the website and author(s) as the source of the material. You may not, except with our express written permission, distribute or commercially exploit the content, nor may you transmit it or store it on any other website or other form of electronic retrieval system.

Differential Item Functioning between Native and Non-Native English Speakers on the International Personality Item Pool

Reanna M. Poncheri
NC State University / Surface, Ward, & Associates

Adam W. Meade
NC State University

Eric A. Surface
Surface, Ward, & Associates / NC State University

With increasing globalization, organizations have an increased need for assessments administered to employees with diverse cultural and lingual backgrounds. These assessments could be cross-national assessments within a multi-national organization, or the administration of a measure to a diverse population at a single organization or location within a country. In such cases, organizations often must choose whether to attempt to translate the measure to the native language(s) of the respondents. For practical reasons, many organizations choose to administer measures in the predominant language of the organization. This study examined the differential functioning (measurement invariance) of an English language version of the International Personality Item Pool, a commonly used assessment of the Five Factor Model (FFM) of personality, across native and non-native English speaking workers living within the United States. Results indicated that many items of the FFM scales did not exhibit measurement invariance across native and non-native English speakers.

One of the most important changes shaping today's work environment is increased globalization (Dahlin, Weingart, & Hinds, 2005; Earley & Gibson, 2002; Speizer, 2006). Increased interdependence among nations' economies has led to many important changes in the way that organizations are structured and the way business is conducted. Multinational organizations continue to expand by setting up operations in foreign countries, and increasingly rely on outsourcing and offshoring. Moreover, work teams within the United States are becoming more nationally diverse, bringing together team members from different cultural and lingual backgrounds (Burke & Ng, 2006; Dahlin et al., 2005). In order for organizations to survive and thrive in an international economy, it is necessary to overcome language and cultural barriers. For example, organizations which operate in more than one country face challenges associated with selecting, training,

and managing members of a workforce who may have different cultural backgrounds and who may speak different languages (Speizer, 2006). In such environments, organizational decision-makers relying on organizational assessments must choose whether to translate assessments to the respondents' native language or administer them in a common language to all respondents. In multilingual organizations where English is the common language of business, it may not be feasible to translate instruments into every native language represented in the workforce. In such cases, organizations may be forced to administer an English language version to all employees. However, the ramifications of administering English language assessments to non-native speakers are largely unknown.

This study uses an Item Response Theory (IRT) approach to explore the suitability of administering an English-version of a widely used personality measure, Goldberg's (1999)

International Personality Item Pool (IPIP), to native and non-native English speakers. Specifically, we examine differential item functioning (DIF) of IPIP scales for native and non-native English speaking workers. If the IPIP is found to function equivalently across native and non-native English speaking groups, the choice of not translating the IPIP for use in diverse samples is supported. If, however, the IPIP does not function equivalently across native and non-native English speaking groups, use of the un-translated measure in diverse samples must be approached cautiously.

Personality Inventories and Cross-Cultural Research

Personality measures are extremely useful in organizational research and practice because personality traits, such as conscientiousness, have been linked to important organizational outcomes including overall job performance, training performance, organizational citizenship behavior (or contextual performance), and counterproductive work behavior (Barrick, Mount, & Judge, 2001; Hattrup, O'Connell, & Wingate, 1998; Sackett & DeVore, 2002). The predictive power associated with personality measures, coupled with their widespread availability, has led to an increase in their use. Rothstein and Goffin (2006) indicate that approximately 20-40% of US companies currently use personality measures for selection and that this percentage is increasing. The use of personality inventories is also increasing globally and, as a result, personality measures have been of particular interest in cross-cultural research (Salgado, Viswesveran, & Ones, 2001).

The five-factor model (FFM) of personality is perhaps the most widely accepted model of personality structure (McCrae & Costa, 1987) and is frequently studied and used in organizational research and practice (Barrick et al., 2001; Block, 1995; Hough & Schneider, 1996; Rothstein & Goffin, 2006). While not without its critics (e.g., Block, 1995; McAdams, 1992), the popularity of the FFM can, in part, be attributed to findings that the FFM has been shown to meet the criteria for a reasonably good taxonomy (Hough & Schneider, 1996), and that the criterion-related validity of the FFM has been demonstrated across many job contexts

(Barrick & Mount, 1991; Barrick et al., 2001; Salgado, 1997; Tett, Jackson, & Rothstein, 1991).

The five dimensions of the FFM, also referred to as the 'Big Five'¹ are: Extraversion (e.g., sociability, dominance, ambition, positive emotionality, and excitement-seeking), agreeableness (e.g., cooperation, trustfulness, compliance, and affability), conscientiousness (e.g., dependability, achievement striving, and planfulness), emotional stability (e.g., lack of anxiety, hostility, depression, and personality insecurity), and openness to experience (e.g., intellectance, creativity, unconventionality, and broad-mindedness; Barrick et al., 2001; Costa & McCrae, 1992; Goldberg, 1990, 1992). While several measures of the FFM exist (e.g., NEO Personality Inventory; Costa & McCrae, 1992), Goldberg's (1999) IPIP measure of the FFM has recently seen an emergence in organizational research. The IPIP is a public-domain personality measure developed specifically to encourage personality research and to provide a freely available alternative to commercial instruments. This effort appears to have been successful, as the IPIP has been widely used in organizational research and other research domains to assess personality (e.g., Bowling, Beehr, & Swader, 2005; Brown, Cober, Kane, Levy, & Shalhoop, 2006; Cucina, Vasilopoulos, & Sehgal, 2005; Rubin, Munz, & Bommer, 2005).

The five factors in the FFM have been shown to be generalizable when assessed in a variety of cultural settings (McCrae & Costa, 1987) as have the relationships between the FFM and organizationally relevant outcomes (see Salgado et al., 2001 for a review). The growing global use of the IPIP in particular is evidenced by the translation of the items into more than 25 languages (Goldberg et al., 2006). As a result of the increased use of the IPIP and other personality inventories, the psychometric properties of these translated instruments have been assessed at the factor- (e.g., Schmit, Kihm, & Robie, 2000) and item-levels (e.g., Ellis,

¹ Technically, the FFM and Big 5 are distinct models of personality developed via different means. In practice, they are highly similar and thus we use those labels interchangeably.

Becker, & Kimmel, 1993; Orlando & Marshall, 2002).

The basic FFM model as a whole has been shown to be replicable across diverse cultures, but considerably less is known about the equivalence of the relationships between items and individual personality *factors* in cross-cultural research. Thus, there is a need to examine the cross-cultural equivalence of *items* that comprise the FFM scales (Huang, Church, & Katigbak, 1997). This is especially important with the IPIP given its increasing prominence in organizational research and considering that the IPIP was *intended* to be an international instrument for use in multiple countries and contexts.

Instrument Translation in Cross-Cultural Research

One known issue in cross-cultural research is that when measures are translated, differences in language and language usage can result in a measure that has different psychometric properties across cultural groups (Candell & Hulin, 1987; Drasgow, 1987; Ellis, Minsel, & Becker, 1989; Hulin, Drasgow, & Komocar, 1982; Hulin & Mayer, 1986). In other words, often the translated measure is not invariant across cultures due to differences in languages and usage. In particular, slang and colloquial language does not translate well. When a measure is translated from a source language to a target language and administered to diverse cultural groups, differences in the psychometric properties of the instrument could be due to cultural differences, problems in the translation process, actual differences on the latent construct, or all three (Ryan, Horvath, Ployhart, Schmitt, & Slade, 2000). Thus, it is important to evaluate the invariance of the psychometric properties of the measure across groups so that observed score differences can be attributed to either a lack of measurement invariance or latent differences between groups (Meredith, 1993; Vandenberg & Lance, 2000).

