Cross-Cultural Psychometrics and the Revised NEO Personality Inventory (NEO PI-R):

The Challenges of Applying Psychometric Instruments in Cross-Cultural I/O Psychology

Andy Schumacher, MBA

Ph.D. Program – Industrial and Organizational Psychology

Capella University, 2010

Abstract

The continuing globalization of workplaces, workforce, and technology demands of organizational leaders to become cross-culturally competent. The influence of personality as one of the elements shaping the potential for effective international leadership, motivation, and performance appears undeniable. Assessing personality traits within a cross-cultural context using a valid and reliable framework would therefore provide significant benefit when comparing and selecting managers for global, expatriate assignments. The revised NEO Personality Inventory Test (NEO PI-R), a broadly accepted psychometric tool to assess personality dimensions along the five factor model (FFM), will therefore be evaluated to showcase the challenges of applying psychometric measurement tools in cross-cultural I/O psychology. Particular focus of the discussion will rest on possibly existing cultural influences on both validity and reliability of items, scales, and test scores derived from NEO PI-R. It is hypothesized that, such differences will not reduce the tools validity and reliability, if the overall construct of leadership personality and its relevant personality dimensions maintain convergent validity with other, valid cross-cultural research findings (e.g. Hofstede's framework of cultural implications on leadership and motivation).

Table of Contents

Cross-Cultural Psychometrics

In the field of cross-cultural I/O psychology cultural dimensions and determinants of both leadership and performance have been at the forefront of research for a number of years (Hofstede, 1980, 1993; Aycan, 2000; Dickson, Den Hartog, & Mitchelson, 2003). With continuing globalization of workplaces, workforce, and technology, the demands on organizational leaders to become cross-culturally competent are expected to significantly increase over the next decade (Singelis, 2009). Authors such as Harris and Kumra (2000) have cited leader personality as a key contributor to successful expatriate selection, acculturation and international management. According to the researchers international management competencies include the creation of a common business culture based on effective interpersonal interactions. Soft skills such as sensitivity to other cultures, the capacity to communicate non-judgmental respect, as well as demonstrating tolerance for potentially significant ambiguity define such interactions, and appear to be directly linked to the personality of an individual. The assessment of personality traits within a cross-cultural context using a valid and reliable framework would therefore provide significant benefit when comparing and selecting managers for global, expatriate assignments. Such assessment, however, would need to measure a consistent construct of leadership and its personality-oriented characteristics in order to remain cross-culturally valid. Typically, leadership theories differentiate between trait-based and behavioral approaches (Judge, Bono, Illies, & Gerhardt, 2002). While the trait-based theory presumes an individual to have distinct and intrinsic qualities allowing for leader emergence, behavioral theories suggest that, leadership, in fact, can be taught, and thus is a result of observation, mentoring, reflection, and

behavior adaptation (p. 766). As such, aligning measurable personality dimensions with leadership would follow a trait-based theory with the goal to identify those dimensions and individual facets of personality conducive to the management of people. In order to assess leadership personality in a cross-cultural context such traits would therefore need to be stable, measurable, and valid across cultures to define the actual, cross-cultural leadership construct. For the following analysis the personality and leadership framework presented by Judge, Bono, Illies, and Gerhard (2002) will be used to define such construct. Using the five factor model of personality (FFM) the authors found the traits of extraversion, openness to experience, conscientiousness, and agreeableness as 90% generalizable and conducive to determining leadership. However, typical leadership theory would suggest that such traits would be more applicable to predicting leader emergence rather than actual leader effectiveness. As such, these traits need to be operationalized into specific behaviors in order to infer personality-rooted leadership behaviors. Within a cross-cultural context, such traits could for example translate into behavioral facets such as creativity, change-orientation, and adaptability (i.e. openness to experience and extraversion), and cooperation, achievement motivation, and diligence (i.e. agreeableness and conscientiousness). Judge et al. compared the correlations relative to such behaviors to the identified personality traits and found similar positive relationship for leader effectiveness (p. 772). As such, the chosen framework appears to provide a valid, normative baseline for assessing leadership personality.

Using this construct of trait-based leadership personality the following analysis will review the application of the NEO PI-R, a broadly accepted psychometric tool to assess

personality dimensions along the five factor model (FFM), and showcase the challenges of applying psychometric measurement tools in cross-cultural I/O psychology.

