Making Cognitive Ability Selection Tests InDIFferent Across Countries: The Role of Translation vs. National Culture in Measurement Equivalence

Michal F. Gradshtein and Alan D. Mead

Illinois Institute of Technology

Robert E. Gibby

The Procter and Gamble Company

Author Note

Michal F. Gradshtein and Alan D. Mead, Institute of Psychology, Illinois Institute of Technology. We would like to thank Roya Ayman and Scott Morris for providing insightful comments and Rodney McCloy for his support and encouragement in this study. We would also like to thank Susan Nordquist-Mead and Shujaat Ahmed for their help in preparing this manuscript. Correspondence concerning this article should be addressed to Michal F. Gradshtein, Institute of Psychology, Illinois Institute of Technology, 3105 South Dearborn, Chicago, IL 60616. E-mail: chalinka@gmail.com

Running head: EFFECTS OF TRANSLATION VS. CULTURE ON DIF

Abstract

When using organizational measures globally, it is important to ensure measurement equivalence across different language and cultural groups. Although methodology for measurement equivalence has been studied extensively, considerably less research has been conducted as to why these differences occur and such research in the organizational literature has largely ignored cognitive ability tests. The current study assessed the separate roles of translation and national culture in causing measurement non-equivalence of a cognitive ability test used globally for selection. Translation and national culture were examined by conducting dyadic DFIT analyses. The design of the study allowed for separation of the effects of translation and culture by utilizing two samples (995 American taking the test in English and 581 Thai nationals taking the test in Thai) which differ from a third group (366 Thai nationals taking the test in English) by only one variable, national culture or language. Our findings suggest that previous discussion about the role of translation for measurement non-equivalence in cognitive ability tests might be overemphasized as compared to the role of national culture.

Keywords: Measurement Equivalence; DIF; DFIT; Translation; National Culture

Making Cognitive Ability Selection Tests InDIFferent Across Countries: The Role of Translation vs. National Culture in Measurement Equivalence

Organizations employ numerous measurement tools in a global manner. Employee attitudinal surveys, performance appraisals, and selection tests are some examples of such measures. The use of these measurements globally requires organizations to consider potential differences in the statistical properties of scores for all participating groups in order to ensure that conclusions drawn from test scores are accurate for all participating groups. When test scores have the same meaning across groups (so that differences in scores represent true differences, rather than measurement artifacts) we can say that the measures are equivalent for the specific groups compared. Despite the importance of measurement equivalence (MEQ) to multinational organizations, very little research is done in a cross cultural manner. In fact, Gelfand, Raver, and Ehrhart (2002) examined 15 years of I/O research and found only 1% of articles on personnel selection were cross-cultural in nature. In addition, much of the existing research focuses on *whether* differences occur, and *how* to assess both real and statistical differences. However, considerably less research touches on *why* these differences occur (especially in regard to differences in statistical properties).

One of the main barriers to such research is the paucity of theory relating cross-cultural characteristics and statistical properties of selection tests (Gierl & Khaliq, 2001). When cross-cultural measurement research is done in organizations, the MEQ hypotheses are chosen to be tested based upon available theories. For example, Ryan, Horvath, Ployhart, Schmitt, and Slade (2000) tested DIF-related hypotheses based on Hofstede's (1991) dimensions using attitudinal survey data regarding supervisory behaviors. Similarly, Liu, Borg, and Spector (2004) examined

job satisfaction surveys in four cultural groups using Schwartz's (1999) framework of work attitudes.

While performance on attitudinal measures regarding the supervisor or the work itself can be related to existing cross-cultural theories, performance on cognitive ability tests cannot be easily predicted using these same theories. For example, it is hard to draw specific hypotheses using Hofstede's widely used cultural dimensions. Indeed, much of the research on factors affecting MEQ of cognitive ability tests employs SMEs trying to pinpoint the causes of measurement non-equivalence (MNE) rather than hypothesizing about possible factors using a constructed theory (e.g., Elosua & López-Jaúregui, 2007; Yildirim & Berberoglu, 2009).

Overall and regardless of the construct measured by the test, cross-cultural research examining MEQ finds two main factors to play a role in MNE, namely translation and culture. However, two main problems exist. First, translation and culture are often confounded in that culture is looked upon as an aspect of translation (e.g., semantic meaning of translated words) or in that the design does not allow for separating effects of translation and culture. Second and specifically with cognitive ability tests, there is no understanding of the specifics in the culture that might cause these differences. While the second problem is beyond the scope of the current research, the first problem will be addressed in addition to two other issues in cross-cultural MEQ research.

In sum, the current study has two goals. First, translation and national culture are looked at separately as possible factors causing differences in the statistical properties of items (i.e., causes of MNE). Second, as much of the research on cognitive ability tests is conducted by educational psychologists in a non-work context, the current study examines a cognitive ability test used globally for selection purposes by a multinational organization. As no specific theory

exists for measurement equivalence of cognitive ability tests, and current cross-cultural theories do not easily relate to performance on cognitive ability tests, the current study is exploratory in nature and aimed at expanding our understanding of the issue. The next section will discuss MEQ and then culture and translation as they relate to MEQ.

Measurement Equivalence

Measurement can be defined as the assignment of numerical values to individuals in some systematic way in order to represent some properties of these individuals (Allen & Yen, 1979). Measurement equivalence (MEQ) touches on the comparability of the *psychological meaning* of these numerical assignments across groups. In essence, MEQ assures organizations and researchers that the relationship between the latent construct measured and the observed raw score is the same for all groups (Raju, Laffitte, and Byrne, 2002). Drasgow & Kanfer (1985) argue that "without equivalent measurement, observed scores from different groups are in different scales and, therefore, are not directly comparable." (p. 662). Equivalent measures, however, will produce the same expected score for individuals obtaining the same level on the underlying construct independent of the subpopulation to which they belong. Thus, only when MEQ is achieved can meaningful inferences be made about the numerical values obtained by different groups.

Because the ability to make comparisons across groups is of high importance to both organizations and researchers, two main questions need to be asked. One concerns the different methods to establish measurement equivalence, and the second pertains to the factors causing MNE. While the first question has been thoroughly investigated, we were unable to find any literature on the reasons for MNE in cognitive ability selection tests. In a review of MEQ in cross-cultural research, Van de Vijver and Tanzer (1997) stated "the sources of bias in cross-

cultural assessment are manifold and it is virtually impossible to present an exhaustive overview" (p. 267). However, the authors do make an attempt to identify typical sources of item bias and conclude that three main such sources arise: item translation; nuisance factor (e.g., familiarity with item content), and cultural specifics (e.g., connotative meaning). Of these three factors, much of researchers' attention has been given to translation and culture.