Brislin (1986, 1970) is attributed with developing the most widely used model of instrument translation, the back-translation process. Brislin's (1970) model involves two steps: (a) Forward translation (i.e., translation from the source language to the target language)

by a bilingual professional, and (b) translation back to the source language by another bilingual professional. The back-translation is then compared with the original instrument in the source language and the process can be repeated if necessary. Although Brislin's model (1970) has been widely used in some fields, it has also been criticized. The *Standards for Educational and Psychological Testing* (1999), published by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME), do not recommend using back translation exclusively. The publishers of the *Standards* argue that back translation "may provide an artificial similarity of meaning across languages but not the best version in the new language" (p. 92). As the *Standards* imply, back-translation alone does not ensure the equivalent performance of a measure across groups.

Another issue is the practicality or feasibility of translating measures into multiple languages represented in the workforce. As organizations become more multinational and American companies turn increasingly to foreign workers (Earley & Gibson, 2002), the multinational diversity in the US workforce will continue to increase. Translating instruments into multiple languages greatly increases the logistical complexity and cost of assessments. Thus, while translation may be a good option for a bilingual workforce, it can quickly become impractical for a multilingual workforce. Moreover, there are many cases where instruments are not available in respondents' native languages and translators are difficult to find. These logistical issues, coupled with difficulties that may arise during the translation process, make translation a less than optimal solution.

One potential solution to the problems introduced by the inherent difficulties in translation would appear to be to administer instruments in the same language (e.g., English) to all respondents, including bilingual respondents. Although the solution is not without potential problems², it has the advantage

² The *Standards* (1999) indicate that there are important issues to consider when testing a bilingual individual. For example, there is an important distinction between speaking a language and testing in a language, a process which involves both reading and writing.

of being logistically parsimonious in an organization where there is a predominant language for operations. Additionally, differences in languages that might lead to items with different meaning across languages can be eliminated by the use of a common language. This study investigates whether this strategy of administering an English-language assessment to native and non-native speakers of English is a viable option for use of the IPIP.

Investigating DIF on the IPIP

One way to assess measurement equivalence at the item level is to investigate the extent to which items exhibit DIF for two groups (Collins, Raju, & Edwards, 2000). DIF occurs when an item has different psychometric properties for two groups (Camilli & Sheppard, 1994). Several studies have used this approach to investigate the extent to which items on personality inventories exhibit differential functioning for different cultural groups (e.g., Ellis et al., 1993; Huang et al., 1997). These studies have investigated this topic from two major perspectives. The first approach involves examining instruments which have been translated from one language to another. For example, Ellis et al. (1993) examined the extent to which the Trier Personality Inventory, written in German and translated to English, exhibited DIF in a university student sample. The second, less common, approach is to examine the extent to which personality inventories written in one language and administered to native and non-native speakers exhibit DIF for these two groups. As an example of this approach, Huang et al. (1997) examined DIF for items on the NEO Personality Inventory administered to native and non-native English speaking students (American and Filipino). Both of these studies found evidence of DIF between the two groups examined, although there was an important difference. Ellis et al. (1993) found that only 9% of the items exhibited DIF on two different translations of the Trier Personality Inventory, while Huang et al. (1997) found that 40% of the items on the English version of the NEO-PI

exhibited DIF for native versus non-native English speakers. While the findings from the Huang et al. (1997) and Ellis et al. (1993) studies are important, these studies both use university student samples. It is unclear whether these findings would generalize to samples of working adults. Determining the extent of DIF on the IPIP between native and non-native English speakers in the workplace is important for two reasons. First, English versions of instruments are often administered in countries where English is a commonly-spoken second language (Church, 1987; Huang et al., 1997). Second, there a large number of non-native English speakers with limited English proficiency that immigrate to the United States seeking employment (e.g., Garcia-Preto, 1996). Although these individuals are non-native speakers, they are often required to operate within the constraints of an English-speaking workplace (e.g., employees speaking English while on the job; Waxer, 2004). We are unaware of any previous study of DIF for the IPIP across native and non-native English speaking groups of workers.

While many of the difficulties and obstacles associated with instrument translation are known, much less known about administering instruments in a common language to workers with different lingual backgrounds. The purpose of this study is to investigate the extent to which English language items administered to native and non-native English speakers function differentially for the two groups. More specifically, we examined DIF between native and non-native English speaking workers for each of the factors in the FFM in the 50-item version of Goldberg's (1999) IPIP.

Method

Participants

The participants in this study were 1,709 individuals from two separate projects related to training effectiveness. The first project was conducted with a large military organization and most of the participants ($N = 1410$) were military personnel who were participating in foreign language training. The other participants ($N = 95$) from this project were the foreign language instructors who were proficient in both

Bilingual individuals may possess different levels of oral and written proficiency. This difference could go undetected and lead to the faulty conclusion that an individual's ability to "speak" in a language is indicative of one's ability to "test" in the language.

English and the language being taught. The second project was conducted with the American Council on the Teaching of Foreign Languages (ACTFL), a not-for-profit organization. Participants for this project ($N = 204$) were individuals who were participating in training to become certified administrators of the ACTFL OPI[®], a test of language speaking proficiency. ACTFL OPI[®], certified testers are language educators and professionals who are qualified to conduct and rate oral proficiency interviews. Of the 1,709 participants, 1,462 (85.5%) were native English speakers, while the remaining participants ($N = 247$; 14.5%) were non-native English speakers. Individuals with missing data were removed for the purposes of the analyses conducted in this study.

Procedure

Measures. Goldberg's (1999) 50-item measure of personality was used in this study. This measure contains items which assess each of the five factors in the FFM (i.e., extraversion, agreeableness, conscientiousness, emotional stability, and openness to experience) with 10-items per factor. Respondents were asked to indicate the extent to which each item described using the following response categories: 1 = *very inaccurate*, 2 = *moderately inaccurate*, 3 = *neither inaccurate nor accurate*, 4 = *moderately accurate*, and 5 = *very accurate*. Negatively worded items were reverse-coded before conducting analysis. The coefficient alpha internal consistency estimates in this sample were as follows: Extraversion (0.85), Agreeableness (0.82), Conscientiousness (0.80), Neuroticism (0.84), and Openness (0.79).

Analyses

IRT analyses assume unidimensionality, thus exploratory factor analysis (EFA) was used to independently assess the dimensionality of each of the five factors measured by the IPIP. While use of confirmatory factor analysis would have provided an assessment of the equality of the overall structure of the five factor model, previous studies have supported the structure of the model in general (Hendriks et al., 2003), and for the IPIP in particular (Guenole & Chernyshenko, 2005; Lim & Ployhart, 2006). Moreover, the equivalence of the overall IPIP

structure across cultural groups has been supported (Ehrhart, Roesch, Ehrhart, & Kilian, 2006). Our focus, however, was on the relationships between the individual items and the personality scales. Thus, IRT is especially well suited to this task (Maurer, Raju, & Collins, 1998; Meade & Lautenschlager, 2004; Raju, Laffitte, & Byrne, 2002). The entire sample was used to test the dimensionality of the personality scales.

IRT Analyses. Item and person parameters were estimated using the graded response model (Samejima, 1969) via MULTILOG 7.0 (Thissen, Chen, & Bock, 2003). Item and person parameters were estimated separately for each of the five factors in each group (i.e., parameters were estimated for native English and non-native English speakers independently). All subsequent analyses were conducted separately for each of the personality factors.

Before assessing DIF, it was necessary to put the estimated parameters on the same metric. This was accomplished by using a modified version of the test characteristic curve method (Stocking & Lord, 1983) to estimate linking coefficients using the Equate 2.1 program (Baker, 1995). Under the test characteristic curve method, true scores are estimated for each group using information from all item parameters estimated separately in the two groups. Next, linking constants are estimated via an iterative process to minimize the sum of the squared differences in true scores across several points on a general theta distribution (Baker, 1995)³. This was accomplished by linking the scale of the native English speakers to the scale of the non-native English speakers.

DIF and differential test functioning (DTF) were then assessed using the DFIT framework (Raju, 1999). DFIT is a computer program which provides information for assessing DIF at the item-level, using a non-compensatory DIF index (NCDIF) and at the test-level, using a compensatory DIF index (CDIF) and a DTF index. NCDIF values which exceeded 0.096 were considered to be significant and indicative of DIF (Raju, 1999). DTF indices which

³ Note that the original Stocking and Lord (1983) procedure minimized the estimated true scores across all respondents, not a theta distribution.

exceeded 0.96 for each of the personality scales were considered to be indicative of DTF (Raju, 1999).