Personality and Culture – Implications for Psychometric Assessment

For over 30 years researchers have hypothesized a direct relationship between personality and culture (Allik, & Realo, 2009). Specifically, the assumption was that covariations among personality traits in, for example, English-speaking populations could be generalized across diverse cultures and languages. If this is true than the personality trait structure as established in the FFM is essentially universally applicable to all human beings regardless of origin culture (p. 718). Hofstede and Mc Crae (2004) support such linkage between country-level personalities and cultural dimensions as defined in Hofstede's (1980) earlier framework, which related cultural aspects of management to the value system of both the leader and the cultural environment. Their findings appear to provide evidence to the idea that interactions of culture and personality are important not only for effective expatriate selection but also subsequent acculturation.

In contrast, Allik and Realo (2009) opine that the assumption of a universal personality trait structure may not be valid. While cultural beliefs influence the perceived personality traits of the members of such cultural group, they may not be easily validated in psychometric assessments. This is particularly true for Eastern, more collectivistic-oriented cultures. A Western culture-based personality test such as NEO PI-R may therefore introduce biases when interpreting test scores of Eastern cultural test takers (p. 150). For example, Asian cultures may

score lower on personality tests due to method bias and concept bias resulting from socially acceptable (i.e. self-effacing) and socially unacceptable (i.e. self-enhancing) behaviours. As such, a universal application of personality assessments may not provide meaningful information for inferential decision-making when trying to match an expected host-culture personality with the traits of a candidate for an expatriate assignment. In other words, the construct of a leadership personality may not be valid across cultures, if personality traits in themselves are not similarly interpreted, applied, or otherwise demonstrated. In order to increase the validity and reliability of such assessments, test users would have to ensure that any form of bias as a result of cross-cultural application is being eliminated or at least sufficiently minimized (Allik & Realo, 2009).

Linking Personality Traits and Cultural Dimensions relevant to I/O Psychology

Culture is defined as a common set of beliefs and values shared by all members of a specific group (Hofstede, 1980). In his broadly accepted framework, Hofstede classified the individual dimensions of culture by power distance, uncertainty avoidance, individualism-collectivism, and masculinity-femininity. Analyzing each of these dimensions revealed significant differences in perceived leadership emergence and ability between collectivistic and individualistic cultures. As such, psychometric leadership personality assessments need to consider both equivalent construct and facets of leadership personality if they are to be used within a cross-cultural context. In order to achieve such equivalence, a general framework of personality that could be universally applied with meaningful validity and reliability has to be established. The FFM currently represents such framework proven to be a useful tool across cultures. For example, Piedmont and Chae (1997) as well as Cheung, Leung, Fan, Song, Zhang,

and Zhang (1996) demonstrated the successful cross-cultural application of the FFM and its assessment tool NEO PI-R among Koreans and Chinese populations. In order to establish meaningful construct validity across cultures, the researchers needed to consider indigenous characteristics relevant to cross-cultural assessment and comparison. Similar to the approach by Cheung et al. (1996) Piedmont and colleague followed an etic and emic methodology to analyze the instrument's cross-cultural generalizability. Such approach focuses on the identification of both culturally-different traits (i.e. emic) and culturally-similar traits (i.e. etic) in order to isolate and measure indigenous characteristics inherent in each culture. Using a test-retest method with bilingual individuals in both English (Norm version) and translated Korean versions (criterion version) Piedmont and Chae (1997) demonstrated significant congruence coefficients between local and normative test items.

Cheung et al. (1996) provide similar findings of the applicability of the FFM for Chinese populations. The authors were able to develop sufficient intercultural overlap in content and construct descriptions within the context of the five personality dimensions, thus could successfully demonstrate meaningful convergent validity when relating the test items and scales to valid, Western instruments such as the NEO PI-R, MMPI, and the 16-F personality tests (p. 194).

Thus, using the FFM as a valid framework of personality, Hofstede and McCrae (2004) linked five-factor country-level personalities to cultural dimensions in 33 countries. The authors provided correlation evidence between dimensions and traits with the strongest relationships

found between extraversion and individualism, neuroticism and uncertainty avoidance, and conscientiousness and power distance.

In summary, the chosen examples demonstrate the importance of understanding the existing interdependencies and correlations between a cultural system and personality dimensions. Both will affect the validity and reliability of administering and interpreting personality assessments across various cultures due to their inclusion of universal as well as indigenous characteristics, which could introduce bias into cross-cultural, psychometric testing. Such biases will be discussed next.

