DETECTING POTENTIAL CULTURAL AND TRANSLATION BIASES IN KUDER® ASSESSMENTS

A Thesis in
Educational Psychology
by
Deborah Goins

© 2014 Deborah Goins

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

May 2014
The thesis of Deborah Goins was reviewed and approved* by the following:

Hoi K. Suen  
Distinguished Professor of Educational Psychology  
Thesis Advisor

Jerry Trusty  
Professor of Education

Rayne Sperling  
Co-Program Coordinator (PC)  
Associate Professor of Educational Psychology

*Signatures are on file in the Graduate School
**ABSTRACT**

As cross-cultural comparisons are becoming more prevalent, there is an increasing need for linguistic and cultural equivalence of original and translated tests. With increasing popularity of Kuder® assessments in career guidance counseling, it is imperative that great efforts and attention are placed on obtaining cultural and linguistic equivalent tests. Recently, Kuder® assessments have been translated from English to Korean. Based on the Holland theory of Career Choice, test developers constructed equivalent tests using their understanding of existing cultural and linguistic difference and back translation method. However, test translation can lead to construct, method, and item bias. In order to ensure valid assessments, both statistical and judgmental methods are utilized to address issues. In addition to context and cultural sensitivity review by experts, statistical methods for item level analysis are critical in detecting potentially biased items for the Kuder® Skills Confidence Assessment (KSCA).

The purpose of the study was to analyze the KSCA for potentially biased items for test fairness in the translated Korean version. Differential Item Functioning (DIF) method using Logistic Regression (LR) Model was effective at controlling between group differences by including multiple traits in the model and the interaction term in its model. DIF analysis using the LR method was implemented on the 56-item KSCA with 6 subscales (Realistic, Investigative, Artistic, Social, Enterprising, and Conventional) to measure career self-efficacy. Results of the analysis revealed that a majority of the items, with the exception of item 10 in the Conventional scale, have no DIF and that they are adequate for measuring career self-efficacy in relation to Holland type tasks among
Korean and the U.S. students. In addition, cultural sensitivity review was implemented to examine item content. Furthermore, results of this study suggested that future validation of the KSCA versions with revised or newly translated items.
# TABLE OF CONTENTS

List of Figures ........................................................................................................... v  
List of Tables ............................................................................................................. vi  
Acknowledgements ................................................................................................... vii  
Chapter 1 Introduction ............................................................................................... 1  
  Literature Review ...................................................................................................... 3  
  Test Translation ......................................................................................................... 3  
  Methods for Psychometric and Cultural Equivalence .............................................. 5  
  Overview of Cultural Differences between the United States and South Korea ...... 11  
Chapter 2 Career Guidance: Holland Theory of Career Choice ................................. 17  
  Overview of Holland Theory of Career Choice ......................................................... 17  
  Overview of Kuder® Assessments ............................................................................ 22  
  Korean Translation Efforts ...................................................................................... 22  
Chapter 3 Test Bias and Differential Item Functioning (DIF) ........................................ 24  
  Test Bias and Fairness .............................................................................................. 24  
  Type of Bias in Test Translation .............................................................................. 27  
  Test Equivalence ....................................................................................................... 30  
  Consequences .......................................................................................................... 31  
  Differential Item Functioning Methods .................................................................. 33  
Chapter 4 Methods .................................................................................................... 42  
  Participants and Procedures ..................................................................................... 43  
  Measure .................................................................................................................. 44  
  Data Analysis .......................................................................................................... 45  
Chapter 5 Results and Discussion ............................................................................. 48  
  Results ..................................................................................................................... 48  
  Discussion ............................................................................................................... 50  
  Conclusions and Future Studies ............................................................................. 52  
References .................................................................................................................. 55  
Appendix ..................................................................................................................... 67  
  Appendix 1. Realistic Scale Differential Item Functioning (DIF) Indexes by the Logistic Regression (LR) Models ................................................................. 67
Appendix 2. Investigative Scale Differential Item Functioning (DIF) Indexes by the Logistic Regression (LR) Models.................................................................68
Appendix 3. Artistic Scale Differential Item Functioning (DIF) Indexes by the Logistic Regression (LR) Models .................................................................69
Appendix 4. Social Scale Differential Item Functioning (DIF) Indexes by the Logistic Regression (LR) Models .................................................................69
Appendix 5. Enterprising Scale Differential Item Functioning (DIF) Indexes by the Logistic Regression (LR) Models.................................................................70
LIST OF FIGURES

Figure 2-1. Hexagonal Model of Holland Theory. ................................................................. 211
LIST OF TABLES

Table 4-1: Ethnic Groups of U.S. Sample .........................................................................................................44

Table 5-1. Total Scale Score Means, Standard Deviation, and Cronbach's α of KSCA (N=806). .........................................................49

Table 5-2. Conventional Scale Differential Item Functioning (DIF) by Logistic Regression (LR) Models (N=806). .........................................................49

Table 5-3. LR Model Parameters for Conventional Item 10. .................................................................49

Table 5-4. Conventional Item 10 Content ..................................................................................................................50
ACKNOWLEDGEMENTS

I would like to express my utmost gratitude to Dr. Hoi K. Suen for his continuing support, guidance, and encouragement throughout this work and my academic career. I also would like to thank my second reader, Dr. Jerry Trusty, for his expertise and insightful comments. Lastly, I extend my appreciation to the Kuder®, Inc. team, Dr. Spencer Niles, and Dr. Juhu Kim for making this work possible. Lastly, I am grateful for my family and my friends who have taken their time to encourage and support me.
Chapter 1

Introduction

Due to the growing phenomena of testing, test translation is becoming more prevalent in cross-cultural comparisons (Van de Vijver & Hambleton, 1996). In the United States alone, schools are becoming increasingly more racially and culturally diverse. Since 2005, almost 50% of U.S. students can be classified as *culturally different* (Ford, 2010). Also, Lee and Koro-Ljungberg (2007) estimated that by the year 2050, no more than 50% of the U.S. population will be of Anglo ancestry. Beyond the U.S., countries worldwide are administering assessments for international comparisons. Math and science tests were translated in more than thirty languages for the Third International Mathematics and Science Study (TIMSS) to administer to eighty-seven countries (National Center for Education Statistics, 2013). Another large-scale assessment used for international comparison is the Programme for International Student Assessment. There were at least 65 countries that participated in the latest PISA administration. With a growing need for cross-cultural comparisons, the need for unbiased translations is essential to obtaining fair tests. If there is bias in translated assessments, any inferences or use of test scores are invalid. In order to build culturally equivalent tests, multiple methods are necessary for valid translation of tests from one language to another.

One system of assessment widely used in career counseling and planning is the Kuder® Career Assessments based on the Holland theory of six personality and environment types (Armstrong, Day, McVay, & Rounds, 2008; Kuder, 2014).
Throughout the history of Kuder® assessments, there have been various theories used in the construction of Kuder® career planning systems (Harris-Bowlsbey, Suen, Niles, & Trusty, 2013). The earlier Kuder® systems used theories by Donald Super, David Tiedeman, and Martin Katz. However, the most recent career planning system is based on the work of Anne Roe's theory of personality development and career choice and John Holland's Holland Theory. Technology has fostered the widespread access and usage of Kuder® assessments on the web.

By providing career assessments in multiple languages, they assist individuals across all ages to explore career options, plan for education, and prepare for job search. The three types of Kuder® assessments are Kuder® Career Interests Assessment (KCIA), Kuder® Work Values Assessment (KWVA), and the Kuder® Skills Confidence Assessment (KSCA). With its growing popularity, Kuder®, Inc. provides global customization services of their assessments. Through language translation and cultural adaptation, Kuder® Assessments are currently being used all over the world, and they are available through various online platforms.

The most recent development of KSCA originated from an initial pool of 175 potential five-point Likert items (Kuder® Skills Assessment-16, KSA-16). KSCA item contents describe particular tasks related to the six Holland types in which respondents rate their level of career self-efficacy on a 5-point Likert scale: Cannot do at all (1), Slightly certain can do (2), Moderately certain can do (3), Very certain can do (4), and Completely certain can do (5). In order to construct a reliable and valid measurement of career self-efficacy relative to the six Holland types, both judgmental and statistical methods were implemented using a national sample of 2,100 respondents. A panel of 5
national experts in career counseling and guidance reviewed the items for elimination or modification (Harris-Bowlsbey, Suen, Niles, & Trusty, 2013). In addition to a cultural sensitivity review, several item level statistical analyses, such as differential item functioning using polytomous logistic regression model with Zumbo-Thomas criterion and factor analyses such as alpha factoring and parallel analyses with oblimin rotation, were implemented to detect potential bias regarding gender and race/ethnicity. Resulting in the 56-item KSCA, findings from judgmental and statistical analyses provided evidence for high levels of reliability, strong evidence for validity based on content, internal structure, freedom from gender or race/ethnic bias, context, content, or language. As Kuder's newest scale development, it has been translated from English to Korean so it is critical that both versions are culturally and linguistically equivalent.

**Literature Review**

In cross cultural test adaptations, various test translation methods and methods for establishing psychometric and cultural equivalence have been used to ensure valid assessments. Not only is it imperative to use a combination of methods to achieve equivalent assessments, but it is also vital to examine the cultures. Understanding cultural differences will provide a framework within which an assessment can be appropriately constructed for a valid translated assessment.

**Test Translation**

Test developers and researchers ought to be sensitive to both linguistic and cultural elements to construct a fair and valid translated instrument. Geisinger (1994)
recommends the following steps in translating a test from one language to another:
translate and adapt the instrument, review and comment on the translated instrument,
modify the instrument based on reviewers’ comments, pilot and field test the instrument,
standardize or equate the scores, conduct appropriate validation research, create a manual
and other documents for test users, train test users, and collect reactions from test users.

Utilizing both statistical and judgmental methods, the most common translation methods
are back translation, forward translation, comparison of parallel texts, and a combination

Back translation consists of at least two independent sets of bilingual translators.

One group translates the test from source to target language. Afterward, the second group
of translators translates from the target language back to the source language. At the end,
they use expert judgment to compare the two versions at the item level. However, back
translation has a few disadvantages (Brislin, 1970). Translators may share the same rules
for translating non-equivalent words and phrases. For instance, “amigo” and “friend” are
used as the standard translated versions of each other despite different underlying
connotations and implications (Suen & Pun, 2012). Secondly, due to some poorly
translated items, some back translators may be able to determine the original version.
Lastly, the translated version may have retained grammar structure of the source
language that allows the translator to replicate the original version. Despite the
disadvantages, back translation is the most commonly used translation method. Expert
judgments such as cultural sensitivity review can supplement such a method to help
reduce potential bias from cross-cultural test translation.
In an attempt to address some of the issues related to back translation, forward translation adds an additional element with pilot and field testing (International Test Commission, 1993). Similar to back translation, forward translation begins with the translation of the test from the source to the target language. However, the translated version is then given to a sample of the target language population. Examinees of the target population are asked to interpret the meaning of each item. Following the test administration, experts or impartial judges review and evaluate the qualitative information. If they conclude that the target population interpreted most of the items consistently with the original meaning of the source language version, the information strengthens the equivalence of the two versions of the test.

Finally, in literal and idiomatic translation, translators compare parallel texts to retain the organic nature of words and phrases while avoiding undesirable literal translations (International Test Commission, 1993). Another advantage of using parallel texts for translation is that it reflects the notion that there is no single way an idea can be represented. Therefore, idiomatic translation is encouraged particularly when merging several texts to create a single text. In this way, the original meaning is preserved, but the product can be more organic. However, pilot testing is also recommended to facilitate revisions to the translated tests particularly when used for international achievement tests.