An iterative linking process was used if items in any of the personality scales were found to exhibit DIF. This process involved estimating linking coefficients using all scale items, identifying DIF items using DFIT, then re-estimating linking coefficients by omitting the items exhibiting DIF as anchor items. These new linking coefficients were then used to transform the native English speaker group item parameters onto the metric of the non-native speaker group. DIF was then examined for all items using these newly transformed item parameters. This process was repeated until the same set of items was identified as DIF items in consecutive runs. Iterative linking has been shown to provide more accurate DIF detection than single-stage linking (Kim & Cohen, 1992; Park & Lautenschlager, 1990). Convergence in the list of DIF items was achieved after only two iterations for all scales.

Results

Unidimensionality of the Personality Scales

Tables 1-5 show the means and standard deviations for all of the items on each of the personality subscales along with the linked item parameters for each group. The EFAs for each of the personality subscales provided evidence of unidimensionality. All of the first eigenvalues for each of the scales accounted for more than 35% of the total variance while the second factor accounted for less than 12% (see Table 6). Inspection of the scree plots for each of the five scales indicated clear unidimensionality as well.

DIF and DTF Results

None of the Big 5 personality scales exhibited significant DTF. However, several individual items exhibited significant DIF, with results varying greatly by scale. Table 7 provides C-DIF, NC-DIF, and DTF values for each of the items on the five scales. No items on the extraversion scale exhibited DIF, one item on the agreeableness scale exhibited DIF (i.e., Item 1), three items on the conscientiousness scale exhibited DIF (i.e., Items 2, 9, and 10), five items on the emotional stability scale

exhibited DIF (i.e., Items 3, 4, 7, 8, and 9), and two items on the openness scale exhibited DIF (i.e., Items 1 and 8). Results for each of the five factors are discussed in more detail below.

The only agreeableness item with significant levels of DIF was Item 1 (“Feel little concern for others”), which showed large differences in the a parameter. Thus, this item was much more reflective of the agreeableness construct for the native speaking sample than the non-native speaker sample. Put differently, the relationship between latent agreeableness and this item was lower in the non-native speaker sample.

Item 2 on the conscientiousness scale, “Leave my belongings around”, had a somewhat higher a parameter in the native than the non-native speaking sample, however this item also had minor differences in b parameters. In particular, it took somewhat lower levels of the latent trait for non-native speakers than native English speakers to indicate a (reverse coded) response of two or higher. Conversely, both Items 9 (“Follow a schedule”) and 10 (“Am exacting in my work”), had slightly higher a parameters in the non-native than the native speaker group. For both Items 9 and 10, however, item b parameters for the lowest and highest response options were very different across groups. In the native English speaking group, there was much more distance between the boundary response functions (BRFs; see Figure 1 for Item 9) than in the non-native group. Such differences in b parameters indicate that if latent levels of conscientiousness were the same in the two groups, considerably more variability in observed scores would be seen in the non-native speaker group with more central tendency in the native speaker group.

Five emotional stability items indicated significant DIF. Items 3 (“Worry about things”) and 4 (“Seldom feel blue”) showed DIF primarily on the a parameter. Item 4 exhibited the largest DIF of all items (NCDIF = 0.494) with the item performing extremely poorly in the non-native group ($a = 0.290$; see Figure 2). Interestingly, although this item exhibited a large amount of DIF, its counterpart (Item 10: “Often feel blue”) did not exceed the NCDIF cutoff (NCDIF = 0.045). This comparison suggests that the DIF associated with Item 4 may be related to use of the word “seldom” since

this is the only word that differs between the two items. Conversely, Items 7, 8, and 9 showed differences primarily on *b* parameters. Item 7 uniformly “favored” the non-native group as across all levels of theta, such that lower levels of emotional stability were needed to indicate higher levels of (reverse coded) agreement on the item than in the native speaker group. Items 8 and 9 tended to have more widely spaced *b* parameters in the native speaker group (again indicating more probable central tendency with that group) with the (reverse coded) highest response option on Item 9 being particularly more unlikely to be chosen for native than non-native speakers (see Figure 3).

Lastly, two items on the openness to experience scale indicated DIF. Item 1, (“Have a rich vocabulary”), exhibited *b* parameter DIF, in which members of the non-native speaker group were more likely to agree with the item than members of the native speaker group with the same openness to experience levels. In other words, more of the underlying trait is necessary in the native speaker group to have the same probability of response as the non-native speaker group. This makes sense if non-native English speakers are considering their vocabulary in both their native language and their secondary language when responding to this item. However, this explanation is purely speculative and would require further investigation. Item 8 (“Use difficult words”), had somewhat low *a* parameters in both groups, with the non-native sample *a* parameter being particularly low. It is possible that for this item, the very definition of a difficult word varies considerably across samples.

Discussion

The results of this study showed that 11 out of the 50 items (22%) on the IPIP (Goldberg, 1999) exhibited DIF when comparing native and non-native English speakers. This proportion of items with significant DIF was high considering both the previous validation work involved with the IPIP (e.g., Goldberg, 1999; Lim & Ployhart, 2006) and that the IPIP was designed for international and cross-cultural use. These findings indicate that researchers and practitioners should use caution when administering the English version of Goldberg's

(1999) IPIP measure to non-native English speakers. Although the majority of items did not exhibit DIF, and there was no DTF found, the items that did exhibit DIF required careful examination in order to determine the nature of the DIF and thus the appropriateness of administering these items to non-native English speakers. Moreover, the size of the DIF encountered was quite large for many items.

One interesting finding was the degree to which the personality factors differed by the presence of items exhibiting DIF. In general, when factor analyzed, the Big 5 emerge in order of strength as Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and finally Openness to Experience. We found no DIF items for Extraversion, only one DIF Agreeableness item, three DIF Conscientiousness items, and five DIF Emotional Stability items. Openness countered this trend with only two DIF items. Still, the inverse relationship between the presence of DIF and the strength of the factor was notable.

It is difficult to say the extent to which the results from this study generalize to other (non-personality) measures administered to native and non-native English speakers. It is possible that the use of colloquial language in the IPIP could be more prone to different interpretations than would be other organizational assessments using more formal language. To some extent, this argument is bolstered by our findings that DIF was not evenly distributed across the five personality factors (i.e., some factors had more items with DIF than others). The particularly low *a* parameters for some items indicate their inappropriateness for use in non-native English speaking samples. However, the IPIP is based on the most widely accepted model of personality (the five factor model; Costa & McCrae, 1992) and has previous validation work (Goldberg, 1999; Lim & Ployhart, 2006) as is typical of the most widely used personality assessments. To this extent, use of any English language personality measures in non-native English speaking samples would seem likely to show DIF. This would seem potentially problematic for multi-national organizations that wish to use the same validated predictors for employee selection across countries, or within a single country for native and non-native speakers. The

likelihood of DIF in non-personality measures is more difficult to judge, yet remains a distinct possibility. In any case, we strongly recommend that researchers test for DIF for any measure when administering the measure to a non-native speaker group.

Translated instruments are known to be prone to DIF because they are susceptible to both differences in cultures and to differences in language. Administering an assessment in English to both native and non-native English speakers seems, on the surface, to be an appropriate strategy for eliminating one potential source of DIF. However, as indicated in this study, this strategy may be inappropriate to the extent that items administered in the same language need not function equivalently across groups.

When faced with such a dilemma, organizations are advised to first evaluate the existence of DIF during pilot testing across native and non-native speakers. If no DIF is found, it makes sense to administer the instrument to both groups in the common language. If minor DIF is found on very few items, the content of those items should be reviewed and revised if possible. If pervasive DIF is found, the organization should consider the possibility of translation. If the organization has a large number of employees and there are relatively few languages represented in the workforce, then it probably makes financial sense to translate. However, when there are many native languages represented in the workforce, translation is less desirable. Organizations should consider translating to as many languages as there are sufficient group members to justify the investment.