Bias and Equivalence in Cross-Cultural Psychometric Application

In order to apply mono-culturally developed psychometric tools within a cross-cultural context, the concepts of test bias and test equivalence need to be considered (Van de Vijver, & Poortinga, 1997). Test equivalence is guaranteed, if construct, measurement, and scalar characteristics within cross-cultural testing contexts remain similar to those applied in mono-cultural conditions (Van de Vijver, & Tanzer, 2004). Specifically, construct equivalence of leadership personality requires similar definitions, and consequently similar traits and items across cultures to reliably represent such construct (p. 122). Moreover, measurement equivalence tries to incorporate known offsets between scales so achieved scores can be made comparable. To provide an analogy, a capo applied on a guitar in the second fret will shift all chords up by one full key making it easier for smaller hands to play certain barred chords. Knowing such offset

can make the actual chord playing comparable (i.e. achieving the same chord) to a guitar player with larger hands who is not using a capo.

Lastly, the highest level of equivalence is scalar or full scale equivalence. It assumes completely bias-free measurement, thus requires equivalence already achieved in construct and measurement (Van de Vijver, & Tanzer, 2004). Essentially, full scale equivalence uses the same scale across cultures, thereby maintaining the same unit of measure (p. 122). Naturally, such equivalence can only be achieved if scales are universally used and accepted to hold the same universal meaning (e.g. Fahrenheit or Celsius scale).

Even the quality of translation of actual tests and test items alone can have a significant impact on the equivalence of a psychometric test, particularly if cultural and linguistic meanings of test items and scales differ. Such discrepancies can introduce substantial test bias leading to invalid constructs or unreliable results. According to Van de Vijver and Pootinga (1997), three specific types of biases need to be predominantly considered; these being construct bias, method bias, and item bias.

Construct bias surfaces whenever definitions of the concept that is to be assessed do not provide a 100% match among cultures. In other words, if the construct of leadership personality defined by both FFM and applied frameworks such as that of Judge, Bono, Ilies, and Gerhardt (2002) does not fully overlap with leadership personality attributes prevalent in the foreign culture, psychometric assessments are not culturally universal, possibly invalid and not reliable as they measure dissimilar constructs.

Method bias, in turn describes the phenomenon that cultural factors irrelevant to the psychometric test can actually affect item and test scores. Van de Vijver and Poortinga (1997)

attribute the occurrence of such bias to differing aspects of social desirability (p. 30). For example, in collectivistic cultures, leadership personality may be associated with self-effacing, community-supporting traits and behaviours. As such, test items assessing self-presenting or self-enhancing traits would be viewed as socially undesirable and rated lower. Yet, such test scores would be interpreted as incongruent to possible leadership emergence, traits and effectiveness in individualistic, Western cultures.

The last prevalent form of bias revolves around the actual items themselves. Item bias or differential item functioning occurs when test takers show varying familiarity with individual test items ((Van de Vijver, & Poortinga, 1997). For example, a German student will answer a question about the capital of Germany more often correctly than a student from Micronesia. In other words, differential item functioning is related to the perceived difficulty of and familiarity with test items.

The findings above suggest that validity and reliability of psychometric tests does not merely transfer, as scores may have different psychological meanings within and across cultures (Van de Vijver, & Tanzer, 2004). Rather, cross-cultural test adaptation appears to require a "culture by culture" approach including comprehensive re-testing every time a test is to be applied in a different culture (p. 120). This is underlined by the fact that construct bias was found to be the most problematic issue when comparing collectivistic Eastern and individualistic Western cultures (p. 125). Obviously, a universally shared theoretical construct is often difficult to establish, define, and interpret. Particularly leadership personality assessments require indigenous content considerations in order to avoid construct bias. As already outlined earlier, to achieve cross-cultural equivalence, test scales and items need to include in their translation the

cultural meaning, and not just the linguistic meaning of the content. Realizing such dynamics in culturally universal test adaptation, some generic strategies providing a bias-free adaptation protocol will be discussed in a later chapter. First, the revised NEO Personality Inventory (NEO PI-R) (McCrae, & John, 1992) will be introduced as the chosen example for evaluation of its potential for cross-cultural application in determining leadership personality.