**Methods for Psychometric and Cultural Equivalence**

In general, there is no standard of which method is the most effective in test translation. The International Test Committee (ITC) guidelines have been widely used for test translation, cross cultural studies, and international comparisons (International Test Commission, 1993; Suen & Pun, 2012). The ITC guidelines provide methods in relation
to the context, test development and adaptation, administration, and documentation or score interpretation of a translated test. ITC recommends that translators and experts work in teams to identify potential linguistic and cultural differences and dialects to preserve meaning. Rating scales for translators may facilitate the process to achieve content, construct, and linguistic equivalence. Test wisdom and environment ought to be considered for the administration of a translated test. There may be differing cultural norms of what is an acceptable speed or time limit of a test. Also, multiple designs or assessment to measure the same construct may increase strengthen overall quality of validity evidence. For example, one can use multiple IQ tests for one student. By reporting different results, a variety of explanation can help facilitate the validation process as external factors surface. Expanding upon the IQ test example, external factors such as qualitative information provided by the student’s teacher may increase validity for placement of the student in an appropriate program. Altogether, decisions about a person on one test score may be based on multiple sources of data.

In addition to the ITC guidelines, test developers may desire a concurrent and iterative development of both the source and target language versions of the test. To simultaneously develop both versions of the test, decentering allows the source language to be modified to match the meaning of the target language with the translation of target language to the source language. Some words or phrases lack or differ in meaning from one culture than another (Suen & Pun, 2012). In such a case, those words or phrases are either removed or revised. The iterative nature of decentering may help reduce linguistic differences while increasing cultural equivalence.
Another important aspect of test translation is the selection and training of translators. Not only should translators be knowledgeable in the language, but they also ought to be familiar of cultures involved in the test translation process (Hambleton & Patsula, 1998). If translators are not familiar with both the source and the target cultures, this can lead to biased translation and inhibit cultural equivalence. Lastly, training the translators on principles of test development can facilitate translation to enhance validity of the translated test. For instance, if a translator is unaware of how to effectively construct a multiple-choice item, the distracters may be translated in a way that they result in similar meaning (Hambleton & Patsula, 1998).

When source and target examinees perform differently on the test, this may be an indication of non-equivalent tests due to biased test translation or cultural differences. In order to identify and achieve psychometric and cultural equivalence, there are both statistical and judgmental methods. Common statistical analyses are differential item functioning (DIF), confirmatory factor analysis (CFA), structural equation modeling (SEM), hierarchical regression analysis, and distracter analyses (van de Vijver & Tanzer, 2004).

DIF is a statistical method that allows one to determine whether the items are functioning in the same way for all subgroups of interest. If there is a DIF item, the item may be biased. Although there is no direct connection between DIF and validity of a test, item analysis using DIF facilitates deeper examination of construct representation (Holland, Wainer, & ETS, 1993). To analyze construct and content equivalence, CFA is helpful in showing if the original and translated tests show similar internal structure (Hui & Triandis, 1985). SEM allows comparison of different models to test if certain factors
lead to varying test responses of subgroups of examinees. Although it may be impossible to include all extraneous factors, Poortinga and van de Vdver (1987) state that including known confounding variables in a study may increase validity of cross-cultural comparisons; hierarchical regression analysis statistically controls confounding variables so that cultural differences are excluded in the analysis. Additionally, when translating a multiple-choice test, developers can intentionally write distracters to identify typical errors of the examinees (van de Vijver & Tanzer, 2004).

Another way to control for individual differences and achieve psychometric and cultural equivalence is the use of experts or impartial judges. Judgmental method involves linguistic, cultural, and psychological analyses. Taking cultural and linguistic background into consideration, judges can examine both the items and study the cultures. When evaluating items, judges can look at the grammatical structure, underlying meanings of words or phrases, item format, test layout and directions, and meaningfulness of the construct and score interpretation. In terms of cultural and psychological analysis, there may be underlying connotations or emotions that certain words evoke for the examinees of the target population that are different than examinees of the source population. Surveys or further study of the target culture can shed light in such cultural differences.

Testing behavior can also affect the validity of translated tests. For instance, some cultures are more reserved so when examinees respond to certain items, a lower score may not indicate low performance (Wang & Mallinckrodt, 2006). Another potential source of error is attitude of examinees. Arvey, Strickland, Drauden, and Martin (1990) found that the Caucasian and African American examinees had different test taking
attitudes. Caucasian examinees were more motivated to work and show efforts during test administration, whereas the African American examinees had a more positive attitude toward test preparation. All in all, judgmental methods require extensive knowledge of both the target language and culture.

Wang and Mallinckrodt (2006) conducted a study to assess the cultural differences between American and Taiwanese about adult attachment. The Experiences in Close Relationships Scale (ECRS) was translated from English to Chinese to assess adult attachment. As a result, a simple translation did not lead to a valid assessment. Western ideas and values were reflected in the original ECRS about what constitutes secure adult attachment. In the United States, avoidance is defined as “fear of intimacy, discomfort with self disclosure, reluctance to seek support or depend on others, and reticence about expressing feelings,” and anxiety is defined as “worries about being alone, concern about a partner’s disapproval, frustration when a partner is not available, and desire to be closely merged with one’s partner” (Brennan, Clark, & Shaver, 1998, pp. 46-76). Such ideas did not have equivalent meaning in the Taiwanese culture because of different values and perspectives.

Consequently, Wang and Mallinckrodt (2006) determined that self-construal was similar to the construct of adult attachment in the Taiwanese culture. Assessing individualistic and collectivistic cultural factors, self-construal consisted of independence and interdependence that relate to avoidance and anxiety dimensions for the Taiwanese examinees. Unlike the United States culture, emotional constraint and interdependence on family are valued in the Taiwanese culture. For instance, a low score on an item stating “I prefer not to show a partner how I feel deep down” has a different interpretation for an
American and a Taiwanese examinee. For an American examinee, a low score on the item may indicate high avoidance, but for a Taiwanese examinee, a low score does not necessarily indicate high avoidance. Along the same lines, a high score on an item stating, “I worry about being alone” may be evidence for high interdependence but may not necessarily mean high anxiety for a Taiwanese examinee.

To demonstrate linguistic and cultural equivalence, test and retest internal reliabilities of a bilingual sample were compared on the original and translated ECRS versions. For criterion-related validity, the study found significant correlations in expected directions with other measures of adult such as preferences for touch and sexual behavior. In terms of construct validity, statistically significant coefficient alpha between same language versus mixed language forms to support linguistic equivalence. Moreover, there were significant score differences between U.S. and Chinese examinees in directions consistent with their cultural characteristics. Qualitative evidence by directly asking examinees the meaning of items may strengthen validity of generalization. If examinees’ interpretations of item meanings are found to be consistent with the original meaning of items, the test results may be generalizable to the target population. In conclusion, regardless of which test translation method is used, test developers ought to take cultural values and linguistic differences into consideration. Teams of researchers, translators, and test developers ought to ensure linguistic and culturally equivalent assessments for valid score interpretations. By striving to be culturally sensitive, test developers can minimize potential for test bias in cross-cultural test adaptation.
Overview of Cultural Differences between the United States and South Korea

Because Kuder® Assessment measures career self-efficacy, interests, and skills, such constructs may stem from underlying value systems. In particular, KSCA was designed to measure career self-efficacy by asking respondents to self-report their levels of confidence completing various tasks. In doing so, content in translated items may not have the same effect or meaning to the respondents from U.S. and Korea resulting in linguistic and cultural inequivalence. Essentially, general differences in values, culture, and lifestyles may have an impact on how individuals interpret and respond to items on the Kuder® Assessments in relation to perceptions of self and social relationships, family, and education or work.

Perceptions of self and social relationships are influenced by the structural dynamic of a society. More specifically, the notions of collectivistic or individualistic society can be used to understand conception of self. In a collectivist society like Korea, there is more value placed on group cooperation and modest individual behavior. On the other hand, greater value is placed on individual freedom and self-reliance in an individualistic society like the U.S. Consequently, there is a tendency for Koreans to give more weight to personal relationships than written laws or procedures when interacting with others whereas written rules are assumed to be fair and applicable to every individual among interactions for Americans. In order to achieve collectivism, even if there may be differing opinions, Koreans strive for harmony in community by avoiding direct confrontation and any topics which may embarrass another individual. Americans, on the contrary, tend to be more willing to confront others directly to express individual opinions and views. Reciprocity is also more heavily emphasized among Koreans to
maintain a continuous sense of community. Finally, Korean culture tends to foster a smaller number of close friendships with a sense of mutual obligation to help each other while American culture allows for a larger number of friends and acquaintances with limited sense of mutual obligations. Moreover, there is a greater emphasis on building and maintaining a harmonious relationship over accomplishing tasks among Koreans (AIEF, 2013).

Family dynamics and expectations reflect the general differences between an egalitarian society of U.S. and hierarchical society of Korea (AIEF 2013; Kim, Pan, & Park, 1998; Lee et al., 2011). In an egalitarian society like U.S., each individual is seen as equal with regard to status and importance; Americans perceive social status as a place that can be achieved. However, in a hierarchical society, each individual accept his or her role in which includes certain expectations of behavior. Furthermore, status is usually based on inherited traits such as age, gender, and family. Although there are general tendencies between Americans and Koreans, family and social dynamics can be more deeply understood by understanding Korea's history.

In Korea, Confucian roots have influenced the way in which people relate to each other and their value systems. Within a hierarchical society, Confucianism values morality, duty, obedience, filial piety, and loyalty to family relationships (Park, 1991). As such values manifest in Korean families, they affect the development of young people's identity and sense of belonging in the context of familial dynamics (Kim, 1997). Duty, respect, and honoring one's family come first before individual desires or goals. Also, Yi, Coale, Choe, Zhihu, and Li (1994) found that children leave their parents' home to live on their own at a later age in East Asian countries versus western countries. In terms of
family dynamics, Lee and Koro-Ljungberg (2007) found that it is common for Korean parents to encourage their children to devote themselves to education and studies. Furthermore, collectivistic attitude of Korean families encourage parents to be more active stakeholders in children's education as they tend to be more involved in their children's decisions than American parents (Tsuya & Bumpass, 2004).

Interestingly, the make-up of Korean and U.S. populations further shape family dynamic and identities. As Korean culture forms within a racially homogenous society, Koreans tend to perceive their identities in relation to their race. Additionally, Lee and Koro-Ljungberg (2007) found that a more racially homogeneous society lead to strong attachment and obligation to one's family. On the other hand, the U.S. is racially heterogeneous with allow Americans to identify more with ethnic groups, based on common cultural values and behavior, than a single race or family.

In addition to the emphasis of filial piety, morality, and obedience to family relationships in Confucian philosophy, the industrialization of Korea led to an increased need to educate and prepare future workers for the industry (Sandefur & Park, 2007). As a result, there was a rapid educational expansion in Korea at primary, secondary, and college (tertiary) levels particularly during the 1980's and 1990's with the economic boom. Especially at the college level, for instance, 40 percent of students who graduated high school continued onto college in 1980 while in 1998, 80 percent of high school graduates went to college (Korean Educational Development Institute [KEDI], 2000).

The educational system in Korea and the U.S. is also different in its attendance requirements. In Korea, primary school is comparable to elementary school in U.S., lower secondary school is comparable to middle school in the U.S., and secondary school
is comparable to high school in the U.S. Only the first six years of school are compulsory for Korean students whereas in the U.S., enrollment in elementary, middle, and high school up to age sixteen is required. With a standardized educational system in Korea, students are sorted into either academic high school or vocational high school at the completion of lower secondary school. Even though it is not as currently prevalent, vocational schools prepare students for working in the labor market industry with specific skill training whereas academic high school students are not trained in occupationally specific skills (Chang, 2001; Sandefur & Park, 2007). Upon high school graduation, students can either attend a two year college to gain further occupational skills or a four year university. In the U.S., tracking for students is more common throughout secondary school and vocational schools as well as four year colleges are available.