Limitations, Future Research, and Implications

As with all studies, this study has limitations. First, while we believe that our use of a sample engaged in training in an organizational setting represents a much needed extension to previous work on the IPIP using student samples, it should be noted that our sample was largely from a military organization. As such, this organization differs somewhat from other organizations (such as commercial businesses, not-for-profit organizations, etc.). We note, however, that personality measures are

not inherently organization specific and that we did see considerable variance on all five factors investigated as would be expected with non-military organizations. Additionally, we were interested in the invariance of the items across groups, not in comparing groups themselves. Thus, even when observed or latent mean differences are anticipated, we would not expect to necessarily see DIF on the IPIP items.

We also note that our sample of non-native English speakers was somewhat heterogeneous in that non-native speakers were drawn from a large and diverse number of languages and cultures. One limitation of this diverse nature of non-native English speakers is that it is difficult to determine potential cultural influences that may have affected respondents' interpretation of the items. However, this diversity can also be seen as a great advantage such that with a large and diverse number of cultures being represented, it is unlikely that any one cultural influence was responsible for the DIF witnessed in this study.

The present study lays the initial groundwork for the examination of DIF across native and non-native English speakers, but much future work is needed. First, while this study examined the IPIP as a specific example, it would be beneficial to investigate the differential functioning of other measures commonly used in organizational research to see if they also exhibit DIF across native and non-native English speaking groups. Additionally, a replication of the performance of the IPIP in other organizational or research contexts would be useful to see if the items that exhibited DIF in this study exhibit DIF in other settings. If future studies find that the same items exhibit DIF, these items may need to be removed from the measure or re-worded if the measure is going to be used with a sample of both native and non-native English speakers. If different items exhibit DIF, then more extensive work will be needed to determine the appropriateness of administering English versions of this measure to non-native English speakers.

In any case, the implications of the current study are clear. Simply administering a measure in a common language to individuals with different cultural and lingual backgrounds is not sufficient to ensure the identical functioning of

the measure across those groups. When an English language measure is intended for use in a non-native English speaking sample, the examination of the invariance of that measure must occur before respondents from differing groups can be adequately compared.

References

- American Council on the Teaching of Foreign Languages. (2006, August). www.actfl.org
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: Author.
- Baker, F. B. (1995). EQUATE 2.1: Computer program for equating two metrics in item response theory: Madison: University of Wisconsin, Laboratory of Experimental Design.
- Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*, 1-26.
- Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next? *International Journal of Selection and Assessment, 9*, 9-30.
- Block, J. (1995). A contrarian view of the five-factor approach to personality description. *Psychological Bulletin, 117*, 187-215.
- Bowling, N. A., Beehr, T. A., & Swader, W. M. (2005). Giving and receiving social support at work: The roles of personality and reciprocity. *Journal of Vocational Behavior, 67*, 476-489.
- Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology, 1*, 185-216.
- Brislin, R. W. (1986). The wording and translation of research instruments. In W. J. Lonner & J. W. Berry (Eds.), *Field methods in cross-cultural research* (pp. 137-164). Beverly Hills, CA: Sage.
- Brown, D. J., Cober, R. T., Kane, K., Levy, P. E., & Shalhoop, J. (2006). Proactive personality and the successful job search: A field investigation with college graduates. *Journal of Applied Psychology, 91*, 717-726.
- Burke, R. J., & Ng, E. (2006). The changing nature of work and organizations: Implications for human resource management. *Human Resource Management Review, 16*, 86-94.
- Camilli, G., & Sheppard, L. A. (1994). *Methods for identifying biased test items*. Thousand Oaks, CA: Sage Publications.
- Candell, G. L., & Hulin, C. L. (1987). Cross-language and cross-cultural comparisons in scale translations: Independent sources of information about item nonequivalence. *Journal of Cross-Cultural Psychology, 17*, 417-440.
- Church, A. T. (1987). Personality research in a non-Western culture: The Philippines. *Psychological Bulletin, 102*, 272-292.
- Collins, W. C., Raju, N. S., & Edwards, J. E. (2000). Assessing differential functioning in a satisfaction scale. *Journal of Applied Psychology, 85*, 451-461.
- Costa, P. T., Jr., & McCrae, R. R. (1992). *NEO-PI-R professional manual*. Odessa, FL: Psychological Assessment Resources.
- Cucina, J. M., Vasilopoulos, N. L., & Sehgal, K. G. (2005). Personality-based job analysis and the self-serving bias. *Journal of Business and Psychology, 20*(2), 275-290.
- Dahlin, K. B., Weingart, L. R., & Hinds, P. J. (2005). Team diversity and information use. *Academy of Management Journal, 48*, 1107-1123.

- Dragow, F. (1987). Study of the measurement bias of two standardized psychological tests. *Journal of Applied Psychology, 72*, 19-29.
- Earley, P. C., & Gibson, C. B. (2002). *Multinational work teams: A new perspective*. Mahway, NJ: Erlbaum.
- Ehrhart, K. H., Roesch, S. C., Ehrhart, M. G., & Kilian, B. (2006). *A test of the factor structure equivalence of the 50-item IPIP five-factor model measure across gender and ethnic groups*. Paper presented at the 21st annual conference of the Society for industrial and organizational psychology, Dallas, TX.
- Ellis, B. B., Becker, P., & Kimmel, H. D. (1993). An item response theory evaluation of an English version of the Trier Personality Inventory (TPI). *Journal of Cross-Cultural Psychology, 24*, 133-148.
- Ellis, B. B., Minsel, B., & Becker, P. (1989). Evaluation of attitude survey translations: An investigation using item response theory. *International Journal of Psychology, 24*, 133-148.
- Garcia-Preto, N. (1996). Latino families: An overview. In M. McGoldrick, J. Giordano, J. K. Pearce (Eds.), *Ethnicity and family therapy* (2nd ed., pp. 169-182). New York: Guilford Press.
- Goldberg, L. R. (1990). An alternative "description of personality": The Big-Five Factor structure. *Journal of Personality and Social Psychology, 59*, 1216-1229.
- Goldberg, L. R. (1992). The development of marker variables for the Big-Five Factor structure. *Psychological Assessment, 4*, 26-42.
- Goldberg, L. R. (1999). A broad-bandwidth, public domain, personality inventory measuring the lower-level facets of several five-factor models. In I. Mervielde, I. Deary, F. De Fruyt, & Ostendorf (Eds.), *Personality Psychology in Europe*, Vol. 7 (pp. 7-28). Tilburg, The Netherlands: Tilburg University Press.
- Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., et al. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality, 40*, 84-96.
- Guenole, N., & Chernyshenko, O. S. (2005). The suitability of Goldberg's Big Five IPIP personality markers in New Zealand: A dimensionality, bias, and criterion validity evaluation. *New Zealand Journal of Psychology, 34*, 86-96.
- Hattrup, K., O'Connell, M. S., & Wingate, P. H. (1998). Prediction of multidimensional criteria: Distinguishing task and contextual performance. *Human Performance, 11*, 305-319.
- Hendriks, A. A. J., Perugini, M., Angleitner, A., Ostendorf, F., Johnson, J. A., De Fruyt, F., et al. (2003). The Five-Factor Personality Inventory: Cross-cultural generalizability across 13 countries. *Journal of Personality, 17*, 347-373.
- Hough, L. M., & Schneider, R. J. (1996). Personality traits, taxonomies, and applications in organizations. In K. R. Murphy (Ed.), *Individual differences and behavior in organizations*. San Francisco: Jossey-Bass. pp. 31-88.
- Huang, C. D., Church, A. T., & Katigbak, M. S. (1997). Identifying cultural differences in items and traits: Differential item functioning in the NEO Personality Inventory. *Journal of Cross-Cultural Psychology, 28*, 192-218.
- Hulin, C. L., Dragow, F., & Komocar, J. (1982). Application of item response theory to analysis of attitude scale translations. *Journal of Applied Psychology, 67*, 818-825.
- Hulin, C. L., & Mayer, L. J. (1986). Psychometric equivalence of a translation of