The revised NEO Personality Inventory (NEO PI-R)

Personality tests are very common assessments in the workplace. Inventories related to the FFM have become the baseline for many personality-oriented sections of pre-employment tests. Traits such conscientiousness and openness to experience have been found to influence coping with change-related stress, job attitudes, and turnover intentions (Maynard, Joseph, & Maynard, 2006). As such, assessing personality traits using a valid and reliable framework provides significant benefit when looking to select employees that are expected to exhibit high achievement motivation, diligence, and overall organizational citizenship behaviors.

The NEO PI-R conceived by Costa and McCrae (1978) assesses five major domains of adults normal, positive personality along their popular Five Factor Model; these being neuroticism, extraversion, openness to experience, agreeableness, and conscientiousness, with each represented by six lower level facet scale scores. The test aims to reliably predict interests, psychological well-being, as well as person-characteristic coping behaviors. Using a five-point Likert scale ranging from “Strongly agree” to “Strongly disagree” a self-report personality profile is being created which plots the individual item scores within each dimension against standardized norms.

According to McCrae and John (1992) the reliability of the NEO PI-R was assessed using both test-retest techniques as well as Cronbach's Alpha coefficient. Domain reliability ranged from 0.86 to .095. Kaplan and Saccuzzo's (2009) concluded that reliability of test instruments should range between 0.6-0.85, thus suggesting high relatedness, yet still offer new information. As such, the NEO PI-R exhibits excellent reliability. Individual personality facet levels also showed good reliabilities and fell between 0.56 and 0.90 for both self- and observer-report forms of the NEO-PI-R. In test-retest situations the domains of neuroticism, extroversion, and openness to experience showed long-term reliability. Since personality traits are expected to remain relatively stable over time, test-retest analysis provides additional data pointing to the reliability of the instrument (McCrae, Costa, Del Pilar, Rolland, & Parker, 1998).

Another influence on both an instrument's validity and reliability consists in the selecting proper norms and scales used to determine and assess participants' scores. Costa and McCrae (1978) established the norms based on a sampling of 1,000 male and female subjects, which matched the US Census projections for age, gender, and race. Such alignment of sampling characteristics provided a credible reflection of the total and multicultural population, thus creating norms that can be referred to as typical, standardized scores. In comparison, the scales used in the NEO PI-R were deemed valid as they correlated with other, in-kind scales used in instruments widely accepted for their validity, such as Jungian Types (e.g. Meyers-Briggs Type Indicator), needs and motives (e.g. Personality Research Form), or even psychopathology (i.e. Minnesota Multiphasic Personality Inventory).

While the psychometric properties of the NEO PI-R are impressive, they are not necessarily a guarantee for the tests' culturally universal application capabilities. As such, the

following chapters will evaluate its potential for equivalent cross-cultural application within the context of validity and reliability considerations.

Evaluating the NEO PI-R as a Cross-Cultural Assessment Tool

If the NEO PI-R is to provide meaningful information in cross-cultural personality assessments, various aspects related to its adaptation need to be considered. Among them are validity and its various sub-forms (i.e. construct, convergent, content, etc.) and reliability. These issues will be discussed next.

Validity and Reliability Considerations

Since the norms of the NEO PI-R are based on Census demographics, it could be argued that this personality assessment tool acknowledges and supports diversity in test takers. Issues, however, could arise, when this test is used within an international context, particularly in cross-cultural settings. Cheung et al. (1996) state that the challenge in cross-cultural personality assessment revolves around the creation of test inventories suited to represent both local and universal (i.e. norm) needs so that both validity and reliability of the original tool are being retained. In other words, are the traits of agreeableness and extroversion found to be determinants of the construct of leadership personality also indicative of leadership potential in other, more collectivistic cultures? In order to resolve such potential issues, various steps have to be taken before administering the test. First, an equivalent construct of leadership personality needs to be defined and assessed within its cultural parameters. This establishes the baseline for a benchmark

profile forming the norm against which cross-cultural candidates are being measured. Cheung and colleagues (1996) developed such construct validity by coupling extensive literature reviews with both qualitative and quantitative research methodologies. This allowed for the incorporation of indigenous characteristics and their categorization into the five personality dimensions of the FFM (p. 184).