Although there are differences in the educational systems between Korea and the United States, there is a common value placed on making connections by attending the few top elite colleges and universities. Value on admission to the most elite colleges affect the level of emphasis placed on exams for college admission. For instance, despite only a few available elite colleges and universities, Koreans expect that such educational attainment leads to connections which are valued to enhance success in the labor market (Lee & Brinton, 1996). Thus, although there may be many other higher education institutions available, there is much pressure and competition to be admitted into the top colleges determined by a high score on a single Korean national entrance exam (Lee, 1999; Sorensen, 1994).

Along the same lines, families in the U.S. also strive for a college admission to the elite schools such as the Ivy League schools. In the U.S., standardized test scores such
as the SAT and ACT play a role in college admission. However, the there are other factors such as high school GPA, extracurricular activities, community service hours, personal statements or essays, and letters of recommendation that can increase the probability of U.S. students to attend college. In essence, there are multiple methods to enter college in the U.S. whereas there is only one way to enter college in Korea through the national entrance exam. As a result, the pressure to score high on the college entrance exams (e.g., SAT or ACT) in the U.S. may not be as prominent as high performance on the national entrance exam in Korea. Moreover, the different level of emphasis on exams for college admission may affect students' lifestyles.

Because college admissions are affected by students' lives and activities outside of academic settings (i.e., community service, sports, etc.), U.S. parents and students may generally devote less time and financial resources on SAT and ACT preparation. Instead, U.S. families and students may devote more time to participate in extracurricular activities, socialize with friends, and gain work experience. On the other hand, a single path to college in Korea through the national entrance exam may limit experiences and activities for Korean students to strictly academic settings.

Private tutoring (hakwon) outside of regular school hours are commonplace among Korean families (Park, Byun, & Kim, 2010). As of 2007, 8 out of 10 students in primary and secondary schools in Korea were enrolled in some form of private tutoring. Consequently, Korean high school graduates usually either attend college directly after graduating high school, enter the labor market, or just spend time preparing for the national entrance exam. The heavy emphasis on educational achievement and competition for college admission affects the lifestyle of Korean students. As Korean
families push for admission to higher ranked universities for connections to succeed in the labor market, value systems and lifestyles of Korean students are also affected in a competitive society. Admission to a higher ranked university is ever more critical because educational opportunities are not as prevalent among older adults (Park, 2013).

Although Confucianism has had a great impact on Korean value system, the impact of traditional Confucian ideology has been changing in the last century (Lee, 1999). With the growth of the economy and industrialization, western and other foreign influences have challenged and influenced traditional Koreans culture. Instead of its traditionally hierarchical society, Korean society and culture has become more similar to U.S. culture as social relationships and individual goals and needs become more socially acceptable and commonplace (Park & Cho, 1995).

Due to general differences in societal, relational, family, and educational dynamics and values, individuals in Korea and the U.S. may have difference experiences. For example, it may be more socially acceptable and commonplace for U.S. high schools to start working as well as attend school to gain work experience because U.S. society also values work experience for college admission. On the other hand, students in Korea have so much pressure to score well on the single way to college that any time outside of regular school hours are committed to private tutoring in preparation for the national entrance exam; hence, they may not have similar amount of social and work experiences as U.S. students upon high school graduation. In such a case, linguistic and cultural equivalence of Kuder® assessments may be even more critical to help Korean students to explore future career opportunities with limited work experiences.
Overall, it is imperative to examine and consider differences in values which influence the lifestyles and experiences of individuals when adapting an assessment across cultures. Specifically, since Kuder® assessments were constructed in the U.S., there are cultural and linguistic elements embedded in the item content and choice of words that are appropriate for individuals from the U.S. However, both traditional and changing cultural values and system are important to consider when adapting Kuder® assessments to Korean.

Chapter 2

Career Guidance: Holland Theory of Career Choice

Overview of Holland Theory of Career Choice

John Holland, a career counselor, made a significant contribution to the field of Career Counseling. In an attempt to effectively help individuals who needed career guidance, the Holland theory explains how personal and environmental characteristics lead to satisfying and dissatisfying career decisions and career changes or retention. Holland addresses how personal and environmental characteristics affect career stability or change and work performance over one’s lifetime. The Holland theory explains that one can understand possible vocational outcomes with the congruence of personality types and environments because the choice of vocation is an expression of personality (Holland, 1997). Based on a particular personality type, Gottfredson and Holland (1996) developed a list of occupations that correspond to each type.
Holland grounds his theory based on four assumptions. Firstly, within a culture, personal preferences, attitudes, and skills develop so that most people can be categorized as one of the six types. Secondly, there are six model environments that correspond to the six personality types; each environment is dominated by a given personality type. Consequently, people with a particular personality type seek out an environment in which their interests, competencies, and perspective are most effectively used. Likewise, the third assumption states that people seek out environments that allow individuals to exert their skills and abilities and express their attitudes and values. Lastly, one's behavior is the outcome of personality to environment interaction.

With two major components, personality and environment, people are characterized by their resemblance of six personality types: Realistic, Investigative, Artistic, Social, Enterprising, and Conventional. The realistic type prefers hands-on activities using objects, tools, and animals. The investigative type prefers activities that consist of investigation of physical, biology, and cultural phenomena. The artistic type prefers unsystematized and ambiguous activities using materials to create art. The social type prefers work to train and care for others in an ordered and systematic way using materials. The enterprising type prefers to work with others to achieve organizational goals or economic gain. Finally, the conventional type prefers to work in an orderly and systematic fashion using data and materials. For each personality type, Holland examined vocational preferences, values and goals, belief and perception of self, and problem solving style. In addition, there are six environments that are consistent in characterization of the personality types.
As Holland began his work with the Vocational Preference Inventory (VPI) in combination with the transmission of the environmental influences through other people, three main principles are used to explain individual vocational and educational choice, vocational stability and achievement, personal competence, social behavior, and susceptibility to influence (Holland, 1958; Linton, 1945). The three main principles are differentiation, congruency (similarity), and consistency (construct).

Differentiation is related to how a person can be meaningfully defined by his or her resemblance to a particular personality or environment type. While a person may have one dominant personality type, all individuals exhibit varying degrees of the six Holland personality types. Thus, a personality pattern numerical score is used to analyze the absolute difference between a person's highest and lowest RIASEC scale scores ranging from 0 to 14. For example, person A may highly resemble a realistic personality type with a score of 14 while having low resemblance to a social personality type with a score of 2. In such a case, the individual's degree of differentiation is 2. On the other hand, person B with a highest score of 7 on the realistic scale and the lowest score on the social scale of 3 has a lower degree of differentiation of 4. Thus, person A is more clearly defined than person B based on his or her dominant personality and environment type. In essence, the degree of differentiation examines the meaningfulness of the difference in personality profile scores. Assuming that the classification of people can be made using the concept of differentiation, Holland exerts that a higher degree of personality and environment match leads to positive outcomes in the predictability of vocation.

Congruency is related to the compatibility or consistency between a person and his or her environment (Chartrand & Walsh, 1999). The fit of a person within a certain
environment is assessed by examining the congruence of his or her personality type to the types of activity preferences within a particular environment. When there is congruence, a person may find stability, satisfaction, and achievement in one's career because people tend to choose a career that align with their set of identities, goals, and values. By seeking out environments with opportunities to engage in certain activities that are suitable to their personalities, they find the environment to be rewarding. Likewise, persons seek out to work with others who are similar to them. As a consequence, a person's dominant personality type drives one's preferences for an environment which predict career choice. While the congruency principle examines the relationship between a person and an environment, the consistency principle describes the degree of relatedness among personality or environmental types within a person or an environment. Statistical analyses are implemented to identify which types have more in common with each other. Higher consistency between personality and environmental types lead to higher predictability for career choice. For example, realistic and investigative persons have more common with each other than realistic and social persons.

Congruency and consistency can be understood by Holland's six type personality hexagonal model (Cole, Whitney, & Holland, 1971; Holland, Whitney, Cole, & Richards, 1969). The distance among the personality and environment types in the model are inversely proportional to their theoretical relationships (Figure 2-1). Consequently, an individual personality pattern is considered to be psychologically consistent if a person's personality types share common traits or closer together in the model. For example, a person with a Realistic and Investigative dominant types has better consistency and congruency than a person with Realistic and Social dominant types.
Because differentiation, congruency, and consistency principles serve as estimates of the level of clarity of a respondent's Holland profile, they are fundamental in understanding the Kuder® assessment. More specifically, the results of Kuder® assessments are based on a Euclidean distance similarity index. As represented in the geometrical hexagonal model, the distances define the level of consistency of a person and an environment as well as the congruence of a person and environment (Holland, 1997). For example, an individual with a Holland profile of Realistic-Investigative-Conventional exhibit a greater degree of differentiation, congruency, and consistency because the distance among the types are less than an individual categorized as Realistic-Investigative-Enterprising. Since the six Holland areas are essentially used for classification purposes, the degree of career changes calculated by the Euclidean similarity index, reflect the distances among the Holland types. Consequently, such results from Kuder® assessments can be used to predict or interpret a person's behavior and activities in relation to the occupations or environments in their corresponding areas.

Figure 2-1. Hexagonal Model of Holland Theory.
Overview of Kuder® Assessments

Most recently, an Atlas model of Occupational Reinforcer Patterns (O*NET) based on Holland Theory has been used to classify occupations (Armstrong, Day, McVay, & Rounds, 2008; McCloy et al., 1999). Based on job and organizational research, the atlas model looks at the relationships among personality types consisting of a person's abilities, values, skills, education, experience, activities, occupational outlook, and occupation specific tasks (National Center for O*NET Development, 2013). O*NET classifies occupations and job titles to form a taxonomy of skills and interests. There are a total of 16 clusters and within each cluster, there are 79 career pathways. Within each pathway, there are numerous occupations.

Specifically, Kuder® Skills Confidence Assessment (KSCA) measures confidence level in performing, or learning to perform, the skills required in various occupations. Created within the framework of the Holland Theory, respondents receive scores for the six Holland areas. KSCA scores are used to match O*NET occupations and 16 National career clusters and pathways for respondents (National Association of State Directors of Career Technical Education Consortium, 2013). Identified occupations are then divided into career clusters which are placed in the corresponding six Holland areas. Thus, the Holland profile from KSCA scores is representative of his or her personality pattern resembling the six Holland personality types showing the top matching career clusters.

Korean Translation Efforts

The Holland Theory gained popularity among the various personality theories because it took complex multi-faceted aspects of personality and environment and
provided a common language. Such an approach is relevant to the way in which the results of Kuder® Assessments are used worldwide. Using a back translation method, Kuder® assessments hired a translation company to translate the KSCA from English to Korean for high school students seeking higher education. Additionally, Dr. Hoi Suen and Dr. Juju Kim implemented a cultural sensitivity review.

Furthermore, the Korean translated version was adapted based on a cultural taxonomy established by UNESCO for educational affairs because the O*NET in the Korean version will not function in the same way. Occupations in United States are different from Korea so they may not fall in the same or either categories as in the English version. To prevent such inaccuracy and strive for cultural equivalency, Dr. Suen and Dr. Kim have changed some items because item contents are not appropriate for the Korean culture versus the American culture by using a database by the labor department in Korea.

In collaboration with the International Centre for Career Development and Public Policy (ICCDPP), Kuder® is an active leader in advancing career development around the world using evidence-based assessment and career guidance tools (Global Career Planning Services, International Support, 2013). The English version of KSCA is generally based on western cultural values whereas the Korean version is based on eastern cultural values. Thus, the Korean version will probably be more applicable to other countries such as countries in the Asia Pacific region. Translation of Kuder® from English to Korean is a significant contribution to Kuder® as they expand their services worldwide because assessments ought to be culturally appropriate for a variety of
countries. In the near future, Kuder® plans to conduct reliability analyses for the Korean version of KSCA.