Culture effects on MEQ

Culture is a broad term which can be defined in numerous ways. One of the most common definitions of culture was presented by Hofstede (2001). Hofstede defines culture as "the collective programming of the mind that distinguishes the members of one group or category of people from another" (p. 9). These patterns are manifested in four main ways (i.e., symbols, heroes, rituals, and values) with 'values' being the most deep and meaningful. Hofstede (1991) defines values as the tendencies towards these patterns he described as culture. Triandis (1996) provides a more anthropological definition to culture as the man-made elements of our environment that are shared and which "provide the standards for perceiving, believing, evaluating, communicating, and acting among those who share a language, a historic period, and a geographic location" (p. 408).

The operationalization of culture can take two main forms – one is through measuring the values themselves or by defining the groups' boundaries and assuming cultural differences due to the group membership. The boundary itself however is not defined in an absolute way and can take many forms. Hofstede (1991) discussed the layers of culture and indicated that boundaries can be defined on numerous levels such as country, region, religion, gender, generation, etc. Moreover, Hofstede indicated that in global research, the use of geographical boundaries of country as an operationalization of culture is the most common. For example, Hofstede's

cultural dimensions, the GLOBE cultural dimensions (House, Javidan, Hanges & Dorfman, 2002), and Schwartz's (1999) dimensions were all derived by examining national cultures.

Indirectly, geographical boundaries of country imply a dominant system of language, education, law, politics, and more which contribute to a similar socialization process for all individuals under the same nation (Schwartz, 1999). These similar socialization mechanisms in turn, contribute to the similarities in the mental programming of all citizens within a country (Hosftede, 1991). In addition to the common mental programming citizens might share, using national borders as the operationalization of culture can be beneficial when the end goal is to make inferences at the country level. For example, Yildirim and Berberoglu (2009) wanted to examine MEQ in a math test which is used to make cross-cultural educational comparisons (between Turkey and the US).

Some researchers disagree with equating culture with a nation (e.g., McSweeney, 2002). Certainly, many countries have multiple cultures and cultures may cross borders. However, organizations often establish policies, build facilities, hire personnel, etc. within a country and rarely (never, as far as we know) define operational units based upon some more amorphous cultural group either within a country or across countries. Thus, a multinational company will typically interpret American, Indian or Chinese survey results, regardless of the fact that these countries have diverse cultures within their national boundaries. Therefore, we feel that, given the organizational nature of our research, we are justified in considering national culture (and in so doing, we follow a long tradition of cultural researchers by making this simplification), even as we acknowledge limitations of this perspective.

Cross-cultural research on MEQ conducted by I/O psychologists has formulated hypotheses by using conceptual relations between the measured constructs and cultural

dimensions. For example, Ryan and her colleagues (2000) hypothesized about the relationship between Hofstede's (1991) cultural dimensions and the response patterns to an attitudinal survey on job satisfaction with one's supervisor. The authors indicated that items chosen to be used for the research were guided by their relevancy to the cultural value at hand. Thus, the rationale for the hypotheses was derived from a large body of research relating leadership behaviors and cultural values.

On the other hand, trying to hypothesize the relationship between cultural dimensions and performance on a cognitive ability test is not as direct. For example, the current study examines differences between Thai and American nationals. Looking at the GLOBE research in an attempt to develop specific hypotheses for the specific cultural dimensions that might effect MNE resulted in a dead end. Gupta, Surie, Javidan, and Chhokar, (2002) showed there are three dimensions on which the Thai and US clusters differ. However, only one dimension (in-group collectivism) showed an extreme difference. In-Group Collectivism is defined as "the degree to which individuals express pride, loyalty, and cohesiveness in their organizations or families" (House et al, 2002, p.5). Theoretical relations between collectivism and organizational and leadership behaviors can be made, but we were unable to derive any hypotheses about MNE in our context.

An important conclusion which can be drawn from the above discussion is that in order to hypothesize about the effect of specific cultural values on the responses to items/tests, the relation between the cultural value and the content of the item/test has to be clear. If such relationship is not clear one might need to use subject-matter experts which are familiar with both the culture and test at hand to come up with hypotheses (e.g., Elosua & López-Jaúregui, 2007; Yildirim & Berberoglu, 2009). Thus, if the measure at hand is a numerical reasoning test

(such as the case in the current study) one would need some understanding as to the specifics in the culture that might affect performance on numerical reasoning tests, irrespective of ability, in order to identify specific hypotheses.

Unfortunately, although cognitive ability tests are used extensively by many organizations for selection purposes, we were unable to find any literature about national (psychological) cultural characteristics/dimensions that might affect performance on these tests beyond the effect of real ability differences. For example, Yildirim and Berberoglu (2009) found that familiarity with the cognitive skills measured by a test caused MNE for a Turkish sample, they attributed this to differences in the educational system. However, detailed information about the educational systems in different countries is not readily available to organizations, is not easily compared, and does not have any clear link to any theory of national culture. In addition, using national boundary as the frame in which culture is defined includes more than education (e.g., values), but as of now there is no way to know if these affect performance on cognitive ability tests in general, or numerical ability tests specifically. Thus, the goal of the current research is to explore the role of culture (defined according to geographical boundary) on MNE in cognitive ability tests.

Translation effects on MEQ

Translation seems to be one of the most recognized factors studied in the context of measurement equivalence. This extensive attention given to translation resulted in twenty-two specific guidelines published by the International Test Commission (ITC, 1992). Prior to these guidelines "the test translation and adaptation process appeared to be incomplete … and there was substantial evidence that current practices were far from ideal" (Hambleton, 2001, p.164). These practices mentioned by Hambleton include the use of a single translator, use of

translation-back translation (only) designs, and the use of bilinguals as SMEs. Each of these practices has serious flaws, and thus there is no such thing as a perfect translation. For example, bilingual SME's might be different from monolingual individuals on important aspects resulting in translation which might not relate equally to monolinguals. So, while ITC's guidelines should be followed, they are not a guarantee of measurement equivalence (Hambleton, 2001). In order to ensure measurement equivalence, the translated measure must be empirically compared to the original measure.

The specific ways in which translation will affect MEQ is still unclear. Much of the research on the effect of translation on MEQ includes a comparison among groups that differ not only on language but also on culture. For example, Allalouf, Hambelton, and Sireci (1999) analyzed DIF across two languages of the Israeli Psychometric Entrance Test (PET): Hebrew and a Russian translation. These groups, however, did not just differ on the language but also represented two separate cultures. In fact, the authors themselves indicate "since 1990, approximately one million people have immigrated to Israel from the former Soviet Union" (p.188). This statement, along with the timing of the research, demonstrates that it is highly likely that the Hebrew group was comprised of mostly native Israelis, while the Russian group was comprised mainly of Russian immigrants. Thus, culture was a confounding variable to translation. Trying to provide some guidance to research on translation, the authors suggested that translation should be examined in a similar way to culture so that instead of ethnicity or gender, language would be defined as the grouping variable. The current study attempts to test this idea using groups distinct in their culture and language. The next section will discuss the methodology we used and then describe the current study.