Another influence on construct validity stems from the translation of individual test items defining the leadership personality construct. Such translations need to reflect both the linguistic and cultural meaning of the original test to ensure content and convergent validity. Van de Vijver and Poortinga (1997) recommend the process of back-and-forth translation by bi-lingual and bi-cultural individuals in order to ensure internal consistency; i.e. that items portray similar meaning relevant to the construct. Various researchers support this approach and have used it in their cross-cultural adaptation of NEO PI-R (Cheung et al., 1996; Mottus, Pullman, & Allik, 2006; Piedmont & Chae, 1997). For example, Mottus, Pullman, and Allik (2006) used subject matter experts to first determine pertinently adapted test items and then asked translators unfamiliar with both scales and items to provide their understanding and interpretation. This provided additional support for construct validity.

In addition, in their adaptation of the NEO PI-R for the Korean population, Piedmont and Chae (1997) ensured construct validity by (1) developing translated scales that remained internally consistent, (2) keeping the factor structure of scales aligned with the five major FFM dimensions of personality, as well as (3) leaving factor loading for facet scales consistent with rational placement and original normative structure. With validity scores of 0.63 on average, and reliability values ranging from 0.80 to 0.92 the FFM and its assessment through NEO PI-R were

found to generalize well to the Korean culture. These results also seem to support McCrae and colleagues' (1998) conclusions of personality only varying in intensity, not in actual dimension across cultures.

Another aspect of an equivalent cross-cultural adaptation centers on convergent validity among psychometric tests. Such validity is established if test items show strong correlations with other items featured in other valid psychometric assessments used to measure similar constructs. In the case of cross-cultural adaptation of NEO PI-R convergent validity between adapted and original versions was established by both parallel test administration (original and adapted versions) and relating items and scales of the adapted version to other valid personality measurement tools such as the MMPI or the 16-F (Cheung, et al., 1996). Cheung and coworkers reported convergent validity values of 0.70 for scales and test items. Mottus, Pullman, and Allik's (2006) administration of NEO PI-R in Estonia showed convergent validity values of 0.73 when compared to their local personality assessment tool (EPIP) and 0.89 when related to the International Personality Item Pool (IPIP) used effectively throughout Eastern Europe. These correlation values demonstrate good validity and transferability of NEO PI-R items when assessing the personality of the Estonian culture (p. 155).

Lastly, cross-cultural application of the NEO PI-R needs to ensure acceptable reliability of both test items and overall scales. Leedy and Ormrod (2010) define reliability as an instrument's potential to produce similar results at different times. It is the quasi quality seal of testing instruments assuring the researcher that measured results actually produce an accurate representation of the truth. Within the context of cross-cultural application reliability assesses if a culturally adapted version of a mono-cultural test effectively and consistently measures a similar

construct (i.e. leadership personality). As such the norms used in NEO PI-R need to be adjusted to reflect a standardized mean score that is reliable for the population to be tested, so to establish a meaningful benchmark against which the culturally adapted trait construct of leadership can be measured. De Fruyt, De Bolle, McCrae, Terraciano, and Costa (2009) deployed their abbreviated version of the NEO PI-R (NEO PI-3) across adolescents aged between 12 and 17 from 24 individualistic and collectivistic cultures. Using Cronbach's alpha, the researchers measured the internal consistency of test items to find moderate item reliability of 0.59 (p. 305). Other adaptations, such as those of Mottus, Pullman, and Allik (2006) produced much higher reliability values ranging from 0.88 to 0.95 across the five personality domains. McCrae, Costa, Del Pilar, Rolland, and Parker's (1998) study developed similar reliability values among Western and Eastern cultures leading the authors to conclude that personality as described in the FFM is a universal construct of human beings, thus not very susceptible to cultural differences (p. 172). More specifically, cultural differences revolve more around the actual intensity and type of personality dimension (i.e. extraversion versus introversion) rather than individual facets within these dimensions (p. 182). This would imply that personality becomes a trans-cultural phenomenon, with the cultural component being relevant to the intensity of adaptation of particular dimensions.

Bias and Equivalence

According to classical testing theory every test taker can produce a true score if the testing instrument was 100% reliable and without errors. Since attaining such reliability seems an impossible feat, psychological testing can only produce an observed score; in essence, a score

that possibly only reflects a close approximation of the true score (Kaplan & Saccuzzo, 2009). Within the context of cross-cultural application of psychometric tests, however, producing a consistent reliability across international test subjects is critical, if one aims to assess and compare behaviors, performance, or traits against meaningful norms. Van de Vijver and Poortinga (1997) found that, equivalence is best achieved when pursuing an integrated, triangulated approach, as only then all three major types of biases, which often appear to happen simultaneously, can be controlled.