Chapter 3

Test Bias and Differential Item Functioning (DIF)

Test Bias and Fairness

As the use of assessments continues to grow, there is a widespread concern for test fairness. The Standards for Education and Psychological Testing (1999) was the first edition to contain a chapter on test fairness. An updated version of the Standards are set to be published in Spring 2014 which ought to have changes in the standards for fairness such as issues related to educational accountability, technological advances in testing, and workplace testing and credentialing (AERA, APA, & NCME, 2014). This document provides standards for addressing test bias related to evaluating tests, testing processes, and test use.

Fairness in testing is particularly emphasized in relation to subgroups in the population such as race, ethnicity, gender, licensing and certification, test accommodations, and linguistic diversity (Camilli, 2006). Interestingly, there is no single definition for the term “fairness.” However, fairness is related to the use of a test and its consequences for various groups. All members of groups ought to have equal opportunity to prepare, perform, and take the test in equal testing conditions. The key element in establishing fairness is in the context of comparing two groups. Holland and Thayer (1988) stated that the “performance of two groups may be compared in terms of a test
item, a total test score, or a prediction regarding success on a criterion” (p. 130). The outcome of a test, whether it is expected to show differences between examinees or not, ought to be due to factors that are related to the construct that is being measured by the test.

Depending on the social, legal, or technical context of which an assessment is used, fairness in testing may be interpreted as a lack of bias and as equitable treatment in the testing process. However, it is important to distinguish between fairness and bias; they are not synonymous. Bias, or statistical bias, is a systematic error in which members of different subgroups score differently on an item given the same latent trait the test was built to measure (Kamata, 2004). Therefore, when the test score has different meanings for different subgroups among the examinees, the test becomes unfair to one group over another. In other words, test fairness is in relation to examining whether the observed group differences or test outcomes lead to accurate interpretations within the context of the test.

To strive for fairness in testing, it is important for assessments to be free of bias. Especially in cross-cultural comparisons, tests may be used in linguistically diverse settings. In such a case, bias free translation is essential for an assessment to function with its original intended purpose to provide equal opportunity for all groups of examinees and to make accurate comparisons. Simply translating a version of a test does not necessarily produce an equivalent form of the original test in content, difficulty, reliability, and validity (AERA, APA, & NCME, 1999). If there is bias in translation of tests, the validity of the test is threatened. The 1999 Standards requires a collection of evidence to support the validity of an assessment which shows that there is equality in
conditions of testing and the test scores have consistent meaning for various groups of examinees.

When establishing linguistic and cultural equivalence to make valid comparisons, both linguistic and cultural equivalents affect the following aspects of validity: construct, content, response processes, and consequences. Messick (1989) states that a test measures only construct relevant differences between subgroups. Thus, in the development of an assessment, a construct first ought to be defined in order to avoid construct underrepresentation or construct irrelevant variance (AERA, APA, & NCME, 1999). If there is any bias in the translation, this can lead the translated test to measure a different construct than the original intended construct. For instance, words in different languages may not have equivalent meaning when translated. As a result, the test score may be inferring to a different construct. On the other hand, non-equivalent items may be used to assess construct validity. In addition, when translating a test from one language to another, the construct ought to be similar in both cultures to establish cultural equivalence.

Content validity is also affected by biased test translation. Content related evidence of validity looks at the internal structure of the test. If there are biased items in test translation, the items may be testing unintended content that is different from the original test. Items may be testing unintended content that is different from the original test. More specifically, the difficulty of the test can be inconsistent from original test to the translated test. An item or subscale may be more difficult for members of one group than members of another even though the ability level of both groups is generally equal. In terms of criterion related evidence of validity, the direction of correlation ought to be
consistent as anticipated direction of relationship based on cultural knowledge. Also, the answer to some items may require examinees to have prior cultural knowledge of the dominant cultural group; such an instance may be a threat to content validity.

The 1999 Standards also requires a collection of evidence to support the validity of an assessment which show that there is equality in conditions of testing and the test scores have consistent meaning for various groups of examinees. More specifically, Standards 1999 also established that the test and testing practices ought to be designed in a way to “reduce threats to the reliability and validity of test score inferences that may arise from language differences” (AERA, APA, & NCME, 1999). A deep understanding of the target culture is essential to ensure equal testing conditions. Surveying examinees of the target population, training test administrators, providing clear instructions, and repeated test administrations may strengthen the validity of scores for the target population.

**Type of Bias in Test Translation**

Test translation has important implications for bias and fairness of an assessment. Simply translating a version of a test does not necessarily produce an equivalent form of the original test in content, difficulty, reliability, and validity (AERA, APA, & NCME, 1999). Furthermore, valid translation of a test consists of psychological, linguistic, and cultural aspects (van de Vijver & Hambleton, 1996). In order to make valid interpretations and use of test scores, test translation ought to be free of bias that may arise from linguistic and cultural differences. Thus, for a valid test to be constructed, the test items ought to be equivalent. Hambleton and Bollwark (1991) defines equivalence as “the direct comparability of test items and [test] scores…in terms of psychometric
meaning…[measuring] the same behaviors across the populations of interest” (p. 13). A valid translated test ought to have construct (functional) equivalence, item equivalence, and scalar equivalence.

In test translation, there are three types of bias: construct, method, and item bias (van de Vijver & Hambleton, 1996). In test development, there must be a well-defined construct in which the test will be designed to measure (Camilli, 2006). Based on the construct, test items are created and the test is administered. After the test is administered, one ought to be able to make accurate inferences and interpretations back to the construct. Test fairness is achieved when the test measures only construct relevant differences between subgroups (Messick, 1989). Construct bias occurs when a construct is not defined similarly in both the source and target cultures; only part of the construct may be meaningful in both cultures. In other words, if there is construct irrelevant variance, the test may no longer be measuring the same construct as intended by the original test resulting in unfair inferences. For instance, words in different languages may not have equivalent meaning when translated, and they may have different levels of difficulty. Linguistically, some words may not even exist in the target culture.

The Wellness Evaluation of Lifestyle (WEL) scale was translated from English to Korean (Chang & Myers, 2003). In Korean, there is no word for “Gender” so translators substituted it with a phrase describing qualities of being a female or male. Also, some behaviors in one culture may not be associated with that particular construct. A nomological network is typically used to examine construct validity of a test. However, various aspects of the construct itself may not be representative of the culture (van de Vijver & Tanzer, 2004). Concurrent development, or decentering, in two languages is
recommended to avoid construct bias, but it may be more expensive (Werner & Campbell, 1970). Instead, test developers can further study the target culture by administering local surveys.

While enhancing cultural understanding, surveys may help developers to avoid any ethnocentric words or concepts. A team of researchers, test developers, and translator is recommended to acquire construct equivalence by removing culturally specific words and phrases (Hambleton, Merenda, & Spielberger, 2005). For example, Kim, Lee, Suen, and Lee (2010) used the Delphi consensus-seeking process to survey the meaning of a socially and culturally constructed concept of *school readiness* in early childhood education within the Korean culture. With a list of 23 items describing specific aspects of what *readiness* may mean within the Korean context, twenty-two experts in early childhood education in Korea rated each item with four rounds of the Delphi process. Contrary to Confucian tradition, researchers found that *school readiness* in Korean culture was related to specific functions for social adjustment to school instead of academic abilities such as understanding basic math concepts. Therefore, experts’ consensus gave insight to the values of education in Korea. Another common method to test for construct equivalence is using factor analyses methods such as exploratory factor analysis, confirmatory factor analysis, and simultaneous components analysis (Kiers & Ten Berge, 1989; Watkins, 1989).

A second type of bias is method bias. Method bias is related to test administration, sample, and instrument. Such a bias may include unwarranted differences between source and target samples, different administrative and environmental testing procedures, unclear instructions, tester and examinee communication, varying social norms and
expectations, and varying degrees of test format familiarity. Gathering additional information through triangulation of cultures or examination of test score differences between repeated test administrations may shed light on potential sources of method bias in the target population (Lipson & Meleis, 1989). In addition, training test administrators and detailed testing protocol can help reduce method bias (van de Vijver & Tanzer, 2004). Similar to establishing construct equivalence, a team of experts may help reduce or eliminate method bias in translated tests. Lastly, item bias can arise from poorly worded items, inappropriate content, inaccurate translations, and culture specific factors such as idioms.

Lastly, the translated version of the test ought to have the same psychometric properties as the original test at the item level. Allalouf, Hambleton, and Sireci (1999) conducted a study on possible causes of DIF items in test translation when translating the verbal subset of Psychometric Entrance Test (PET). After translating the verbal subset of PET and running the Mantel-Haenszel DIF analysis, they found four causes of biased items. They found that words or sentences became easier or more difficult, the meaning of items changed, the format of the item and location of words in the item changed, and there were differences in cultural content. Allalouf et al.’s study emphasizes the importance of the accuracy of test translation ensuring bias free items when adapting a test from one language to another. Therefore, bias in test translation could result in an unfair advantage of a certain subgroup.

**Test Equivalence**

Test bias threatens the comparability, or equivalence, of scores for different linguistic and cultural groups. In *Adapting Educational and Psychological Tests for*
Cross-Cultural Assessment, van de Vijver and Poortinga (2005) stated four primary levels of equivalence in test translation: construct equivalence, structural equivalence, measurement unit equivalence, and scalar equivalence. Construct equivalence is obtained when both the original and translated versions of a test measures the same construct. If construct bias exists, the assessments will have construct inequivalence.

Secondly, structural equivalence is related to the internal structure of tests. Patterns of factor and item correlations ought to be the same for each of the subgroups. Factor analyses techniques, such as exploratory factor analysis and confirmatory factor analysis, are commonly used to examine the number of factors and whether or not rotated factor loadings for the items are similar across different groups.

Thirdly, scales in each group ought to be on the same metric or level of measurement but they have a different origin to obtain measurement unit equivalence. Though differences within groups can be compared, between group differences cannot be compared due to different origins. In this study, for example, the English and Korean KSCA items are on ordinal scales. However, if only measurement unit equivalence is obtained, a Holland profile of a Korean student and an American student cannot be interpreted in the same way. Finally, scalar equivalence indicates that both tests have the same psychometric properties. In other words, test scores of both groups are not only on the same metric scale, but they also can be similarly interpreted.

Consequences

In addition to types of bias and levels of equivalence, there are intended as well as unintended social consequences that may arise with assessments. More specifically, 1999 Standards also established that the test and testing practices ought to be designed in a
way to “reduce threats to the reliability and validity of test score inferences that may arise from language differences” (p. 97). When a test is translated from one language to another, bias in a test can affect all aspects of validity. The test should equally serve any group of examinees. If the test score resulted from biased items due to inaccurate test translation, then any use or interpretation of the test score would be inappropriate.

Particularly in education, stigmatization from mislabeling or misdiagnosis can lead to adverse effects for examinees. For instance, a biased intelligence test score can be harmful to minority students (Ford, 2004). Low IQ scores can result in the student being identified as mentally retarded when the test lacks linguistic and cultural equivalence. Such a label affects a student’s education and may lead to unfair and negative social treatment from peers. In the field of psychology, many instruments are readily used to diagnose patients in order to identify appropriate treatments (Wang & Mallinckrodt, 2006). If a psychologist uses a test that is biased from test translation or is not culturally equivalent, the use of such a test may cause the patient to be misdiagnosed. In such a case, an inappropriate treatment may be implemented that may further harm the patient.

At a larger scale, international comparisons may also lead to unintended social consequences for examinees. For example, No Child Left Behind is a national phenomenon that has changed the lives of teachers as well as students (No Child Left Behind [NCLB], 2002). To be internationally competitive and enhance education, political leaders felt that teachers and students ought to be held accountable to boost education of the students. Even though original intentions may have been positive, there were also negative consequences that arose from the tests.
Schools wanted to hold teachers accountable in order to get federal funding so many teachers’ salaries became contingent on their students’ test performance. As a result, teachers felt more pressure to ensure good performance on their students. Unfortunately, such pressure led many teachers to cheat that resulted in the Atlanta Scandal (Sainz, 2012). Even though politicians may have begun with good intentions, the heavy emphasis of assessments ignited a domino effect resulting in negative consequences of many in society. Also, since tests are found on socially constructed concepts, the meanings of the same construct may differ in different cultures. Accordingly, non-equivalent tests may lead to different stigmatizations of examinees and unfair consequences.