Assessing MEQ

Two of the most common approaches for assessing MEQ are Structural Equations Modeling (SEM) and Item Response Theory (IRT) Differential Item Functioning (DIF). Comparisons of these two approaches (e.g., Raju et al., 2002) have suggested that there are many similarities between the two methods and concluded that neither method is necessarily to be preferred under all conditions. However, IRT DIF is considered more appropriate for dichotomous data. A more recent study (Stark, Cherneyshenko, & Drasgow, 2006) showed that when these two methods were implemented in very similar ways, they reached very similar conclusions. This research also showed that IRT DIF was preferable for dichotomous data. Because our data were correct and incorrect responses to a selection test (and thus dichotomous), we chose an IRT DIF method for the current study.

The IRT DIF method allows for the separation between the effects of item characteristics and true ability on the response pattern. In IRT DIF terms 'impact' is the difference in (true) ability between groups. IRT DIF procedures remove the effect of impact before assessing DIF thus allowing researchers to examine MEQ without assuming similar ability distributions in all groups. In two group analyses, one group is termed the *reference* group and the other the *focal* group; impact is usually removed by equating the focal group metric to that of the reference group prior to assessing DIF.

Using IRT terminology, MEQ can be defined as the lack of DIF. DIF is said to occur for an item if the probability of individuals with the same ability to succeed on that item differs for individuals from different groups (Raju & Ellis, 2002). In other words, group membership should not have an effect on the probability of obtaining a correct response, beyond the effects of true differences in ability. There are many ways to asses DIF; we used a relatively recent method proposed by Raju and his colleages (Raju, Van der Linden, & Fleer, 1995). This method allows for

differential functioning (DF) at both the item (i.e., DIF) and the test level (termed Differential Test Functioning or DTF).

Differential Item and Test Functioning (DFIT)

The DFIT method (Raju et al, 1995; Raju, Fortmann-Johnson, Kim, Morris, Nering, & Oshima, 2009) relies on the comparison between the expected test score if the individual belongs to the reference or focal group. These scores are derived from IRT item parameter estimates (that have been linked, so they are on comparable scales). Two conditional probabilities are computed for each of the individuals for each of the items – once using the reference group parameters ($P_{iR}(\theta)$) and then using the focal group parameters ($P_{iF}(\theta)$). The difference between these two probabilities is denoted as d_i. Moreover, these item level probabilities are then used to compute two test-level expected number correct scores (i.e., expected test score) – one as if the individual belongs to the reference group (T_{sF}), and the other as he/she belongs to the focal group. If T_{sF} - $T_{sR} \neq 0$, group membership is said to have an effect on test performance (i.e., the test exhibits DTF).

Including a test level measure of differential functioning can be highly valuable in practice as in many cases only overall test scores are used to guide decision making (Drasgow, CITE). Defining DF at the test level (i.e., DTF) created in turn two separate approaches to DIF at the item level: compensatory DIF (CDIF) and non-compensatory DIF (NCDIF). Non-compensatory DIF (NCDIF) is derived with the assumption that other items on the test are bias free and has been shown to be similar to previous DIF statistics such as Lord's chi-square (Raju et al, 1995). However, in most cases this assumption about lack of bias in other items on the test is an unreasonable assumption and one of the main benefits of using the DFIT framework is CDIF. The computation of CDIF "begins with a definition of DTF and then decomposes DTF

into differential functioning at the item level (CDIF)" (Raju et al, 1995, p. 355). This decomposition leaves us with an estimate of DIF at the item level which encompasses DIF present in other items, which may cancel to some degree with the current item. The CDIF values sum to DTF. Therefore, rather than trying to remove all DIF items, DFIT encourages researchers to first assess the statistical significance of DTF. If the value is significant, then the item with the largest CDIF is removed and all statistics are recalculated. This procedure is repeated until DTF is not significant. In this way, test scores can be made to measure equally across groups while removing a minimum of items from the scale (Raju & Ellis, 2002).

To summarize, strengths of this method include the ability to assess differential functioning for both items and tests, the simplicity and obvious importance of the terms being compared (expected scores on items and tests), the use of both CDIF and NCDIF, and the fact that DFIT provides effect size estimates in addition to statistical significance. Complete details about the DFIT framework and "IPR" Monte-Carlo method of determining statistical significance can be found in the DFIT literature (see Raju et al, 1995; Raju et al., 2009; the IPR method is described in Oshima, Raju, & Nanda, 2006).

Current study

The goal of the current research is to better understand the effects of translation and national culture on MEQ. In order to ensure the effects of translation and culture are independent, three separate samples were used: American nationals taking the test in English; Thai nationals taking the test in English; and Thai nationals taking the test in Thai. For the current research, the Thai-English group was chosen as the reference group because it is most similar to the other two groups; the Thai-English group differs from the US group only on

national culture and it differs from the Thai-Thai group only on translation. This separation between translation and national culture allows the investigation of two research questions:

Research Question 1: What is the effect of culture on the magnitude of DIF exhibited by numerical reasoning ability items while controlling for translation?

Research Question 2: What is the effect of translation on the magnitude of DIF exhibited by numerical reasoning ability items while controlling for culture?

Method

Samples

The data set analyzed in the current study was composed of archival applicant data from a large multinational organization headquartered in the United States. The data included item-level test responses for applicants to managerial, office administrative, and lab researcher positions between the years of 1991 and 1998. Applicants for the relevant positions chose the language in which they could best read from a catalog of 41 languages and were given the test in their chosen language. The focal group was comprised of applicants residing in Thailand who chose to take the test in English (Thai-English; N=366). The two reference groups were Thai residents who chose to take the test in Thai (presumably their native language; Thai-Thai, N=581), and American residents who chose to take the test in English (US; N=955). While it is possible that individuals were not citizens of the country in which they took the test, employment often relates to citizenship and thus it can be expected that most of the individuals in the US and Thailand were citizens of these two countries.

No additional demographic information was available (data such as gender, educational background and socio-economic status were removed to comply with privacy policies). However, all applicants were applying for similar jobs, so it can be expected that applicants are

rather similar in regard to educational background. Also, the personal reasons behind individuals' choice of language are also not clear. For example, it is possible that applicants from Thailand thought taking the test in English would impress their potential US-based employer, but there is no way to evaluate these possible reasons.