- the Job Description Index into Hebrew. *Journal of Applied Psychology*, 71, 83-94.
- Kim, S.-H., & Cohen, A. S. (1992). Effects of linking methods on detection of DIF. *Journal of Educational Measurement*, 29, 51-66.
- Lim, B., & Ployhart, R. E. (2006). Assessing the convergent and discriminant validity of Goldberg's International Personality Item Pool: A multitrait-multimethod examination. *Organizational Research Methods*, 9, 29-54.
- Maurer, T. J., Raju, N. S., & Collins, W. C. (1998). Peer and subordinate performance appraisal measurement equivalence. *Journal of Applied Psychology*, 83, 693-702.
- McAdams, D. P. (1992). The five-factor model: Issues and applications. *Journal of Personality*, 60, 329-361.
- McCrae, R. R., & Costa, P. T., Jr. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*, 52, 81-90.
- Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58, 525-543.
- Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods*, 7, 361-388.
- Orlando, M., & Marshall, G. N. (2002). Differential item functioning in a Spanish translation of the PTSD checklist: Detection and evaluation of impact. *Psychological Assessment*, 14, 50-59.
- Park, D. G., & Lautenschlager, G. J. (1990). Improving IRT item bias detection with iterative linking and ability scale purification. *Applied Psychological Measurement*, 14, 163-173.
- Raju, N. S. (1999). DFITPS6: A Fortran program for calculating DIF/DTF [Computer Program]. Chicago: Illinois Institute of Technology.
- Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology*, 87, 517-529.
- Rothstein, M. G., & Goffin, R. D. (2006). The use of personality measures in personnel selection: What does current research support? *Human Resource Management Review*, 16, 155-180.
- Rubin, R. S., Munz, D. C., & Bommer, W. H. (2005). Leading from within: The effects of emotion recognition and personality on transformational leadership behavior. *Academy of Management Journal*, 48, 845-858.
- Ryan, A. M., Horvath, M., Ployhart, R. E., Schmitt, N., & Slade, L. A. (2000). Hypothesizing differential item functioning in global opinion surveys. *Personnel Psychology*, 53, 531-562.
- Sackett, P. R., & DeVore, C. J. (2002). Counterproductive behaviors at work. In N. Anderson et al. (Eds.) *Handbook of Industrial, Work, & Organizational Psychology, Volume 1: Personnel Psychology*. Thousand Oaks, CA: Sage.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, No. 17.
- Salgado, J. F. (1997). The five factor model of personality and job performance in the European Community. *Journal of Applied Psychology*, 82, 30-43.

- Salgado, J. F., Viswesvaran, C., & Ones, D. S. (2001). Predictors used for personnel selection: An overview of constructs, methods and techniques. In N. Anderson, D. S. Ones, H. K. Sinangil, & C. Viswesvaran (Eds.), *Handbook of Industrial, Work, and Organizational Psychology* (vol. 1). London: Sage Publications.
- Schmit, M. J., Kihm, J. A., & Robie, C. (2000). The development of a global measure of personality, *Personnel Psychology, 1*, 153-193.
- Speizer, I. (2006). The state of training and development: More spending, more scrutiny. *Workforce Management, 25-26*.
- Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201-210.
- Tett, R. P., Jackson, D. N., & Rothstein, M. (1991). Personality measures as predictors of job performance: A meta-analytic review. *Personnel Psychology, 44*, 703-742.
- Thissen, D., Chen, W., & Bock, D. (2003). MULTILOG 7.0 [Computer Program]. Lincolnwood, IL: Scientific Software International.
- Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4-69.
- Waxer, C. (2004). English-only policies can translate into problems for employers. *Workforce Management*, pp. 57-59.

Table 1

Descriptive Statistics and Item Parameters for Extraversion

| | Native English Speakers | | | | Non-Native English Speakers | | | |
|--|-------------------------|------|------|------------------------|-----------------------------|------|------|------------------------|
| | M | SD | a | b1 b2 b3 b4 | M | SD | a | b1 b2 b3 b4 |
| 1. Am the life of the party. | 3.09 | 1.02 | 1.33 | -2.13 -1.13 0.78 2.59 | 3.00 | 1.18 | 1.15 | -1.94 -0.88 0.60 2.34 |
| 2. Don't talk a lot. (RS) | 3.32 | 1.08 | 1.91 | -2.32 -0.86 0.21 1.54 | 3.31 | 1.12 | 1.37 | -3.04 -0.95 0.13 1.47 |
| 3. Feel comfortable around people. | 4.04 | 0.85 | 1.38 | -3.77 -2.56 -1.30 0.88 | 4.18 | 0.87 | 1.45 | -3.54 -2.63 -1.42 0.31 |
| 4. Keep in the background. (RS) | 3.28 | 0.95 | 1.54 | -2.97 -1.16 0.40 2.10 | 3.00 | 1.06 | 1.02 | -2.84 -1.02 0.92 2.50 |
| 5. Start conversations. | 3.79 | 0.86 | 2.24 | -3.16 -1.68 -0.58 1.27 | 3.86 | 0.97 | 1.73 | -2.90 -1.91 -0.71 0.82 |
| 6. Have little to say. (RS) | 3.59 | 0.92 | 1.80 | -3.37 -1.52 -0.13 1.53 | 3.65 | 1.03 | 1.02 | -4.10 -2.14 -0.35 1.35 |
| 7. Talk to a lot of different people at parties. | 3.63 | 1.02 | 1.98 | -2.54 -1.32 -0.31 1.27 | 3.28 | 1.21 | 1.67 | -1.86 -0.90 -0.07 1.41 |
| 8. Don't like to draw attention to myself. (RS) | 2.54 | 0.95 | 0.90 | -2.37 0.09 2.25 4.86 | 2.65 | 1.17 | 0.84 | -2.11 -0.07 1.63 3.27 |
| 9. Don't mind being the center of attention. | 3.20 | 1.00 | 1.31 | -2.62 -1.19 0.53 2.39 | 3.00 | 1.19 | 1.20 | -1.86 -0.80 0.52 2.34 |
| 10. Am quiet around strangers. (RS) | 3.06 | 1.03 | 1.58 | -2.32 -0.67 0.60 2.30 | 2.88 | 1.13 | 1.71 | -1.84 -0.37 0.65 1.86 |

Note. RS = reverse scored.

Table 2

Descriptive Statistics and Item Parameters for Agreeableness

| | Native English Speakers | | | | | Non-Native English Speakers | | | | | | | | |
|---|-------------------------|------|------|-------|-------|-----------------------------|-------|------|------|------|-------|-------|-------|-------|
| | M | SD | a | b1 | b2 | b3 | b4 | M | SD | a | b1 | b2 | b3 | b4 |
| 1. Feel little concern for others. (RS) | 4.10 | 1.02 | 1.80 | -3.71 | -2.67 | -1.85 | -0.67 | 3.75 | 1.48 | 0.61 | -3.42 | -2.07 | -1.27 | -0.01 |
| 2. Am interested in people. | 4.03 | 0.84 | 1.84 | -3.97 | -3.07 | -1.92 | -0.12 | 4.27 | 0.94 | 1.43 | -3.16 | -2.53 | -1.63 | -0.05 |
| 3. Insult people. (RS) | 3.94 | 1.04 | 0.96 | -5.45 | -3.45 | -1.72 | -0.23 | 4.56 | 0.89 | 1.10 | -4.15 | -3.10 | -2.20 | -1.22 |
| 4. Sympathize with others' feelings. | 3.66 | 0.91 | 2.44 | -3.35 | -2.26 | -1.26 | 0.40 | 4.30 | 0.83 | 1.89 | -3.05 | -2.57 | -1.51 | 0.02 |
| 5. Am not interested in other people's problems. (RS) | 3.68 | 0.95 | 1.96 | -3.82 | -2.38 | -1.23 | 0.27 | 3.91 | 1.10 | 1.32 | -3.41 | -1.81 | -0.93 | 0.43 |
| 6. Have a soft heart. | 3.08 | 1.02 | 1.31 | -3.31 | -1.87 | -0.12 | 1.43 | 3.74 | 1.01 | 0.74 | -4.85 | -3.45 | -0.70 | 1.63 |
| 7. Am not really interested in others. (RS) | 3.90 | 0.88 | 2.33 | -3.78 | -2.63 | -1.53 | -0.03 | 4.13 | 1.05 | 1.45 | -2.95 | -2.20 | -1.17 | 0.03 |
| 8. Take time out for others. | 3.77 | 0.77 | 1.55 | -4.41 | -3.13 | -1.64 | 0.85 | 3.82 | 1.08 | 0.87 | -3.73 | -2.70 | -1.22 | 1.17 |
| 9. Feel others' emotions. | 3.36 | 0.96 | 2.00 | -3.11 | -2.06 | -0.82 | 0.93 | 4.14 | 0.84 | 1.32 | -3.61 | -2.98 | -1.65 | 0.57 |
| 10. Make people feel at ease. | 3.70 | 0.75 | 1.24 | -5.53 | -3.78 | -1.37 | 1.15 | 4.18 | 0.81 | 1.27 | -4.84 | -3.25 | -1.51 | 0.42 |

Note. RS = reverse scored.