**Differential Item Functioning Methods**

Particularly with translated assessments, translation errors can cause DIF items (Ellis, 1989). Using both statistical and judgmental methods, the detection of differentially functioning test items were originally developed to detect bias for different groups tested in the same language (Allalouf, Hambleton, & Sireci, 1999). However, with increasing use of translated tests worldwide, processes to screen for biased items are even more critical to detecting potentially unfair items and preventing unintended consequences. Thissen, Steinberg, and Wainer (1993) indicated that DIF methods can be used to detect any difference between any groups in their responses to any kind of item. DIF may be the result of construct irrelevant variance or extraneous influences that lead to inaccurate interpretations which in turn can have unintended consequences for a subgroup of examinees (Messick, 1995).
Through statistical procedures, item bias statistics, or differential item functioning statistics, indicate if two subgroups of examinees have unexpected differing performance on items. If the level of difficulty of the same item is different across two groups, the item is deemed to contain DIF. However, statistical indices alone are not sufficient for analyzing test bias. Test developers ought to investigate whether differing item performance is relevant to the intended construct being measured by the test and interpret DIF within the context of the test. If the source of bias is determined to be irrelevant to the intended construct, the item is determined as a biased item. Therefore, procedures to detect biased items ought to consist of both measuring the effects of items and identifying possible causes and explanations for bias (Camilli & Shepard, 1994).

To determine if the test is biased, it is helpful to examine the test at the item level. Several statistical methods can be used to detect item bias. Generally, in DIF analysis, one subgroup of examinees is assigned as the focal group and the other as a reference group. The focal group is the group of interest and the reference group is the comparison group (Holland & Wainer, 1993). Given the same underlying trait or ability, examinees from both the source and target culture ought to have the same expected score (Camilli, 2006). In essence, biased items require DIF as well as the effect of leading to differences in test performance associated to group characteristics or constructs irrelevant to the intended purpose of a measure.

According to Zumbo (2007), examination of biased items began with three major stages in the history of DIF. In the beginning, a DIF item indicated a situation in which one group of examinees outperformed another group on the same item with implications of bias. However, DIF evolved to include both item bias and the impact of an item
demonstrating that a DIF item did not necessarily equate to item bias. The term *item impact* was used to describe a situation which a DIF item showed the differences of item scores that were due to a true difference between the two comparison groups, whereas, item bias indicated that an item displayed DIF due to differences that were not relevant to the underlying trait or ability that was being measured by a test. Hereafter, the term DIF became the accepted terminology to refer to differences in performances between two groups of examinees due to extraneous factors.

In the third major stage, aspects of testing situations in addition to item characteristics that were irrelevant to the underlying ability or skill of interest were also understood as possible causes of DIF. The following three methods were primarily used to model item responses: contingency tables and/or regression models, item response theory models, and multidimensional models.

In the contingency tables or regression models, there is an *a priori* assumption to DIF analysis that examinees of two subgroups have equal trait or ability levels. Examinees belonging to each group are essentially matched based on their underlying traits which are represented by their test scores; any differences between the scores of reference and focal group members indicate DIF. Therefore, DIF consists of both uniform (main group effect) and non-uniform DIF (interaction of group and ability effect). For uniform DIF, the probability of correctly answering an item for one subgroup of examinees is consistently higher than the other subgroup across all trait or ability levels while holding the underlying trait or ability constant (Brennan, 2006; Camilli & Shepard, 1994; Holland& Wainer, 1993). On the other hand, non-uniform DIF looks at the
probability of correctly answering an item for one subgroup of examinees is higher and lower than the other subgroup across various points of the trait or ability levels.

Within the contingency and regression modeling framework, the Mantel-Haenszel (MH) and logistics regression (LR) methods have been developed. The original MH statistic was a variance statistic using the odds ratio statistic calculated from contingency tables (Holland & Thayer, 1988; Mantel & Haenszel, 1959; Zumbo, 2007). Restricted to comparing two matched groups, MH usually uses a two-by-two contingency table for each total test score in which item responses are counted as either correct or incorrect (item score) and with two levels of group membership (focal or reference). At a given underlying trait level as indicated by controlling for total score, an item shows no DIF if the odds of responding to an item correctly were equivalent for both focal and reference group members. Other forms of MH procedures were generalized Mantel-Haenzel test for nominal responses and the Mantel test for ordinal responses.

For purposes of easier interpretation, the MH index was later converted into a log scale and became MH D-DIF index. MH D-DIF used Bayesian procedures based on the delta metric (Holland & Thayer, 1985; Mantel & Haenszel, 1959). As a result, positive MH D-DIF values favor the focal group making it easier to interpret. An advantage of MH D-DIF over the original MH procedure is its stability across populations. Although MH methods may be intuitive, the MH procedure lacks power in detecting non-uniform (interaction) bias especially with a sample size of less than twenty-five and large group mean score differences (Hambleton & Rogers, 1989; Strout, 1990; Wang & Su, 2004).

Another DIF method is based on logistic regression (LR) which was first introduced by Swaminathan and Rogers (1990). By using the total subscale score as a
continuous matching variable, item responses are analyzed as dichotomous variables representing the focal and reference groups (Clauser & Mazor, 1998; Zumbo, 1999). In the LR procedure, the dependent variable is scored as either a binary item response (e.g. correct or incorrect) or ordinal (e.g., five point Likert type scale). The independent variables include dummy coded grouping variable for the reference and focal groups, total scale score variable as each subject's total scale score, and a group by total scale score interaction (see equation 1). By setting the item response as the dependent variable, one can test the significance of the relationship between item response and total subscale score \( b_1 \), effects of uniform DIF \( b_2 \), and effects of non-uniform DIF \( b_3 \).

Equation 1. LR equation

\[
\text{logit } [p (Y \leq j)] = b_0 + b_1(\text{total}) + b_2(\text{group}) + b_3(\text{total})(\text{group})
\]

*where \( Y \) is the answer given by an examinee \( j \)

The logistic regression (LR) procedure is effective at controlling between group differences by including multiple traits in the model and the interaction term in its model (Zumbo, 2007). Not only does the LR procedure take into consideration the continuous underlying trait or ability of examinees, but it also allows one to simultaneously test for uniform and non-uniform DIF. Zumbo (1999) stated three advantages of the LR method over other DIF analysis methods: no need to categorize a continuous criterion variable, efficient test for both uniform and non-uniform DIF by including total subscale score and group membership as independent variables, and extension of the basic binary LR model for analyzing ordinal item responses. In addition, the modeling strategy for ordinal LR is equivalent for both binary and ordinal items. Personality or career inventories such as the Kuder® assessments typically have Likert-type responses. For instance, in the case of
KSCA, item responses are five-point Likert-type scales. Thus, the LR method for DIF analysis is efficient and flexible.

The total subscale score variable is sequentially entered into the LR model as the conditioning variable followed by the grouping variable and the interaction term (see Equation 2).

Equation 2. Sequential Modeling for DIF-LR

Model 1

\[
\text{logit} \left[p \left(Y \leq j\right)\right] = b_0 + b_1(\text{total})
\]

Model 2

\[
\text{logit}\left[p \left(Y \leq j\right)\right] = b_0 + b_1(\text{total}) + b_2(\text{group})
\]

Model 3

\[
\text{logit} = \left[p \left(Y \leq j\right)\right] = b_0 + b_1(\text{total}) + b_2(\text{group}) + b_3(\text{total})(\text{group})
\]

*where $Y$ is the answer given by an examinee $j$*

To evaluate the items for DIF, a chi-square test for LR is used for test for DIF with no missing data and a sample size of at least 200 per subgroup (Zumbo, 1999). By subtracting the model 3 (log-likelihood) chi-square value from model 1 chi-square value $(df=2)$, one can simultaneously test for uniform and non-uniform DIF (Swaminathan & Rogers, 1990). Furthermore, DIF-LR method has an effect size measure accompanying the test statistic. Zumbo emphasized the importance of determining whether or not an item has DIF by looking at both the LR chi-square test statistic and its corresponding effect size (Zumbo & Thomas, 1997).

The Zumbo-Thomas effect size measure examines the R-squared change between the models. For an item to be flagged as a DIF item, the chi-square test between model 3
and model 1 \((df=2)\) ought to have a p-value less than or equal to 0.01 and for the Zumbo-Thomas effect size measure to be greater than or equal to 0.13. If both the p-value is significant and the effect size is larger than the threshold, one can determine whether the item has uniform DIF or non-uniform DIF. Since model 2 tests for the effect of group membership on item response while controlling for the total scale score, comparing the p-value from the LR chi-square test and R-squared change between model 2 and model 1 measures the significant of uniform DIF. For this study, uniform DIF points to item bias due to cultural or linguistic differences. For non-uniform DIF, the LR chi-square value and R-squared change value of model 3 and model 2 are compared. For any DIF items, DIF items are eliminated from the measure in subsequent purification analyses. Total scale scores are then recalculated as the new matching criterion for all items.

When evaluating the items in the Korean translated Kuder\textsuperscript{®} Assessment, it is essential to take into consideration the various cultural and individual differences between students in the United States and South Korea as well as potential language differences. Yasuda, Lei, and Suen (2007) examined DIF of the Multiple Affect Adjective Check List-Revised (MAACL-R) in the English and the Japanese translated version using the logistic regression (LR) procedure. The MAACL-R is an instrument used for research and clinical purposes to assess human affect. There was a total of 66 items consisting of 132 adjectives and 5 subscales.

Using the back translation method, the instrument was originally in English and then three people translated to Japanese. Afterward, three other people translated the Japanese version back to English. To test for psychometric equivalence at the item level, the researchers used the LR procedure for the DIF analysis in addition to a multiple group
confirmatory factor analysis (MGCFA) to increase the accuracy of identifying DIF items. After running the analysis, Yasuda et al. identified 36 out of 66 DIF items using LR procedure and 31 DIF items using the MGCFA procedure. Possible limitations of the study were smaller sample size and a need for further evaluation at the test level. However, results showed that possible reasons for group differences were due to translation errors and cultural relevance.

A second method for DIF analysis is within the Item Response Theory framework. Angoff and Ford (1973) stated that DIF analysis ought to be based on multivariate matching strategies. Within the Item Response Theory framework, differential item functioning (DIF) is a statistical method that allows one to determine whether the items are functioning in the same way for all groups of interest (Zumbo, 2007). Based on the DIF-IRT, item difficulty and item discrimination are calculated to determine if items are biased for each group. In a bias free item, item difficulty and item discrimination should be equal in both groups (Camilli, 2006). Individuals with the same trait level in different groups should have equal probabilities of responding to a test items. In other words, an item is biased when members of two subgroups score differently on that item given that they have the same underlying ability or trait.

IRT produces visual representations for DIF analysis by examining the area between item characteristic curves (ICCs). ICC is a nonlinear regression function of an item score on an underlying trait or ability that is being measured by a test (Hambleton & Swaminathan, 1985). When an item displays uniform DIF, the two curves for both groups have the same slope but they differ in their thresholds (point along the underlying continuum where there is fifty percent probability of examinees of endorsing an item. If
ICCs intersect, the item demonstrates non-uniform DIF in which discrimination and/or guessing parameters are taken into account. An advantage of using DIF within the IRT framework over MH method is that matching criteria is not necessary. Conversely, there are few challenges with DIF-IRT. Firstly, scaling a latent underlying variable on a theta scale is arbitrary. Secondly, if the two subgroups have different ability distributions, the arbitrary underlying scales are different making DIF interpretations difficult.