Originally the US sample had 16,127 examinees. However, in order to maintain similar sample sizes, a random sample of 1,000 individuals was drawn from the original US dataset. After randomly selecting 1,000 individuals, five individuals who tested in a language other than English were deleted; thus, the final US sample had 995 individuals. Unfortunately only 24 Americans (out of the 16,127) chose to take the test in Thai and so a fourth group was not available for the current study.

Instrument

The 11-item numerical reasoning test examined in the current study was one of three components of a 50-item cognitive test that also included data interpretation and paragraph comprehension items. The test was developed for selection of applicants globally to positions in a large multinational organization based in the United States. The specific test form used for this research was administered by the company to all managerial, office administrative, and lab researcher candidates from 1991 to 1998.

For the present study, only the numerical reasoning subtest was available to study. The test was comprised of eleven items and was administered in a paper and pencil form in different test locations. The items were structured with the intent that they increase in difficulty from easiest to hardest within the test. Questions in the test were all multiple choice questions with 5 response options. An example of the kind of item would be: "A restaurant orders 20 cases of soup at $10 per case. The first 10 cases were sold at $17 each. The remainder were marked

down to 10% less than cost and sold at that price. How did proceeds from sale of soup compare with their cost?" and this stem was presented with five numerical dollar-amount response options.

The test was developed by US item writers in English with consideration of broad cultural differences between the US and the rest of the world. For example, the metric system was used in the questions, so 'liters' instead of 'ounces', and 'meters' instead of 'miles' were used (currency was presented in dollars). In addition, items flagged in subsequent cultural review were either eliminated or altered. Besides the cultural review, the test was also translated into more than 35 languages through a process of using local translation experts from a 3rd party vendor combined with an internal review of the translation by local employees fluent in English and the translated language.

Procedure

Tests were administered to individuals as part of a job application process. The paper and pencil test forms were completed under supervised conditions by trained administrators using a standard protocol. Each candidate received 65 minutes to complete the entire 50-item test. Upon arriving at the testing facility, individuals were informed of their ability to take the test in the language they read best. Most of the US sample chose to answer the test in English (and only Americans responding in English were included in the US sample). With the Thai sample, while most chose to complete the Thai version of the test (N=588), a number of Thai chose to respond to an English version of the test (N=366).

Results

Descriptive statistics using classical test theory were first conducted using BILOG – MG 3.0 (Zimowski, Muraki, Mislevy, and Bock, 1996) to ensure that there were no problems

with the response data. As shown in Table 1, no problems were detected except that Item 8 was very difficult and had poor item-total correlations. The low item-total correlation may well be attributable to the difficulty of the item; the low item-total values were similar for all three language groups, and were not due to mis-keying. Further investigation of the item revealed that item 8 is the only item in which the correct response option is the one specifying there is no way to answer the question. This unique feature of item 8 had additional implications for the IRT and DIF analyses which will be discussed later.

In most cases, the US sample demonstrated higher probabilities of success as compared to both Thai samples. However, two items (3 and 6) showed similar probabilities of success for all three groups. Interestingly, in most cases, the second group most likely to succeed was the Thai-Thai group who took the translated version of the test rather than the Thai-English group who took the test in its original language. This pattern appeared at the scale level, with the US sample having the highest mean (7.31), followed by the Thai-Thai sample (mean of 6.04), and the Thai-English sample (mean of 5.73). An ANOVA ($F(2) = 90.253$, $p < 0.001$) and post-hoc analyses showed that the US scored significantly higher than both Thai groups ($p<0.01$) but the two Thai groups were not significantly different from each other.

IRT based analyses

Prior to IRT scaling, we conducted Lord's (1980) heuristic check of the dimensionality of the numerical reasoning data. We analyzed the combined data of all three groups using principle axis factoring followed by an oblique rotation (promax, kappa=3). The resulting scree plot had a distinct elbow between the first and second eigenvalues and the first eigenvalue was about 1.5 times the second eigenvalue. We interpreted this as adequate unidimensionality to proceed with the IRT analyses (details of this analysis are available from the authors).

The IRT 3PL model was fit to each item independently for each group using BILOG-MG 3.0. We used BILOG's default parameters, except 31 quardrature points were used and the EM and Newton cycles were set to 25 and 10 respectively. Also, when estimating the item parameters for the Thai-Thai group, we were forced to include a weak prior on the location parameter (M=0, SD=2), and so we included that configuration on all analyses. We allowed BILOG to impose its log normal default prior distribution for the slope parameters. The resulting IRT item parameter estimates are presented in Table 2. In addition to item parameters, the MAP (Bayes modal) theta-hat values were calculated using BILOG with default configuration. For both item parameters and theta-hat estimations, the iterative estimation processes converged.

In order to evaluate model fit, BILOG "fit plots" were requested. These plots show the IRT item characteristic curve with confidence intervals overplotted with the actual proportion correct within bands defined at different theta levels. Visual analysis of the fit plots for all items and groups revealed that actual proportion correct values generally fell within the confidence intervals and residuals were distributed without any apparent pattern that might indicate non-logisticity or other systematic misfit.

Because BILOG-MG imposed an arbitrary metric in each analysis, IRT parameter estimates must be translated to a common metric before different groups' estimates can be compared. To link the focal group metrics to that of the reference group, we used Baker's (1993) implementation of the Stocking-Lord (1983) Test Characteristic Curve (TCC) linking algorithm. In essence, the algorithm uses the item parameters to produce linear transformation coefficients *A* and *K* which minimize the differences between the reference and focal group TCCs. The *A* coefficient reflects the degree to which the focal group (US or Thai-Thai) theta distribution variability should be adjusted to match that of the reference group (Thai-English). The closer *A* is

to 1, the more similar the two groups in their underlying ability. The *K* estimate reflects the real mean differences between the groups on the underlying ability. Smaller *K* means smaller real group differences.

Following Candell and Drasgow (1988), we used an iterative linking process: In each cycle the *A* and *K* estimates generated in the previous cycle were used to link metrics and *all* items were screened for DIF. Linking was then repeated without these DIF items and again *all* items were screened for DIF. This procedure was repeated until a stable set of DIF items is observed. The *A* and *K* coefficients of the final iteration represent real group differences (i.e., unaffected by DIF) and were used to link metrics for the final DFIT analysis. In the current study, two linking processes were undertaken – one for linking the US parameters to the Thai-English group, and the second to link the Thai-Thai group to the Thai-English group. Linking metrics to that of the Thai-English group allowed us to compare all three groups' parameter estimates directly. The iterative cycles needed for both linkage processes, including the *A* and *K* parameters and items flagged as containing DIF, are presented in Table 3.