Table 3

Descriptive Statistics and Item Parameters for Conscientiousness

| | Native English Speakers | | | | Non-Native English Speakers | | | |
|--|-------------------------|------|------|-------------------------|-----------------------------|------|------|-------------------------|
| | M | SD | a | b1 b2 b3 b4 | M | SD | a | b1 b2 b3 b4 |
| 1. Am always prepared. | 3.94 | 0.72 | 1.55 | -3.91 -2.52 -1.19 1.38 | 4.00 | 0.89 | 1.07 | -4.47 -3.01 -1.37 0.91 |
| 2. Leave my belongings around. (RS) | 3.97 | 1.03 | 1.71 | -2.86 -1.70 -0.77 0.46 | 3.63 | 1.27 | 0.90 | -3.61 -1.57 -0.42 0.74 |
| 3. Pay attention to details. | 4.21 | 0.70 | 1.99 | -3.48 -2.55 -1.56 0.55 | 4.13 | 0.90 | 1.56 | -3.35 -2.47 -1.25 0.36 |
| 4. Make a mess of things. (RS) | 4.14 | 0.86 | 1.98 | -3.34 -2.21 -0.97 0.36 | 4.12 | 1.06 | 1.60 | -2.85 -2.07 -1.01 0.01 |
| 5. Get chores done right away. | 3.70 | 0.89 | 1.57 | -3.40 -1.84 -0.47 1.40 | 3.85 | 0.98 | 1.73 | -3.38 -1.85 -0.63 0.74 |
| 6. Often forget to put things back in their proper place. (RS) | 3.94 | 0.97 | 1.90 | -2.88 -1.65 -0.80 0.62 | 3.90 | 1.17 | 1.51 | -3.04 -1.47 -0.73 0.22 |
| 7. Like order. | 3.82 | 0.82 | 1.37 | -4.02 -2.48 -0.75 1.42 | 3.77 | 1.14 | 0.94 | -3.21 -2.33 -0.94 1.00 |
| 8. Shirk my duties. (RS) | 4.36 | 0.81 | 1.22 | -5.07 -3.61 -1.58 -0.18 | 4.10 | 1.08 | 0.74 | -5.10 -3.63 -1.35 -0.03 |
| 9. Follow a schedule. | 3.72 | 0.84 | 1.30 | -3.82 -2.19 -0.72 1.82 | 4.16 | 0.92 | 1.89 | -2.77 -2.05 -1.22 0.27 |
| 10. Am exacting in my work. | 3.78 | 0.74 | 1.46 | -4.80 -2.71 -0.62 1.63 | 4.09 | 0.88 | 1.83 | -3.02 -2.34 -1.04 0.43 |

Note. RS = reverse scored.

Table 4

Descriptive Statistics and Item Parameters for Emotional Stability

| | Native English Speakers | | | | Non-Native English Speakers | | | |
|------------------------------------|-------------------------|------|------|------------------------|-----------------------------|------|------|------------------------|
| | M | SD | a | b1 b2 b3 b4 | M | SD | a | b1 b2 b3 b4 |
| 1. Get stressed out easily. (RS) | 3.96 | 0.97 | 1.68 | -2.87 -1.41 -0.25 1.11 | 3.54 | 1.29 | 1.33 | -2.21 -1.23 -0.14 0.80 |
| 2. Am relaxed most of the time. | 3.88 | 0.90 | 1.02 | -3.97 -2.10 -0.80 1.93 | 3.40 | 1.13 | 0.80 | -4.06 -1.53 -0.18 2.24 |
| 3. Worry about things. (RS) | 3.22 | 1.09 | 1.17 | -2.68 -0.43 0.83 2.52 | 2.62 | 1.11 | 0.66 | -2.99 0.21 2.24 4.02 |
| 4. Seldom feel blue. | 3.64 | 1.08 | 1.06 | -2.87 -1.49 -0.10 1.92 | 2.92 | 1.22 | 0.29 | -5.64 -2.44 2.31 7.37 |
| 5. Am easily disturbed. (RS) | 3.90 | 0.93 | 1.71 | -2.85 -1.45 -0.27 1.32 | 3.61 | 1.09 | 1.35 | -3.53 -1.46 -0.21 1.06 |
| 6. Get upset easily. (RS) | 4.00 | 0.95 | 2.24 | -2.25 -1.27 -0.34 1.04 | 3.82 | 1.11 | 2.19 | -2.26 -1.42 -0.48 0.57 |
| 7. Change my mood a lot. (RS) | 3.74 | 0.93 | 2.15 | -2.41 -1.18 0.11 1.54 | 3.94 | 1.04 | 1.99 | -2.80 -1.73 -0.53 0.42 |
| 8. Have frequent mood swings. (RS) | 4.19 | 0.91 | 2.51 | -2.62 -1.43 -0.42 0.64 | 4.15 | 1.01 | 2.73 | -2.33 -1.73 -0.76 0.07 |
| 9. Get irritated easily. (RS) | 3.74 | 0.99 | 2.23 | -2.31 -0.91 0.01 1.44 | 3.89 | 1.18 | 2.53 | -1.99 -1.31 -0.56 0.30 |
| 10. Often feel blue. (RS) | 4.22 | 0.81 | 2.13 | -3.01 -1.88 -0.66 0.76 | 4.07 | 1.04 | 1.88 | -2.89 -1.86 -0.74 0.16 |

Note. RS = reverse scored.