Lastly, multidimensional methods are used to detect DIF items that are designed to measure multiple constructs or factors (Ackerman, 1992). In a multidimensional test, the total subscale score is a weighted function of all underlying abilities or traits (Camilli, 1992). Simultaneous item bias test (SIBTEST) is effective at examining various multidimensional scenarios as potential sources of DIF (Shealy & Stout, 1993). Also, Roussos and Stout (1996) described a theory-based approach in which meaningful hypotheses are made for possible reasons for DIF. In a multidimensional framework, such a theory-based approach may be advantageous by identification of secondary dimensions. For either of these methods, clearly defined constructs along with factor analysis and confirmatory analysis facilitate DIF analysis for multidimensional items.
Chapter 4

Methods

Culture can be defined as “socially constructed phenomenon…[that serves as a lens] for understanding life…” (p. 6, Gopaul-McNicol & Armour-Thomas, 2002). However, culture is a concept that affects everyday decisions. From the quality of test development and translators to the validation methods of translated tests, linguistic and cultural differences ought to be carefully evaluated. Hambleton (2001) describes linguistic and cultural differences between source and target language populations as cultural distance. Cultural distance might consider differences in language, family structure, religion, lifestyle, and values. By assessing cultural distance, test developers and translators may be able to reduce linguistic and cultural differences that the test is attempting to measure. For example, Fernandez-Ballesteros, Hambleton, and O’Neil (2001) found that more than 10% of test items in a field test of cognitive, psychological, and life skills were found to contain potential bias due to error in translation in seven countries (Fernandez-Ballesteros, Hambleton, & O’Neil, 2001).

As cross-cultural comparisons are becoming more prevalent, there is an increasing need for linguistic and cultural equivalence of original and translated tests. Unbiased test translation is a vital start to designing equivalent tests. However, test translation can lead to construct, method, and item bias. ITC guidelines facilitate test developers and translators to obtain fair cross-cultural comparisons. Even though there is no single definite method to construct a valid test, statistical and judgmental methods are
utilized to address issues with test translation in order to establish linguistic and cultural equivalence. Also, careful examination and study of the target culture is imperative for valid score interpretations. Despite great efforts to make valid cross-cultural comparison, there may be unforeseen linguistic or cultural differences that ultimately have the biggest impact on the examinees. Therefore, test translation and cultural equivalent tests may be most effectively achieved in a collective effort for valid cross-cultural comparisons.

Even though it may be impossible to obtain complete fairness, it is imperative to strive for equality in testing with the growing movement of cross-cultural comparisons. Unfortunately, there are not many studies that have been done examining bias in test translation. The purpose of this study is to examine if there is psychometric equivalence at the item level between the English and Korean versions of Kuder® Skills Confidence Assessment (KSCA) using participants from the United States and South Korea.

**Participants and Procedures**

The U.S. participants consisted of 499 high school students (222 boys and 277 girls; 95 freshmen, 63 sophomores, 97 juniors, and 243 seniors). Among the 414 participants who indicated their ethnicity, the following ethnic groups were represented in this sample: Hispanic, Non-Hispanic, Black, White, Asian, American Indian, and Hawaiian/Pacific Islander (see Table 4-1). From the Republic of Korea, 307 participants were recruited from advanced elementary, junior high, to senior high schools (1 in sixth grade, 21 in seventh grade, 8 in eighth grade, 74 in ninth grade, 135 in tenth grade, 15 in eleventh grade, 53 in twelfth grade). There were 133 boys and 174 girls in the Korean sample.
Table 4-1: Ethnic Groups of U.S. Sample

<table>
<thead>
<tr>
<th>Ethnicity</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hispanic</td>
<td>53</td>
</tr>
<tr>
<td>Non-Hispanics</td>
<td>8</td>
</tr>
<tr>
<td>Black</td>
<td>11</td>
</tr>
<tr>
<td>White</td>
<td>9</td>
</tr>
<tr>
<td>Asian</td>
<td>2</td>
</tr>
<tr>
<td>American Indian/Black</td>
<td>1</td>
</tr>
<tr>
<td>American Indian/White</td>
<td>1</td>
</tr>
<tr>
<td>Hispanic</td>
<td>28</td>
</tr>
<tr>
<td>White</td>
<td>16</td>
</tr>
<tr>
<td>Black</td>
<td>6</td>
</tr>
<tr>
<td>Asian</td>
<td>2</td>
</tr>
<tr>
<td>American Indian/White</td>
<td>1</td>
</tr>
<tr>
<td>American Indian/Black/White</td>
<td>1</td>
</tr>
<tr>
<td>Asian/Hawaiian/Pacific Islander/White</td>
<td>1</td>
</tr>
<tr>
<td>Hawaiian/Pacific Islander</td>
<td>1</td>
</tr>
<tr>
<td>Non-Hispanic</td>
<td>301</td>
</tr>
<tr>
<td>White</td>
<td>178</td>
</tr>
<tr>
<td>Black</td>
<td>92</td>
</tr>
<tr>
<td>Asian</td>
<td>11</td>
</tr>
<tr>
<td>Black/White</td>
<td>6</td>
</tr>
<tr>
<td>American Indian/White</td>
<td>5</td>
</tr>
<tr>
<td>American Indian</td>
<td>3</td>
</tr>
<tr>
<td>American Indian/Black/White</td>
<td>2</td>
</tr>
<tr>
<td>Asian/White</td>
<td>2</td>
</tr>
<tr>
<td>Asian/Black/White</td>
<td>1</td>
</tr>
<tr>
<td>Black/Hawaiian/Pacific Islander/White</td>
<td>1</td>
</tr>
</tbody>
</table>

Measure

The 56-item form of Kuder® Skills Confidence Assessment (KCSA) used in this study is a career self-efficacy measure for career planning and counseling. Respondents are asked to rate their level of confidence in performing or learning to perform the skills necessary in various occupations. Item responses are on a five-point Likert scale. The six subscales and the number of items (in parentheses) are Realistic (9), Investigative (15), Artistic (6), Social (10), Enterprising (12), and Conventional (5). Based on the item ratings, a set of six Holland scores are calculated ranging from 1 to 7 for each respondent.
Harris-Bowlsbey, Suen, Trusty, and Niles (2013) reported Cronbach's $\alpha$ estimates for KSCA ranging from 0.74 to 0.91. In the present study for the U.S. sample, Cronbach's $\alpha$ coefficients are .901 (Realistic), .897 (Investigative), .679 (Artistic), .884 (Social), .929 (Enterprising), and .785 (Conventional).

Kuder®, Inc. hired a translation company to implement the back translation method to translate KCSA from English to Korean. In addition, Dr. Hoi Suen, an Asian American consultant to Kuder® and Dr. Juhu Kim, a native Korean professor of educational psychology, conducted a cultural sensitivity review of the item content. In the present study for the Korean sample, Cronbach's $\alpha$ estimates were also comparable to the original version of KSCA with the following coefficients: .897 (Realistic), .915 (Investigative), .670 (Artistic), .879 (Social), .917 (Enterprising), and .760 (Conventional).

**Data Analysis**

After assessment data were collected, differential item functioning (DIF) analysis using the polytomous Logistic Regression (LR) procedure was conducted to identify biased test items (Zumbo, 1999; Swaminathan & Rogers, 1990). Because the matching criterion is a multidimensional composite (KSCA measures six Holland areas), any items identified as having DIF may be due to uncontrolled differences between the two groups (Mazor, Kanjee, & Clauser, 1995). Accordingly, using the DIF-LR procedure controls for total subscale scores and includes an interaction term. Subsequently, the differences between the two groups were removed using the LR model. Any remaining item differences were identified as biased items within both English and Korean versions of the test. As a result, any consistent difference across levels of the underlying trait
(uniform DIF) and inconsistent differences (non-uniform DIF) can be effectively analyzed between groups.

Variables were entered sequentially to create a total of three models. In the first LR model (M1), the total scale score was entered as the matching variable representing each Holland area of the KSCA. Secondly, group membership was added in the second model (M2). Lastly, Model 3 (M3) was constructed by adding the interaction between group membership and the total scale score of item in question corresponding to each scale. There were a total of 56 sets of LR models for each of the 56 KSCA items.

In the first round of analysis, chi-square differences between M3 and M1 \((df=2)\) were examined in a series of simultaneous tests of uniform and non-uniform DIF for the combined American and Korean sample of 806 respondents (Swaminathan & Rogers, 1990; Zumbo, 1999). Additionally, chi-square differences \((df=1)\) between M2 and M1 were examined for uniform DIF and between M3 and M2 for non-uniform DIF. From the chi-square differences, one can examine incremental fit between models which is due to the addition of the group membership or the interaction term between group and total scale score of an item.

An item was considered to show DIF if the chi-square difference test was found to be significant at the alpha level of .01. Alpha level was set at the conservative level of .01 to control for overall Type I error due to conducting multiple tests. Also, the item in question was included in each subscale analysis to reduce the inflation of Type I error (Holland & Thayer, 1988). From the LR procedure, any items that were identified as DIF (uniform or non-uniform DIF) in the initial round of analysis were excluded in the scale scores for a follow up analysis. Such purification processes were conducted until no new
DIF items were identified. For an item to be identified as a DIF item, both the significance of the chi-square test had as well as an R-squared change ($\Delta R^2$) were examined between the models. An item with a Zumbo-Thomas effect size measure ($\Delta R^2$) greater than or equal to 0.13 was flagged as a DIF item (Zumbo, 1999; Zumbo & Thomas, 1997).

An item identified as DIF alone does not necessarily equate to item bias. Hence, following the statistical analyses, expert judgments helped gain insight on potential causes of DIF. A team of Kuder®, Inc. experts, Dr. Juhu Kim, Korean graduate students and high school teachers, fellow Korean colleagues currently residing in the U.S., and the researcher reviewed the content of the identified DIF item. In cross cultural and linguistic test adaptation, such a review for construct equivalence is equally critical if not more to fully understand potential sources of DIF for test bias (Hambleton & Patsula, 1998; Hambleton, Merenda, & Spielberger, 2005).
Chapter 5

Results and Discussion

Results

The means, standard deviations, and Cronbach’s α of KSCA indicate that Korean respondents had higher mean ratings on the Investigative, Enterprising, and Conventional scales (see Table 5-1). From the LR analysis, a majority of the KSCA were DIF free (see Appendix A.1-A.5). Only one item, item 10 on the Conventional scale, was flagged as a DIF item as showing uniform DIF with a $\chi^2 = 167.60$ ($p<.001$ and $\Delta R^2 = .13$) suggesting that there were consistent differences in ratings across groups over the range of conventional tasks (see Table 5-2). Parameter estimates of the LR model and with their corresponding effect sizes for the DIF item 10 indicate that the odds of scoring higher on Conventional scale were 6.767 times higher for Koreans than Americans holding total scale score constant (see Table 5-3). In addition, Korean students tend to have higher confidence in the Korean item 10 ($M=3.85$, $SD=1.07$) than American students in the English item 10 ($M=2.83$, $SD=1.01$).

After filtering the DIF item from the total score, the purification process resulted in no additional DIF items with an increase in Cronbach’s α of the Conventional scale from .76 to .78. Prior to filtering item 10 from the Conventional scale, the mean of the Korean sample was higher ($M=15.22$, $SD=4.06$) than the mean of the U.S. sample ($M=14.13$, $SD=4.10$). However, after excluding item 10, the mean Conventional scale scores for Korean ($M=11.36$, $SD=3.55$) and U.S. ($M=11.30$, $SD=3.60$) samples became approximately equal. Thus, this is further evidence that item 10 favored the Korean respondents.
Table 5-1. Total Scale Score Means, Standard Deviation, and Cronbach's α of KSCA (N=806).