With both linking processes, item 8 prevented the linking analysis from reaching a reasonable solution. Including item 8 resulted in most items being flagged as DIF for both comparisons (US vs. Thai-English and Thai-Thai vs. Thai-English). However, item 8 was not flagged as DIF for either comparison. Thus, an attempt was made to link the groups omitting item 8 from the analysis and indeed, a definite solution emerged in relatively few iterations. Item 8 was included in the DIF analyses.

When linking the US scale to the Thai-English scale, four iterations were required (a fifth iteration produced results identical to the fourth). As can be seen in Table 3, items 4, 5, and 6 were flagged as DIF and so were omitted from the second cycle to produce new *A* and *K*

coefficients. In analysis of the Thai-Thai data, only 2 iterations were required to reach a final linking solution. As shown in Table 3**,** the *K* coefficients from the linking analyses demonstrate that there are real differences in numerical reasoning ability between the US sample and the Thai-English sample ($K = 0.55$), but not as much difference between the two Thai samples ($K = 0.14$). The linked IRT item parameters are presented in Table 4.

DIF analyses

In order to answer the research questions two DFIT analyses were conducted using Raju's (2005) DFITD7 program. The first analysis compared the Thai-English sample to the US English sample and the second compared the two Thai samples. Table 5 presents NCDIF statistics and mean *d*s (i.e., effect size for the NCDIF statistic) associated with each of the items, in each of the comparisons, as well as the overall DTF, assessed in terms of the items that needed to be eliminated to produce non-significant DTF (using the CDIF values of the items).

The comparison between the Thai-Thai sample and the Thai-English sample can provide some answers with respect to the research question on translation while holding culture constant. This comparison produced only one item which was flagged as NCDIF. In addition, all mean *d*s are of very low magnitude (0.05 was the largest mean *d*). Even when examining item 6 which was flagged as DIF, we can see that the *d* magnitude is small (-0.01). Examining item six's parameters we can see that while the *b* parameter for both Thai groups was similar (-0.18 for the Thai-English sample vs. -0.17 for the Thai-Thai) the *a* parameter for the Thai-Thai group is considerably larger than that for the Thai-English group (1.5 vs. 0.66). This suggests that the translated item was psychometrically more effective (i.e., more discriminating)—probably because it was easier for the Thai examinees to understand. In terms of DTF, no items were

flagged for removal. Thus, the test can be considered as equivalent for both Thai groups. This finding is surprising given the attention devoted to translation as a source of DIF.

To examine the role of national culture on DIF, while holding translation constant, we compared the US and Thai-English groups. After linking the scales, three items were flagged for NCDIF. The magnitude of mean *d*s for these items ranged from -0.1 to -0.23. In addition, 4 items were shown to affect DTF. Three of these items are the same items flagged as NCDIF. Interestingly, although item 8 was not flagged as containing DIF according to NCDIF, it was flagged as contributing to DF at the scale level (e.g, DTF). This is even more interesting given the fact that CDIF has been found to flag fewer items than NCDIF (Ellis & Mead, 2000).

Taking all CDIF and NCDIF items from both the first and second comparison, translation seems to have a smaller effect of DIF (only 1 item was flagged) as compared to cultural differences (4 items flagged). In addition to the number of items flagged, none of the four DIF items from the second comparison were flagged when the two Thai groups were compared. These findings show support to the notion that in this context comparing US and Thai culture, national culture has a meaningful effect on DIF, seemingly more than translation.

In order to shed more light on the first two research questions, the linked parameters of all three groups were also examined. Although the US and Thai-Thai groups' scales were not linked, they are both linked to the same group – the Thai-English group. Thus, comparisons of the linked parameters can be made among all three groups. Consistent with previous findings, DIF flagged items (and most of the other items) demonstrated larger differences between the *b* parameters of the US sample and those of both Thai samples than differences between the *b* parameters of the two Thai samples. In addition it can be seen that in a comparison based on

translation it was the *a* parameter that showed the most change. However, when the comparison was based on culture it was the *b* parameter that showed the most change.

Discussion

The main purpose of the current study was to explore the separate roles of translation and national culture on DIF magnitude in a numerical reasoning selection test. Much attention has been given to translation in measurement research but considerably less research had been devoted to national cultural differences (unless related to translation). However, this study offers several other contributions. Ours is one of the few cross-cultural studies on MEQ in cognitive ability tests and moreover the only one we know of touching specifically on selection tests in a work context. Our methodology (DFIT) allows us to examine MEQ at both the test and item levels and provides an illustration of this method. And finally, the design of our study is unique in that it separates the effects of translation and culture.

The design of the current research allowed for a unique examination of national culture and translation as separate factors. To examine the effects of translation, two Thai samples were compared – one that took the test in English, and the other in Thai. However, both Thai samples belong to the same national culture. The effects of culture were examined when the US sample was compared to the Thai-English sample. While both samples took the test in English (thus, no translation effects could have occurred), they differed on national culture. Overall, findings indicate that cultural differences have a noteworthy effect on DIF.

Effect of Translation on DIF

Multinational organizations invest a substantial amount of money on test translation processes. However, some researchers challenge the notion that translated tests are equivalent tests (e.g., Allalouf et al, 1999, suggested translation should be defined as a grouping variable to

ensure MEQ across translations). The current study investigated the role of translation and culture in DIF and found translation to play some role in DIF but possibly not as big as previously suggested. For example, Elosua and López-Jaúregui (2007) found translation to account for DIF in 63% of the items. However, in the current study only 9% of the items were flagged as NCDIF, and none for CDIF. Moreover, group differences between the two Thai groups were not found to be significant and most of the mean *d*s are of very low magnitude (and for the most part smaller than those produced by the US to Thai English comparison).

While shoddy translations could (obviously) produce DIF, our findings question the primacy of translation as a source of DIF when high-quality translation procedures have been followed. For example, Allalouf et al (1999) present a flowchart of actions to reduce DIF. In this flowchart the first question to be asked is – 'is the translation correct?' The last question is – 'are there differences in cultural relevance?' We suggest that differences in cultural relevance should be a primary concern when trying to prevent DIF. Also, 'differences in cultural relevance' should not be confounded with translation. Our findings support the idea that creating an accurate translation is a task distinct from elimination of differences in cultural relevance.

Another interesting finding is that overall the Thai-Thai group performed better than the Thai-English group. These results might indicate that translation done correctly is not only the same as non-translation in relation to DIF, but that it might be preferred. Perhaps translation by the examinees uses cognitive resources that might otherwise be devoted to solving the problems. Or perhaps examinees are simply overconfident in their ability to quickly produce good translations and so suffer from misunderstandings. Whatever the mechanism, this finding prompts several practical suggestions. First, when language differences are present among applicants, organizations should provide translated versions of selection tests whenever it is

economically feasible. Second, organizations should strongly encourage applicants to complete the test form in their native language. It may be that only Thai versions of this selection test should be made available in Thailand. Finally, our findings suggest that in multilingual regions, organizations should be skeptical about claims of a dominant language. For example, South Africa currently has eleven official languages, although Afrikaans and English were the only official languages until 1994 and are thus widely spoken; testing all South Africans in Afrikaans or English may be economical but may disadvantage many candidates who do not speak one of these as their native language.