Table 5

Descriptive Statistics and Item Parameters for Openness to Experience

| | Native English Speakers | | | | Non-Native English Speakers | | | | | | | | | |
|---|-------------------------|------|------|-------|-----------------------------|-------|------|------|------|------|-------|-------|-------|------|
| | M | SD | a | b1 | b2 | b3 | b4 | M | SD | a | b1 | b2 | b3 | b4 |
| 1. Have a rich vocabulary. | 3.53 | 0.99 | 1.23 | -3.27 | -1.82 | -0.23 | 1.72 | 3.85 | 1.01 | 1.18 | -3.54 | -2.26 | -0.81 | 0.93 |
| 2. Have difficulty understanding abstract ideas. (RS) | 3.98 | 0.86 | 1.57 | -3.95 | -2.54 | -0.94 | 0.76 | 3.80 | 1.11 | 1.12 | -4.06 | -1.91 | -0.62 | 0.62 |
| 3. Have a vivid imagination. | 3.95 | 0.87 | 1.54 | -3.62 | -2.47 | -0.93 | 0.86 | 3.91 | 0.97 | 1.14 | -3.53 | -2.69 | -1.08 | 0.93 |
| 4. Am not interested in abstract ideas. (RS) | 3.88 | 0.92 | 1.40 | -3.93 | -2.48 | -0.70 | 0.82 | 3.75 | 1.08 | 0.69 | -4.96 | -3.25 | -0.89 | 1.36 |
| 5. Have excellent ideas. | 3.90 | 0.68 | 1.70 | -4.99 | -3.40 | -0.88 | 1.35 | 3.97 | 0.76 | 2.48 | -4.96 | -2.50 | -0.74 | 0.84 |
| 6. Do not have a good imagination. (RS) | 4.13 | 0.92 | 1.75 | -3.10 | -2.21 | -1.25 | 0.32 | 4.14 | 1.01 | 1.07 | -4.02 | -2.82 | -1.42 | 0.10 |
| 7. Am quick to understand things. | 3.97 | 0.72 | 1.45 | -4.50 | -3.07 | -1.28 | 1.25 | 4.00 | 0.92 | 1.93 | -2.83 | -2.12 | -0.85 | 0.57 |
| 8. Use difficult words. | 2.99 | 1.04 | 0.89 | -3.00 | -1.08 | 0.93 | 3.35 | 2.70 | 1.14 | 0.43 | -3.74 | -0.97 | 2.77 | 6.34 |
| 9. Spend time reflecting on things. | 3.62 | 0.91 | 0.88 | -4.78 | -2.70 | -0.64 | 2.33 | 3.83 | 1.00 | 0.53 | -6.12 | -4.90 | -1.37 | 1.87 |
| 10. Am full of ideas. | 3.92 | 0.74 | 2.40 | -3.36 | -2.44 | -0.77 | 1.01 | 4.04 | 0.87 | 2.69 | -2.55 | -2.11 | -0.83 | 0.50 |

Note. RS = reverse scored.

Table 6

Exploratory Factor Analysis Results

| | First Eigenvalue | % of Total Variance | Second Eigenvalue | % of Total Variance |
|---------------------------|---------------------|------------------------|----------------------|------------------------|
| Extraversion | 4.23 | 42.27 | 1.06 | 10.64 |
| Agreeableness | 3.93 | 39.33 | 1.03 | 10.26 |
| Conscientiousness | 3.62 | 36.15 | 1.10 | 11.04 |
| Emotional Stability | 4.34 | 43.36 | 1.11 | 11.12 |
| Openness to Experience | 3.57 | 35.74 | 1.16 | 11.61 |

Table 7

C-DIF, NC-DIF, and DTF Values for IPIP Items

| Item No. | <u>Extraversion</u> | | <u>Agreeableness</u> | | <u>Conscientiousness</u> | | <u>Emotional Stability</u> | | <u>Openness to Experience</u> | |
|----------|---------------------|--------|----------------------|--------|--------------------------|--------|----------------------------|--------|-------------------------------|--------|
| | C-DIF | NC-DIF | C-DIF | NC-DIF | C-DIF | NC-DIF | C-DIF | NC-DIF | C-DIF | NC-DIF |
| 1 | 0.006 | 0.001 | 0.536 | 0.457* | 0.081 | 0.010 | -0.024 | 0.002 | 0.048 | 0.097* |
| 2 | -0.008 | 0.005 | 0.071 | 0.009 | -0.315 | 0.146* | -0.153 | 0.065 | -0.017 | 0.014 |
| 3 | -0.002 | 0.038 | -0.153 | 0.041 | 0.001 | 0.000 | -0.158 | 0.217* | -0.003 | 0.005 |
| 4 | -0.002 | 0.047 | -0.048 | 0.009 | 0.035 | 0.003 | -0.236 | 0.494* | -0.021 | 0.071 |
| 5 | 0.002 | 0.019 | 0.154 | 0.039 | 0.211 | 0.065 | 0.013 | 0.000 | 0.017 | 0.040 |
| 6 | -0.012 | 0.014 | 0.040 | 0.013 | 0.035 | 0.002 | 0.140 | 0.048 | -0.006 | 0.007 |
| 7 | 0.014 | 0.056 | 0.119 | 0.025 | 0.020 | 0.001 | 0.369 | 0.336* | 0.025 | 0.051 |
| 8 | 0.004 | 0.027 | 0.211 | 0.065 | -0.211 | 0.065 | 0.268 | 0.121* | -0.034 | 0.190* |
| 9 | 0.010 | 0.017 | -0.064 | 0.015 | 0.470 | 0.323* | 0.414 | 0.403* | 0.026 | 0.016 |
| 10 | 0.014 | 0.015 | -0.169 | 0.043 | 0.375 | 0.205* | 0.152 | 0.045 | 0.045 | 0.063 |
| DTF | 0.026 | | 0.698 | | 0.701 | | 0.785 | | 0.082 | |

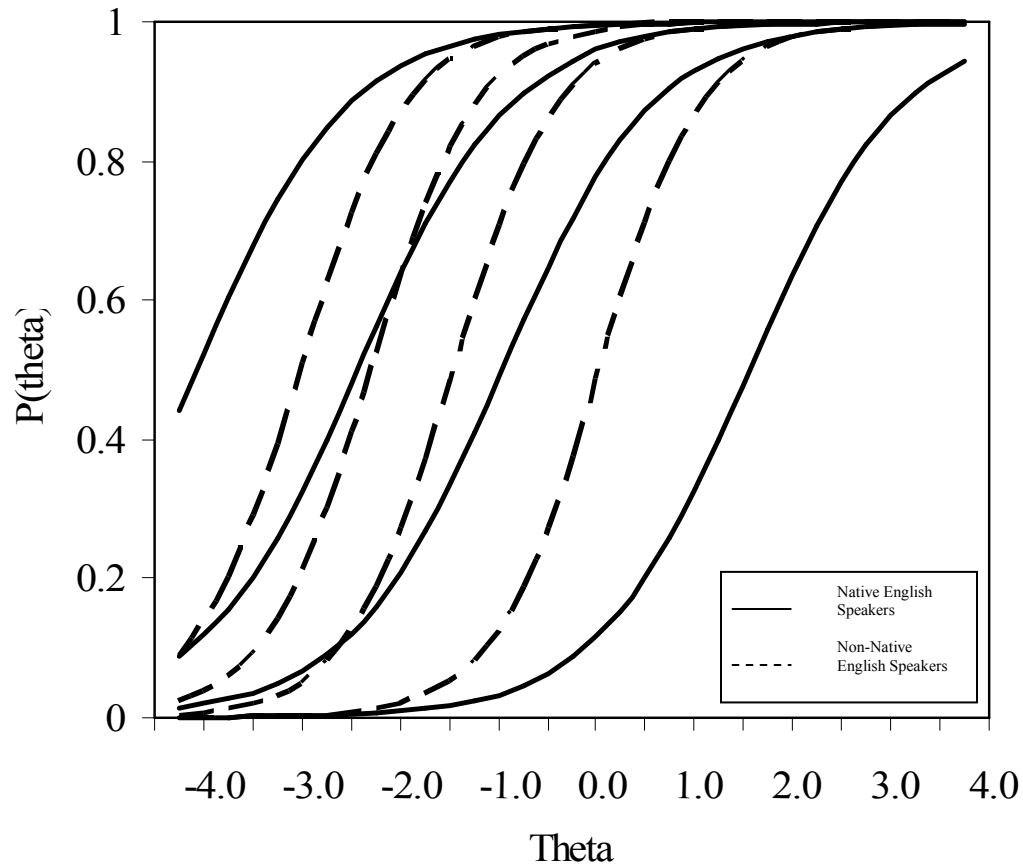
Note: * indicates significant DIF.