<table>
<thead>
<tr>
<th></th>
<th>United States</th>
<th>Republic of South Korea</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>Realistic</td>
<td>27.73</td>
<td>8.24</td>
</tr>
<tr>
<td>Investigative</td>
<td>41.39</td>
<td>10.69</td>
</tr>
<tr>
<td>Artistic</td>
<td>17.53</td>
<td>4.73</td>
</tr>
<tr>
<td>Social</td>
<td>32.38</td>
<td>8.28</td>
</tr>
<tr>
<td>Enterprising</td>
<td>36.99</td>
<td>10.71</td>
</tr>
<tr>
<td>Conventional</td>
<td>14.13</td>
<td>4.10</td>
</tr>
</tbody>
</table>

Table 5-2. Conventional Scale Differential Item Functioning (DIF) by Logistic Regression (LR) Models (N=806).

<table>
<thead>
<tr>
<th>Item</th>
<th>Simultaneous Uniform and Non-Uniform DIF $\chi^2$</th>
<th>$p$</th>
<th>$\Delta R^2$</th>
<th>Uniform DIF $\chi^2$</th>
<th>$p$</th>
<th>$\Delta R^2$</th>
<th>Non-uniform DIF $\chi^2$</th>
<th>$p$</th>
<th>$\Delta R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>19.13</td>
<td>.00</td>
<td>.01</td>
<td>18.91</td>
<td>.00</td>
<td>.23</td>
<td>18.91</td>
<td>.00</td>
<td>.01</td>
</tr>
<tr>
<td>10*</td>
<td>167.64</td>
<td>.00</td>
<td>.13</td>
<td>167.60</td>
<td>.00</td>
<td>.04</td>
<td>167.60</td>
<td>.00</td>
<td>.03</td>
</tr>
<tr>
<td>31</td>
<td>90.51</td>
<td>.00</td>
<td>.06</td>
<td>90.20</td>
<td>.00</td>
<td>.30</td>
<td>90.20</td>
<td>.00</td>
<td>.04</td>
</tr>
<tr>
<td>48</td>
<td>6.06</td>
<td>.05</td>
<td>.00</td>
<td>4.46</td>
<td>.03</td>
<td>.160</td>
<td>4.46</td>
<td>.00</td>
<td>.01</td>
</tr>
<tr>
<td>72</td>
<td>.04</td>
<td>.98</td>
<td>.00</td>
<td>.03</td>
<td>.86</td>
<td>.94</td>
<td>.03</td>
<td>.86</td>
<td>.94</td>
</tr>
</tbody>
</table>

* Significant at $p <.001$, $\Delta R^2 \geq .130$, $\Delta R^2 =$ effect size

Table 5-3. LR Model Parameters for Conventional Item 10.

<table>
<thead>
<tr>
<th></th>
<th>$\beta$</th>
<th>SE</th>
<th>$p$</th>
<th>exp($\beta$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Conventional Score</td>
<td>.342</td>
<td>.020</td>
<td>.000</td>
<td>1.408</td>
</tr>
<tr>
<td>Group</td>
<td>1.912</td>
<td>.155</td>
<td>.000</td>
<td>6.767</td>
</tr>
<tr>
<td>Const. 1</td>
<td>-3.739</td>
<td>.347</td>
<td>.000</td>
<td>.024</td>
</tr>
<tr>
<td>Const. 2</td>
<td>-6.004</td>
<td>.360</td>
<td>.000</td>
<td>.002</td>
</tr>
<tr>
<td>Const. 3</td>
<td>-8.512</td>
<td>.421</td>
<td>.000</td>
<td>.000</td>
</tr>
<tr>
<td>Const. 4</td>
<td>-9.831</td>
<td>.455</td>
<td>.000</td>
<td>.000</td>
</tr>
</tbody>
</table>

$\beta =$ parameter estimate, SE= standard error, exp($\beta$) = odds of scoring higher on for Koreans than Americans holding total scale score constant
Discussion

In order to ensure valid score comparisons across cultural and linguistic backgrounds, it is essential to examine possible causes for such DIF. The LR analyses identified one Conventional item (item 10) as showing DIF. If the sample means of two comparison groups are substantially different and sample size is below the minimum requirement of 200 per group, then DIF can be due to Type I error inflation (Lei, Chen, & Yu, 2006). However, the results in this study show that there was no substantial difference between the Korean and U.S. samples. Additionally, sample sizes over 200 per group give more evidence that the sample means are representative of the general trait differences between the Korean and U.S. populations.

Although there may be several reasons for the systematic differences between the U.S. and Korean samples in their response patterns, translation errors and cultural relevance are possible causes of DIF (Hulin, 1987). In this study, the quality and the content of item 10 were closely examined for possible reasons for uniform DIF in the Korean and English versions (see Table 4). Based on judgments and item review, potential causes of DIF are described in the discussions below for item 10 in the Conventional scale.

Table 5-4. Conventional Item 10 Content.

<table>
<thead>
<tr>
<th>English version</th>
<th>Help decide on the most efficient set of routes for the school’s buses</th>
</tr>
</thead>
<tbody>
<tr>
<td>Korean version</td>
<td>집에서 시청으로 가는 가장 효과적인(효율적인) 경로를 결정하는데 도움을 제공합니다</td>
</tr>
<tr>
<td>English translation of Korean version</td>
<td>Help decide on the fastest route to go from home to the city hall</td>
</tr>
</tbody>
</table>
For Item 10, since few Korean students take school buses, it had been decided during the original cultural review that the task of deciding on the fastest route to go from home to the city hall might be equivalent to the task described in the English version. Hence, the Korean version did not indicate the same identical task. However, the results of the DIF analysis suggest that the two versions may not be equivalent after all.

One possible explanation that item 10 in the Conventional scale functioned differently for each group of examinees is that the Korean version measures students’ social skills and behavior in addition to conventional skills. In order to find the fastest way to go from home to the city hall, for instance, Korean students need to understand a complex public transportation system (i.e. highway and subway systems). Thus, if an individual does not have prior knowledge of the public transportation system, one may be need to ask others for help which requires some level of confidence in his or her social and communication skills.

A second possible cause is in reference to the words *efficient* in the English version and *fastest* route in the Korean version. An *efficient* route can be defined as the route with the least amount of traffic, least crowded bus, or least amount of travel time. However, the *fastest* route implies that the chosen route is one that takes the least amount of time.

Thirdly, the Korean translation uses the word *효과적* (effective) followed by *효율적* (efficient) in parentheses. Perhaps, the difference in definition between these two words affects Korean students’ responses. In this case, the term *효율적* meaning *efficient* may be a more comparable translation to the English version.
Lastly, the Korean version seems to be measuring the Investigative Holland type (prefers tasks that consist of investigation of physical, biology, and cultural phenomena) rather than Conventional type (prefers tasks using orderly and systematic fashion using data and materials). In the English version, a task to help decide on the most efficient set of routes for the school’s buses seems to be more of a demanding and conventional task. Firstly, an individual may need to take more time in exploring and weighing out possible routes to decide on the most efficient set of routes whereas the Korean version is in reference to a decision for a single fastest route. Consequently, finding a single route as stated in the Korean version requires less time commitment and planning than the English version. Secondly, the English version addresses set of routes for school’s buses which requires one to consider multiple routes for a variety of buses which are characteristic of conventional tasks of working in an orderly and systematic way. As previously mentioned, Korean students tended to have higher Investigative scores than U.S. students. Similarly, mean scale ratings showed that Korean students have higher overall confidence in Investigative tasks than the U.S. students. Such results may provide indirect evidence that the Korean item is measuring Investigative types instead of Conventional types.

Conclusions and Future Studies

In analyzing the translated KSCA for cultural and linguistic bias, the results of this study suggest that a majority of the items have no DIF and that they are adequate for measuring career self-efficacy in relation to Holland type tasks among Korean and the U.S. students. At the same time, to ensure comparable scores for test fairness, follow-up steps are necessary to discard or revise the one item that exhibited uniform DIF because it
is not reflecting the underlying conventional trait in the same way across the two language versions of the KSCA. Another option is to create and re-run DIF analysis for a new tryout items. In short, the main causes of DIF suggested in this study were differences in cultural relevance, and changes in content (Allalouf, Hambleton, & Sireci, 1999; Sireci & Allalouf, 2003).

Few studies have been conducted to examine the detection of causes of DIF for translated tests (Allalouf, Hambleton, & Sireci, 1999). However, empirical and judgmental reviews for cultural and linguistic equivalence is critical as international adaptations of tests become increasingly prevalent. In particular, KSCA is a widely used career assessment to aid students in their educational and professional careers. The validation process is an ongoing process to ensure culturally and linguistic equivalent tests as assessments are used in various contexts and time periods. For that reason, future studies may gain more insight into possible causes of DIF for the English and Korean versions of KSCA. In order to revise or replace the identified DIF item, future studies may implement think-a-loud processes for Korean and American high school students. Secondly, empirical studies may be used to understand the nature of DIF (Camilli & Shepard, 1994). Thirdly, a multiple group confirmatory factor analysis can examine the internal structure of KSCA. Lastly, KSCA may include new try out items for the Conventional scale and re-run DIF analyses.

With increasing popularity of Kuder® assessments, it is imperative that great efforts and attention are placed on obtaining cultural and linguistic equivalent tests. While making valid cross-cultural comparisons, there may be unforeseen linguistic or cultural differences that ultimately have the biggest impact on the examinees. The use of
both judgmental methods such as cultural sensitivity reviews for item content and statistical methods such as DIF-LR method are helpful to design equivalent tests. In the end, test translation and cultural equivalent tests may be most effectively achieved in a collective effort for valid cross-cultural comparisons.

Although DIF analysis using LR model is effective in controlling for between group differences, there are a few limitations. Total subscale scores are observed variables which are prone to measurement error. Consequently, the inclusion of total subscale scores into the model also constitutes measurement error. Another possible limitation that may arise with this study is sample size. Typically, a sample size of at least 200 respondents per group is adequate (Zumbo, 1999). Although the sample size of this study is greater than 200 per group, further analysis and studies may be necessary to ensure that the sample means of the two groups accurately reflect the general trait differences between the U.S. and Korean populations. Thus, Type I error may be inflated when both groups differ largely on overall trait score.

Despite its limitations, this DIF study is advantageous because it is both methodological and theory-driven analysis of KSCA. Most DIF studies are simply a methodological exercise in a test validation process. However, this DIF study examines the KSCA within a theoretical framework in addition to its methodological approach. Because there are expected cultural differences between U.S. and Korea, there were expected outcomes prior to DIF analysis. Therefore, the interpretation of DIF analysis results are strengthened due to its theoretical framework.
References


doi:http://dx.doi.org/10.1037/0021-9010.74.6.912


International Test Commission. (1993). *Technical standards for translating tests and establishing test score equivalents.* (Available from the International Test Commission, c/o Dr. Ronald Hambleton, School of Education, University of Massachusetts, Amherst, MA 01003.)


Shealy, R., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*, 159-194.


Yasuda, T., Lei, P. W., & Suen, H. K. (2007). Detecting differential item functioning in
the Japanese version of the Multiple Affect Adjective Checklist -- Revised.


Zumbo, B. D. (2007). Three generations of DIF Analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, 4(2), 223-233.