Effect of National Culture on DIF

It is extremely challenging to explain performance on cognitive ability tests using current national cultural theories. In fact, we could not locate research conducted on DIF in cross-cultural studies of cognitive ability selection tests. However, the results of the current study demonstrate the importance of conducting such research by illuminating the importance of national culture as a factor of DIF in selection cognitive ability tests; in fact, 36% of the items (4 items overall) were flagged as DIF due to cultural differences. Moreover, the comparison between the US and the Thai-English sample showed that the differences between the groups were not due to translation. An interesting finding is that the item flagged as DIF due to translation showed differences in the *a* parameter (i.e., differences in the factor loading of the items) but not the *b* parameter (i.e., in the difficulty of the item). On the other hand, DIF items based on the cultural comparison demonstrated changes in both parameters but a more extreme change in the *b* parameter. These findings support previous research (Ellis, 1995) contradicting Hulin's (1987) theory which postulates that DIF due to differences in the *b* parameters are related to translation while DIF due to differences in the *a* parameters are related to culture.

Taken together the results from the current study indicate culture may often play a primary role in causing DIF in comparisons across cultures. However, more in-depth conclusions about the cultural aspects/dimensions which might affect DIF remain a focus for future research.

Previous research attempting to reveal such factors demonstrated that prior experiences (Scheuneman & Gerritz, 1990) and familiarity with cognitive skills required for the test (Yildrim & Berberoglu, 2009) are important culture-related DIF factors. It can be assumed that experiences and familiarity with different skills will have a strong relationship with national boundary. For example, according to Schwartz (1999), socialization processes within national borders operate on many levels, such as education. Thus, most individuals within the same national borders will have similar educational experiences and test familiarity. Unfortunately, for most organizations and I/O psychologists, we lack a deep understanding of the educational systems in the different countries in which the organization operates. More research can be done to better understand what aspect of the educational systems in different countries contribute to DIF on cognitive ability tests.

Another issue was our use of 'imposed-etic' measures. 'Emic' and 'etic' are concepts borrowed from anthropologists and can be used to describe the cultural relevance of constructs or measures. 'Emic' represents a construct/measure that is culture specific (i.e. meaningful to members of the culture). 'Etic' on the other hand is a construct/measure which is assumed to apply to all cultures (Hui & Triandis, 1985). When a measure is relevant to one culture (i.e. 'emic') but used in a cross-cultural way as if it was 'etic' it is said to be 'imposed-etic' (Ayman, 2004). Gelfand et al (2002) discuss methodological issues in cross-cultural research and indicate that a major problem in cross-cultural research is the use of 'imposed etic' constructs. In fact, Ayman & Korabik (2008) suggest that taking an 'imposed-etic' approach might lead to

measurement-inequivalence. In essence, by using an 'imposed-etic' measure, the current study explores Ayman and Korabik's (2008) assertion.

The results of the current study could be interpreted as indicating the use of an 'imposed etic' measure lead to DIF. However, such inference requires caution as the nature of the test (i.e., 'emic', 'etic', or 'imposed etic') was not tested but assumed due to the method of test development (developed in one place and imposed on another). In addition, the use of only one test did not allow for a real comparison among different types of tests. Thus, a question still remains as for the possibility of obtaining MEQ when an 'imposed etic' approach is taken and additional research is needed.

Limitations and Future Directions

The current study has several limitations that should be considered when results are interpreted. The main limitation was the small number of items studied, which has several implications. First, some statistical approaches could not have been undertaken to explore the research questions presented. For example, it could have been beneficial to conduct a regression analysis trying to predict DIF magnitude by item characteristics. However, conducting such analyses with a sample size of 11 items would not have yielded meaningful results. Second, the number of items flagged as DIF was limited and so only a limited understanding of patterns could be observed. If the test had been longer, more items could have been flagged as DIF or at least we would have had more mean *d*s showing higher variability. Third, the level of confidence we have in the estimates of item parameters and linking coefficients would have increased if more items were available.

A second limitation is the fact that many aspects of the design were not controlled for as the data was extracted from company's archives. For example, Thai applicants chose the

language in which they take the test but the researchers have no understanding of the factors which influenced this decision. It could be that applicants thought to impress their potential US based employer or that they were not Thai. Another example is the lack of an American sample taking the test in Thai. Having such a group in the design would have completed the matrix so that the role of translation and culture would have been more fully investigated. More research should be conducted on MEQ while separating the role of culture and translation in a more controlled manner.

Also, it is possible that the role of translation might be larger for other types of tests; the mathematical nature of our test might make it less prone to translation errors. Future research should seek to separate language and culture, as we have done, and test their influence on other kinds of assessments.

A final limitation is that only one test type (numerical reasoning) and two cultures (US vs. Thai) were compared. Analysis of additional test types and samples are necessary to both replicate our findings as well as identify moderators of our findings.

At least some future studies of the causes of DIF should *not* use operational data because all such studies (including ours) depends upon DIF which happens to have been included in the items. Rather, future researchers may well make *a priori* hypotheses about factors causing DIF, then write items implementing various levels of these factors. For example, we found that an item without a correct answer (where the correct answer was "None of the above") caused very odd results, although it was not flagged as DIF. Yildrim and Berberoglu (2009) found that ambiguous numerical words caused DIF. Future studies could include items with and without such features to explicitly test the causality of factors in creating measurement nonequivalence across cultural groups.

Conclusions

A general conclusion from the current study is that more I/O psychologists should immerse themselves in the understanding of MEQ. Currently the literature suffers from under-representation of such research in organizational settings in general, and particularly in the selection context where MEQ should be of high importance to all multinational organizations. In addition, it might be interesting to explore if there are specific domains of cognitive ability in which tests produce less DIF. For example, in the numerical reasoning test used for the current study very few items were flagged as having translation DIF. Thus, using numerical reasoning tests as a measure of cognitive ability might be preferable in cases where the test is translated to numerous languages.