Figure Captions

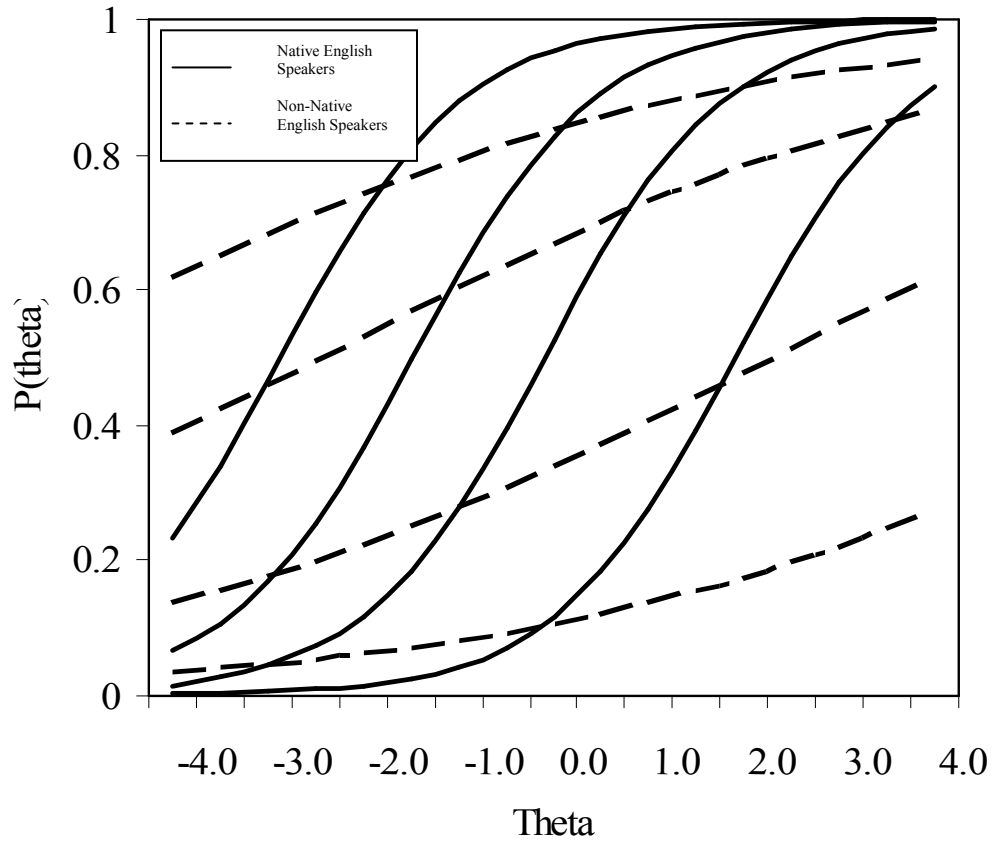
Figure 1. Boundary response functions for Item 9 on the Conscientiousness Scale.

Figure 2. Boundary response functions for Item 4 on the Emotional Stability Scale.

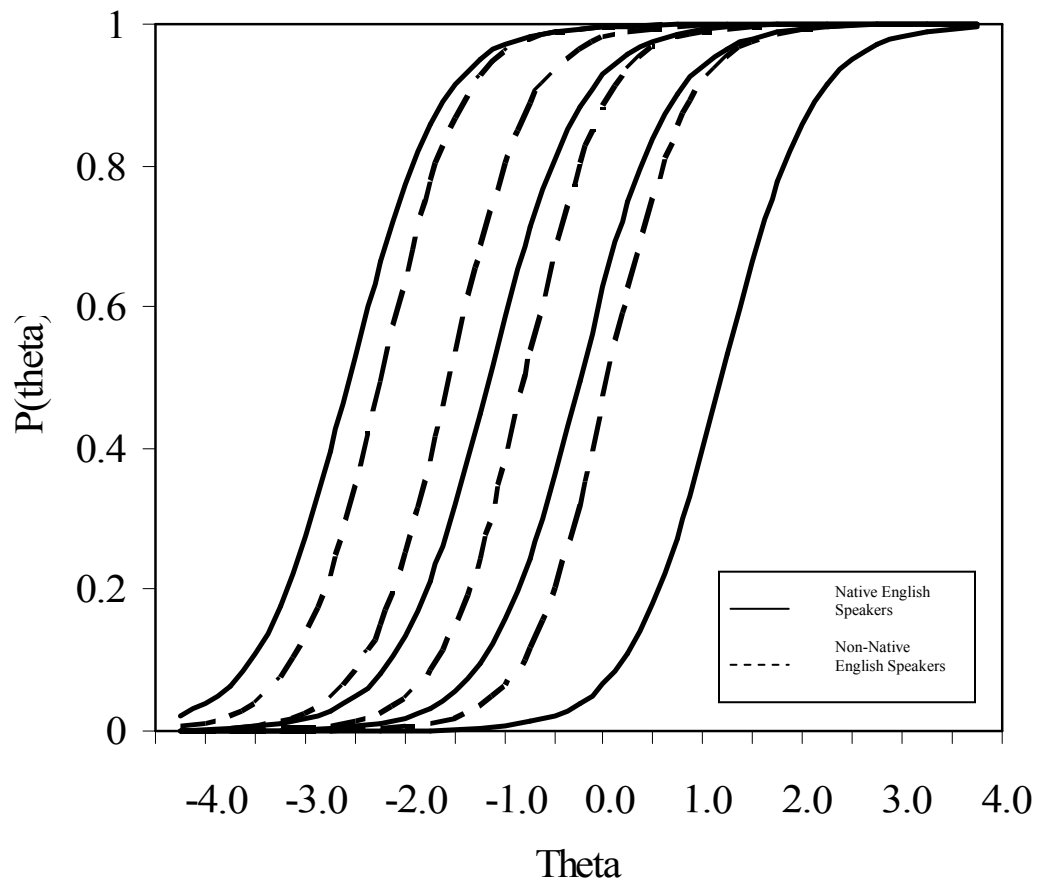
Figure 3. Boundary response functions for Item 9 on the Emotional Stability Scale.



Note. Equated item parameters for the Native English Speakers: $a = 1.295$, $b_1 = -3.824$, $b_2 = -2.189$, $b_3 = -0.722$, $b_4 = 1.819$. Item parameters for the Non-Native English Speakers: $a = 1.888$, $b_1 = -2.768$, $b_2 = -2.049$, $b_3 = -1.217$, $b_4 = 0.274$.



Note. Equated item parameters for the Native English Speakers: $a = 1.058$, $b_1 = -2.868$, $b_2 = -1.487$, $b_3 = -0.098$, $b_4 = 1.917$. Item parameters for the Non-Native English Speakers: $a = 0.290$, $b_1 = -5.642$, $b_2 = -2.438$, $b_3 = 2.305$, $b_4 = 7.372$.



Note. Equated item parameters for the Native English Speakers: $a = 2.232$, $b_1 = -2.308$, $b_2 = -0.908$, $b_3 = 0.010$, $b_4 = 1.437$. Item parameters for the Non-Native English Speakers: $a = 2.531$, $b_1 = -1.994$, $b_2 = -1.311$, $b_3 = -0.556$, $b_4 = 0.299$.

ABOUT SWA CONSULTING INC.

SWA Consulting Inc. (formerly Surface, Ward, and Associates) provides analytics and evidence-based solutions for clients using the principles and methods of industrial/organizational (I/O) psychology. Since 1997, SWA has advised and assisted corporate, non-profit and governmental clients on:

- Training and development
- Performance measurement and management
- Organizational effectiveness
- Test development and validation
- Program/training evaluation
- Work/job analysis
- Needs assessment
- Selection system design
- Study and analysis related to human capital issues
- Metric development and data collection
- Advanced data analysis

One specific practice area is analytics, research, and consulting on foreign language and culture in work contexts. In this area, SWA has conducted numerous projects, including language assessment validation and psychometric research; evaluations of language training, training tools, and job aids; language and culture focused needs assessments and job analysis; and advanced analysis of language research data.

Based in Raleigh, NC, and led by Drs. Eric A. Surface and Stephen J. Ward, SWA now employs close to twenty I/O professionals at the masters and PhD levels. SWA professionals are committed to providing clients the best data and analysis with which to make solid data-driven decisions. Taking a scientist-practitioner perspective, SWA professionals conduct model-based, evidence-driven research and consulting to provide the best answers and solutions to enhance our clients' mission and business objectives. SWA has competencies in measurement, data collection, analytics, data modeling, systematic reviews, validation, and evaluation.

For more information about SWA, our projects, and our capabilities, please visit our website (www.swa-consulting.com) or contact Dr. Eric A. Surface (esurface@swa-consulting.com) or Dr. Stephen J. Ward (sward@swa-consulting.com).