Appendix

Appendix 1. Realistic Scale Differential Item Functioning (DIF) Indexes by the Logistic Regression (LR) Models

<table>
<thead>
<tr>
<th>Item</th>
<th>Simultaneous Uniform DIF $\chi^2$</th>
<th>$p$</th>
<th>$\Delta R^2$</th>
<th>Uniform DIF $\chi^2$</th>
<th>$p$</th>
<th>$\Delta R^2$</th>
<th>Non-Uniform DIF $\chi^2$</th>
<th>$p$</th>
<th>$\Delta R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>8.07</td>
<td>0.02</td>
<td>0.01</td>
<td>7.28</td>
<td>0.01</td>
<td>0.01</td>
<td>0.78</td>
<td>0.38</td>
<td>0.78</td>
</tr>
<tr>
<td>11</td>
<td>13.07</td>
<td>0.00</td>
<td>0.01</td>
<td>13.04</td>
<td>0.00</td>
<td>0.01</td>
<td>0.03</td>
<td>0.86</td>
<td>0.03</td>
</tr>
<tr>
<td>15</td>
<td>2.81</td>
<td>0.24</td>
<td>0.00</td>
<td>0.56</td>
<td>0.45</td>
<td>0.00</td>
<td>2.25</td>
<td>0.13</td>
<td>2.25</td>
</tr>
<tr>
<td>24</td>
<td>1.81</td>
<td>0.40</td>
<td>0.00</td>
<td>1.73</td>
<td>0.19</td>
<td>0.00</td>
<td>0.08</td>
<td>0.77</td>
<td>0.08</td>
</tr>
<tr>
<td>27</td>
<td>7.00</td>
<td>0.03</td>
<td>0.00</td>
<td>7.00</td>
<td>0.01</td>
<td>0.00</td>
<td>0.01</td>
<td>0.94</td>
<td>0.01</td>
</tr>
<tr>
<td>32</td>
<td>1.12</td>
<td>0.57</td>
<td>0.00</td>
<td>0.33</td>
<td>0.56</td>
<td>0.00</td>
<td>0.79</td>
<td>0.37</td>
<td>0.79</td>
</tr>
<tr>
<td>51</td>
<td>33.29</td>
<td>0.00</td>
<td>0.02</td>
<td>33.28</td>
<td>0.00</td>
<td>0.02</td>
<td>0.91</td>
<td>0.92</td>
<td>0.01</td>
</tr>
<tr>
<td>53</td>
<td>87.00</td>
<td>0.00</td>
<td>0.07</td>
<td>86.46</td>
<td>0.00</td>
<td>0.07</td>
<td>0.54</td>
<td>0.46</td>
<td>0.00</td>
</tr>
<tr>
<td>71</td>
<td>12.14</td>
<td>0.00</td>
<td>0.01</td>
<td>1.90</td>
<td>0.17</td>
<td>0.00</td>
<td>10.24</td>
<td>0.00</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Note. $\Delta R^2$ = effect size
Appendix 2. Investigative Scale Differential Item Functioning (DIF) Indexes by the Logistic Regression (LR) Models

<table>
<thead>
<tr>
<th>Item</th>
<th>Simultaneous Uniform and Non-Uniform DIF</th>
<th>Uniform DIF</th>
<th>Non-Uniform DIF</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\chi^2$</td>
<td>$p$</td>
<td>$\Delta R^2$</td>
</tr>
<tr>
<td>6</td>
<td>3.20</td>
<td>0.20</td>
<td>0.00</td>
</tr>
<tr>
<td>8</td>
<td>6.28</td>
<td>0.04</td>
<td>0.01</td>
</tr>
<tr>
<td>12</td>
<td>11.86</td>
<td>0.00</td>
<td>0.01</td>
</tr>
<tr>
<td>18</td>
<td>4.48</td>
<td>0.11</td>
<td>0.00</td>
</tr>
<tr>
<td>19</td>
<td>7.07</td>
<td>0.03</td>
<td>0.01</td>
</tr>
<tr>
<td>30</td>
<td>25.84</td>
<td>0.00</td>
<td>0.02</td>
</tr>
<tr>
<td>35</td>
<td>1.95</td>
<td>0.38</td>
<td>0.00</td>
</tr>
<tr>
<td>40</td>
<td>18.43</td>
<td>0.00</td>
<td>0.01</td>
</tr>
<tr>
<td>42</td>
<td>0.46</td>
<td>0.79</td>
<td>0.00</td>
</tr>
<tr>
<td>43</td>
<td>2.68</td>
<td>0.26</td>
<td>0.00</td>
</tr>
<tr>
<td>49</td>
<td>6.49</td>
<td>0.04</td>
<td>0.00</td>
</tr>
<tr>
<td>52</td>
<td>18.44</td>
<td>0.00</td>
<td>0.01</td>
</tr>
<tr>
<td>60</td>
<td>40.34</td>
<td>0.00</td>
<td>0.03</td>
</tr>
<tr>
<td>65</td>
<td>9.79</td>
<td>0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>67</td>
<td>17.84</td>
<td>0.00</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Note. $\Delta R^2$ = effect size
Appendix 3. Artistic Scale Differential Item Functioning (DIF) Indexes by the Logistic Regression (LR) Models

<table>
<thead>
<tr>
<th>Item</th>
<th>Simultaneous Uniform and Non-Uniform DIF $\chi^2$</th>
<th>$p$</th>
<th>$\Delta R^2$</th>
<th>Uniform DIF $\chi^2$</th>
<th>$p$</th>
<th>$\Delta R^2$</th>
<th>Non-Uniform DIF $\chi^2$</th>
<th>$p$</th>
<th>$\Delta R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.80</td>
<td>0.67</td>
<td>0.00</td>
<td>0.38</td>
<td>0.54</td>
<td>0.00</td>
<td>0.43</td>
<td>0.51</td>
<td>0.00</td>
</tr>
<tr>
<td>7</td>
<td>15.84</td>
<td>0.00</td>
<td>0.02</td>
<td>10.59</td>
<td>0.00</td>
<td>0.01</td>
<td>5.25</td>
<td>0.02</td>
<td>0.01</td>
</tr>
<tr>
<td>17</td>
<td>0.35</td>
<td>0.84</td>
<td>0.00</td>
<td>0.02</td>
<td>0.87</td>
<td>0.00</td>
<td>0.32</td>
<td>0.57</td>
<td>0.00</td>
</tr>
<tr>
<td>28</td>
<td>28.53</td>
<td>0.00</td>
<td>0.02</td>
<td>28.31</td>
<td>0.00</td>
<td>0.02</td>
<td>0.22</td>
<td>0.64</td>
<td>0.00</td>
</tr>
<tr>
<td>36</td>
<td>4.96</td>
<td>0.08</td>
<td>0.00</td>
<td>4.87</td>
<td>0.03</td>
<td>0.00</td>
<td>0.09</td>
<td>0.77</td>
<td>0.00</td>
</tr>
<tr>
<td>64</td>
<td>7.03</td>
<td>0.03</td>
<td>0.00</td>
<td>5.81</td>
<td>0.02</td>
<td>0.00</td>
<td>1.23</td>
<td>0.27</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Note. $\Delta R^2$ = effect size

Appendix 4. Social Scale Differential Item Functioning (DIF) Indexes by the Logistic Regression (LR) Models

<table>
<thead>
<tr>
<th>Item</th>
<th>Simultaneous Uniform and Non-Uniform DIF $\chi^2$</th>
<th>$p$</th>
<th>$\Delta R^2$</th>
<th>Uniform DIF $\chi^2$</th>
<th>$p$</th>
<th>$\Delta R^2$</th>
<th>Non-Uniform DIF $\chi^2$</th>
<th>$p$</th>
<th>$\Delta R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>14.44</td>
<td>0.00</td>
<td>0.01</td>
<td>1.71</td>
<td>0.19</td>
<td>0.00</td>
<td>12.73</td>
<td>0.00</td>
<td>0.01</td>
</tr>
<tr>
<td>29</td>
<td>10.23</td>
<td>0.01</td>
<td>0.01</td>
<td>0.02</td>
<td>0.88</td>
<td>0.00</td>
<td>10.21</td>
<td>0.00</td>
<td>0.01</td>
</tr>
<tr>
<td>39</td>
<td>17.37</td>
<td>0.00</td>
<td>0.01</td>
<td>7.68</td>
<td>0.01</td>
<td>0.00</td>
<td>9.68</td>
<td>0.00</td>
<td>0.01</td>
</tr>
<tr>
<td>45</td>
<td>17.11</td>
<td>0.00</td>
<td>0.01</td>
<td>13.87</td>
<td>0.00</td>
<td>0.01</td>
<td>3.24</td>
<td>0.07</td>
<td>0.00</td>
</tr>
<tr>
<td>46</td>
<td>17.98</td>
<td>0.00</td>
<td>0.01</td>
<td>17.77</td>
<td>0.00</td>
<td>0.02</td>
<td>0.21</td>
<td>0.65</td>
<td>0.00</td>
</tr>
<tr>
<td>50</td>
<td>20.56</td>
<td>0.00</td>
<td>0.03</td>
<td>11.45</td>
<td>0.00</td>
<td>0.01</td>
<td>9.11</td>
<td>0.00</td>
<td>0.02</td>
</tr>
<tr>
<td>54</td>
<td>1.90</td>
<td>0.39</td>
<td>0.00</td>
<td>1.01</td>
<td>0.31</td>
<td>0.00</td>
<td>0.89</td>
<td>0.34</td>
<td>0.00</td>
</tr>
<tr>
<td>55</td>
<td>0.83</td>
<td>0.66</td>
<td>0.00</td>
<td>0.08</td>
<td>0.78</td>
<td>0.00</td>
<td>0.75</td>
<td>0.39</td>
<td>0.00</td>
</tr>
<tr>
<td>62</td>
<td>33.03</td>
<td>0.00</td>
<td>0.02</td>
<td>29.18</td>
<td>0.00</td>
<td>0.02</td>
<td>3.85</td>
<td>0.05</td>
<td>0.00</td>
</tr>
<tr>
<td>70</td>
<td>22.14</td>
<td>0.00</td>
<td>0.01</td>
<td>20.98</td>
<td>0.00</td>
<td>0.01</td>
<td>1.16</td>
<td>0.28</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Note. $\Delta R^2$ = effect size
## Enterprising Scale Differential Item Functioning (DIF) Indexes by the Logistic Regression (LR) Models

<table>
<thead>
<tr>
<th>Item</th>
<th>Non-Uniform DIF $\chi^2$</th>
<th>p</th>
<th>$\Delta R^2$</th>
<th>Uniform DIF $\chi^2$</th>
<th>p</th>
<th>$\Delta R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>22.90</td>
<td>0.00</td>
<td>0.02</td>
<td>22.83</td>
<td>0.00</td>
<td>0.02</td>
</tr>
<tr>
<td>29</td>
<td>2.27</td>
<td>0.32</td>
<td>0.00</td>
<td>1.99</td>
<td>0.16</td>
<td>0.00</td>
</tr>
<tr>
<td>37</td>
<td>7.48</td>
<td>0.02</td>
<td>0.01</td>
<td>7.17</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>38</td>
<td>21.63</td>
<td>0.00</td>
<td>0.01</td>
<td>14.35</td>
<td>0.00</td>
<td>0.01</td>
</tr>
<tr>
<td>41</td>
<td>8.78</td>
<td>0.01</td>
<td>0.01</td>
<td>5.63</td>
<td>0.02</td>
<td>0.00</td>
</tr>
<tr>
<td>44</td>
<td>77.91</td>
<td>0.00</td>
<td>0.05</td>
<td>64.74</td>
<td>0.00</td>
<td>0.04</td>
</tr>
<tr>
<td>56</td>
<td>1.61</td>
<td>0.45</td>
<td>0.00</td>
<td>0.62</td>
<td>0.43</td>
<td>0.00</td>
</tr>
<tr>
<td>57</td>
<td>20.26</td>
<td>0.00</td>
<td>0.01</td>
<td>15.31</td>
<td>0.00</td>
<td>0.01</td>
</tr>
<tr>
<td>59</td>
<td>4.87</td>
<td>0.09</td>
<td>0.00</td>
<td>1.72</td>
<td>0.19</td>
<td>0.00</td>
</tr>
<tr>
<td>61</td>
<td>7.15</td>
<td>0.03</td>
<td>0.00</td>
<td>5.83</td>
<td>0.02</td>
<td>0.00</td>
</tr>
<tr>
<td>63</td>
<td>6.67</td>
<td>0.04</td>
<td>0.00</td>
<td>4.76</td>
<td>0.03</td>
<td>0.00</td>
</tr>
<tr>
<td>68</td>
<td>0.11</td>
<td>0.94</td>
<td>0.00</td>
<td>0.11</td>
<td>0.74</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Note. $\Delta R^2 =$ effect size