As mentioned before, an additional important line of research emerging from the current study is the research on the effects of using 'imposed etic' tests on DIF. For example, Hui and Triandis (1985) recommend a combined etic-emic approach for the construction of measures. According to the authors three steps should be taken when developing a measure for cross-cultural use. First an 'etic' construct should be recognized. Second, in each culture 'emic' ways of measuring this construct should be developed and validated. Ryan and Tippins (2010) recommend involving *cultural experts* in the developmental stage of the assessment tool. However using an 'etic' approach to test development requires the use of *"construct experts"* from different cultures in that stage. The validation should be done by testing the 'emic' measures in the other cultures (i.e. as 'imposed etic'). Lastly, using the results from the second stage an 'etic' construct/measure can be empirically defined by including items which are valid in all cultures involved. This final test would then be referred to as 'derived etic' (Ayman, 2004). Constructing derived 'etic' measures is not usually practical in organizations. However, taking

such an approach might be beneficial for ensuring levels of measurement other than scalar (e.g. conceptual equivalence). First, as measures are developed 'emicly' in different locations, the diverse educational and cultural backgrounds will be represented. Second, the construct itself will be better defined to match a wide range of cultures. Third, language and examples used in questions will reflect the commonalities in educational background and experiences across cultures.

As it is hard to grasp all possible facets of all culture that will affect DIF (especially in cognitive ability testing), it might be beneficial to examine test construction as a cultural aspect (i.e., imposed etic' or 'derived etic' approach to test construction) rather than evaluating the culture itself. For example, an interesting extension of the current study would be to have the test ('imposed' versus 'derived etic') as the grouping variable for the DIF analysis. In essence, the current study used one test and compared two groups. Thinking about it differently it might be interesting to examine two tests in one or more groups. Such research will allow for better clarity as for the role of test construction approach ('emic' vs. 'etic') as a DIF factor.

Finally, research on translation should expand beyond the 'detrimental' effects of translations to explore its benefits. The current study indicates that it is possible that when translation is done correctly, it might not only prevent DIF, but also reduce it. Most literature on translation discusses translation errors without taking into account the possibility that these errors can still occur in non-translated tests when individuals translate their own tests in their heads. More research comparing different levels of translation quality might be beneficial in shedding more light on this question.

The current exploratory study challenges both researchers and practitioners to better take into consideration cultural effects on measurement equivalence. While translation plays an

important role, and was even shown here to reduce DIF, the results demonstrate that culture plays an even more prominent role. The current research also demonstrates the dire need for more cross-cultural research on selection tests used globally by organizations and suggested several possible lines of research.

References

Allalouf, A., Hambleton, R. K., & Sireci, S. G. (1999). Identifying the causes of DIF in translated verbal items. *Journal of Educational Measurement, 36*(3), 185-198.

Allen, M.J. & Yen, W. M. (1979). *Introduction to Measurement Theory*. Belmont, CA: Wadswonth.

Ayman, R. (2004). Culture and leadership. In Charles Spielberger (chief editor), Encyclopedia of Applied Psychology. (Vol.2. pp. 507-519). San Diego, CA, USA: Elsevier Ltd.

Ayman, R. & Korabik, K. (2008). Leadership: why gender and culture matter. Manuscript submitted for publication.

Baker, F. B. (1993). EQUATE 2.1: *Computer program for equating two metrics in item response theory* [Computer program]. Madison: University of Wisconsin, Laboratory of Experimental Design.

Candell, G. L., & Drasgow, F. (1988). An iterative procedure for linking metrics and assessing item bias in item response theory. *Applied Psychological Measurement, 12*(3), 253-260.

Drasgow, F. & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. *Journal of Applied Psychology, 70*(4), 662-680.

Ellis, B. B. (1995). A partial test of Hulin's psychometric theory of measurement equivalence in translated tests. *European Journal of Psychological Assessment, 11*(3), 184-193.

Ellis, B. B., & Mead, A. D. (2000). Assessment of the measurement equivalence of a Spanish translation of the 16PF questionnaire. *Educational and Psychological Measurement, 60*(5), 787-807.

Elosua, P & López-Jaúregui, A. (2007). Potential sources of differential item functioning in the adaptation of tests. *International Journal of Testing, 7*(1), 39-52.

Gelfand, M. J., Raver, J. L., & Ehrhart, K. H. (2002). Methodological issues in cross-cultural organizational research. In S. G. Rogelberg (Ed.), *Handbook of research methods in industrial and organizational psychology.* London: Blackwell, 216-246.

Gierl, M. J., & Khaliq, S. N. (2001). Identifying sources of differential item and bundle functioning on translated achievement tests: A confirmatory analysis. *Journal of Educational Measurement, 38*(2), 164-187.

Gupta, V., Surie, G., Javidan, M., & Chhokar, J. (2002). Southern Asia cluster: where the old meets the new? *Journal of World Business, 37*, 16-27.

Hambleton, R. K. (2001). The next generation of the ITC test translation and adaptation guidelines. *European Journal of Psychological Assessment, 17*(3), 164-172.

Hofstede, G. (1991). *Cultures and organizations: Software of the mind*. London: McGaraw-Hill.

Hofstede, G. (2001). *Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations* (2nd ed.). Thousand Oaks, CA: SAGE Publications.

House, R., Javidan, M., Hanges, P., & Dorfman, P. (2002). Understanding cultures and implicit leadership theories across the globe: An introduction to project GLOBE. *Journal of World Business, 37*, 3-10.

Hui, H. C., & Triandis, H. C. (1985). Measurement in cross-cultural psychology: A review and comparison of strategies. *Journal of Cross-Cultural Psychology, 16*(2), 131-152.

Liu, C., Borg, I., & Spector, P. E. (2004). Measurement equivalence of the German job satisfaction survey used in a multinational organization: Implications of Schwartz's culture model. *Journal of Applied Psychology, 89*(6), 1070-1082.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

McSweeney, B. (2002). Hofstede's model of national cultural differences and their consequences: A triumph of faith – a failure of analysis. *Human Relations, 55*(1), 89–118.

Oshima, TC and Raju, N.S. and Nanda, A.O. (2006). A new method for assessing the statistical significance in the differential functioning of items and tests (DFIT) framework. *Journal of Educational Measurement, 43*(1), 1-17.

Raju, N. S. (2005). DFITPD7: A FORTRAN program for calculating DIF/DTF [Computer software]. Chicago: Illinois Institute of Technology.

Raju, N. S., & Ellis, B. B. (2002). Differential item and test functioning. In Drasgow, F., & Schmitt, N. (Eds.), *Measuring and analyzing behavior in organizations: Advances in measurement and data analysis* (pp.156-188). San Francisco: Jossey-Bass.

Raju, N. S., Fortmann-Johnson, K. A., Kim, W., Morris, S. B., Nering, M. L., & Oshima, T. C. (2009). The item parameter replication method for detecting differential functioning in the polytomous DFIT framework. *Applied Psychological Measurement, 33*(2), 133-147.

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*(3), 517-529

Raju, N. S., Van der Linden, W., & Fleer, P. (1995). An IRT-based internal measure of test bias with applications for differential item functioning. *Applied Psychological Measurement, 19*(4), 353-368.

Ryan, A. M., Horvath, M., Ployhart, R. E., Schmitt, N., & Slade, L. A. (2000). Hypothesizing differential item functioning in global employee opinion surveys. *Personnel Psychology, 53*(3), 531-562.

Ryan, A. M. & Tippins, N. (2010) Indicators of quality assessment. In J.C. Scott & D.H. Reynolds (Eds.), *Handbook of workplace assessment: Selecting and developing organizational talent*. San Francisco, CA: Jossey Bass

Scheuneman, J. D. & Gerritz, K. (1990). Using differential item functioning procedures to explore sources of item difficulty and group performance characteristics. *Journal of Educational Measurement, 27*(2), 109-131

Schwartz, S. H. (1999). A theory of cultural values and some implications for work. *Applied Psychology: An International Review, 48*(1), 23-47.

Stark, S., Chernyshenko, O. S. & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology, 91*(6), 1292-1306.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*(2), 201-210.

Triandis, H. C. (1996). The psychological measurement of cultural syndromes. *American psychologist, 51*(4), 407-415.

Van de Vijver, F. & Tanzer, N.K. (1997). Bias and equivalence in cross-cultural assessment: overview. *European Review of Applied Psychology, 47*(4), 263-279.

Yildirim & Berberoglu (2009). Judgmental and statistical DIF analyses of the PISA-2003 mathematics literacy items. *International Journal of Testing, 9,* 108-121.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R.D. (1996). BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items (Version 3.0) [computer software]. Chicago: Scientific Software International.

Table 1

CTT analyses: items' means, standard deviations, and item-total correlations

	US			Thai-English			Thai-Thai		
Item	Mean	SD	r*	Mean	SD	r*	Mean	SD	r*
1	0.89	0.31	0.33	0.78	0.42	0.27	0.78	0.41	0.21
2	0.72	0.45	0.43	0.60	0.49	0.36	0.65	0.48	0.42
3	0.91	0.29	0.33	0.90	0.29	0.22	0.92	0.27	0.27
4	0.82	0.38	0.26	0.61	0.49	0.20	0.64	0.48	0.14
5	0.86	0.35	0.36	0.61	0.49	0.38	0.66	0.48	0.43
6	0.65	0.48	0.45	0.63	0.48	0.27	0.67	0.47	0.45
7	0.45	0.50	0.31	0.31	0.46	0.25	0.37	0.48	0.29
8	0.38	0.49	0.18	0.13	0.33	0.05	0.15	0.36	0.05
9	0.61	0.49	0.43	0.53	0.50	0.31	0.51	0.50	0.35
10	0.49	0.50	0.30	0.31	0.46	0.10	0.33	0.47	0.08
11	0.53	0.50	0.47	0.31	0.47	0.30	0.36	0.48	0.35
NR Score	7.31	2.40		5.73	2.16		6.04	2.24	

*Corrected item-total point-biserial correlations

Table 2

Item parameters for all three groups before linking.

	US			Thai-English			Thai-Thai		
Item	a	b	c	a	b	c	a	b	c
1	0.93	-1.63	0.19	0.69	-1.06	0.18	0.55	-1.21	0.21
2	1.00	-0.63	0.14	0.93	-0.13	0.15	1.11	-0.34	0.14
3	1.05	-1.68	0.20	0.68	-2.11	0.20	0.97	-1.85	0.20
4	0.56	-1.58	0.19	0.52	0.01	0.21	0.37	-0.17	0.24
5	0.92	-1.38	0.19	0.88	-0.16	0.15	1.10	-0.32	0.16
6	1.01	-0.39	0.10	0.66	-0.18	0.20	1.54	-0.30	0.17
7	0.72	0.60	0.14	1.03	1.32	0.17	1.31	0.94	0.18
8	0.51	1.63	0.20	0.57	4.02	0.11	0.54	3.93	0.12
9	0.99	-0.21	0.12	0.80	0.32	0.19	0.90	0.35	0.18
10	0.91	0.61	0.23	0.86	1.99	0.23	0.71	2.42	0.27
11	1.31	0.06	0.10	1.00	1.09	0.12	1.17	0.88	0.14

Table 3
Linkage Iterations

	US			Thai-Thai		
Cycle	A	K	Items omitted	A	K	Items omitted
1	1.3615	0.7895	8	1.1037	0.1344	8
2	1.4801	0.6487	4,5,6,8	1.0273	0.1429	6,8
3	1.5361	05837	1,4,5,6,8	--		
4	1.5786	0.5453	1,4,5,8			

Note. The reference metric for analyses was that of the Thai-English speakers. Item 8 was removed from equating, because it was found to produce poor results, even though it was not, itself, flagged as DIF.

Table 4

Item parameters for all three groups after linking.

	US			Thai-English			Thai-Thai		
Item	a	b	c	a	b	c	a	b	c
1	0.59	-2.04	0.19	0.69	-1.06	0.18	0.54	-1.10	0.21
2	0.63	-0.45	0.14	0.93	-0.13	0.15	1.08	-0.21	0.14
3	0.67	-2.11	0.20	0.68	-2.11	0.20	0.95	-1.76	0.20
4	0.35	-1.96	0.19	0.52	0.01	0.21	0.36	-0.03	0.24
5	0.59	-1.64	0.19	0.88	-0.16	0.15	1.07	-0.19	0.16
6	0.64	-0.08	0.10	0.66	-0.18	0.20	1.50	-0.17	0.17
7	0.45	1.49	0.14	1.03	1.32	0.17	1.28	1.11	0.18
8	0.32	3.13	0.20	0.57	4.02	0.11	0.53	4.18	0.12
9	0.63	0.21	0.12	0.8	0.32	0.19	0.87	0.51	0.18
10	0.58	1.50	0.23	0.86	1.99	0.23	0.69	2.63	0.27
11	0.83	0.63	0.10	1.00	1.09	0.12	1.13	1.04	0.14

Table 5
NCDIF and Mean d for both dyadic comparisons with Thai-English

	US		Thai-Thai	
Item	NCDIF	Mean d	NCDIF	Mean d
1	0.013*	-0.10	0.000	0.02
2	0.006	-0.05	0.000	-0.01
3	0.000	0.00	0.000	-0.01
4	0.041**	-0.19	0.001	-0.02
5	0.065**	-0.23	0.000	-0.01
6	0.005	0.07	0.013*	0.00
7	0.006	-0.07	0.001	-0.03
8	0.039	-0.20	0.000	-0.02
9	0.000	0.01	0.002	0.05
10	0.015	-0.09	0.001	-0.02
11	0.011	-0.09	0.000	-0.01

Note. * significant at the 0.05 level , ** significance at the 0.001 level