As discussed earlier, Van de Vijver and Poortinga (1997) as well as Van de Vijver and Tanzer (2004) highlight construct, method, and item bias as the three main impediments standing in the way of achieving cross-cultural equivalence. Relating these biases specifically to the NEO PI-R researchers must be cognizant of all latent aspects affecting the interpretation of the observed score of personality. For example, when evaluating the construct of leadership personality within a cross-cultural context, the definitions of such construct must translate universally and completely cover all relevant facets (Van de Vijver & Tanzer, 2004). Additionally, these facets need to be discernable by the five personality dimensions. Furthermore, differential appropriates of traits and behaviors associated with leadership personality need to be considered (p. 124). This is particularly important when specific attributes are not socially desirable thus do not belong with the cultural group to be tested, such as the concept of ‘losing face’ and ‘disrupt harmony’ in Chinese cultures (Cheung et al., 1996).

In terms of method bias, translations of items and scales relating to the individual dimensions need to be free of ambiguity, confirmation bias, and differing levels of social desirability (Van de Vijver, & Tanzer, 2004). For example, De Fruyt, De Bolle, McCrae,

Terraciano, and Costa (2009) found that it is socially desirable in Malaysian, Chinese, and Korean cultures to express conscientiousness by appearing less energetic (i.e. lower extroversion) and more intellectual (i.e. higher introversion). As such, an equivalent method of assessing the construct of leadership personality in these cultures needs to consider traits that suggest intellectual curiosity, yet somewhat docile behavior (p. 310).

Lastly, item bias can render the result of a psychometric test invalid and unreliable. Within the evaluation of the NEO PI-R, items related to individual personality facets or dimensions need to be properly and culturally specific translated in order to reduce any incidental differences in connotative meaning of the item content (Van de Vijver, & Tanzer, 2004). Specifically, items should not invoke additional traits that could be potentially misinterpreted (p. 124).

Strategies for a Culturally Universal Application of NEO PI-R

The NEO PI-R has thus far been reliably administered in over 50 different cultures (Cheung et al., 1996; De Fruyt, De Bolle, McCrae, Terraciano, & Costa, 2009; McCrae, Costa, Del Pilar, Rolland, & Parker, 1998; Mottus, Pullman, & Allik, 2006). At the core of such administrations is the recognition of emic and etic characteristics influencing both test validity and reliability. Van de Vijver, and Tanzer (2004) suggest various strategies to indentify and correct bias in cross-cultural assessments. Most of these have been implemented in the referenced studies.

For example, the authors suggest either decentering or convergent approaches when dealing with construct bias (Van de Vijver, & Tanzer, 2004). Decentering in this context means

to simultaneously develop the same instrument in several cultures (p. 128). Cheung et al. (1996) applied this strategy when developing their Chinese personality inventory assessment along items from the FFM and the NEO PI-R. In contrast, convergence strategies suggest an independent within-culture development of instruments and subsequent cross-cultural administration of all instruments (Van de Vijver, & Tanzer, 2004). Mottus, Pullman, and Allik (2006) used this approach when building correlation data between the NEO PI-R and the Estonian Personality Inventory. Furthermore, all of the above focused on linguistically and culturally appropriate translations of test items and scales, which were performed by both native-speaking, local culture experts as well as bilingual individuals (Hofstede & McCrae, 2004). When using this technique, Hofstede and McCrae found that bilingual individuals who completed the NEO PI-R scored with only little to no mean-level differences on the two administrations in host and origin language (p. 67). It is questionable, however, if such strategy is in fact a valid method of proving cultural universality of a construct, and thus, indicating construct validity. It could, for instance, be argued that bilingual test takers exhibit test-wise behaviors (i.e. remember what a specific item meant in the language in which they are most comfortable), thus draw on the experience and memory when responding to test items. If this is the case, then such strategy would not support the validity assumptions of a cross-culturally administered test.

Van de Vijver, and Tanzer (2004) further suggest extensive (cultural sensitivity) training of test users, along with providing them with detailed protocols, manuals, and instructions on scoring and interpretation to minimize possible method bias. Within this context, test-retest methodologies may be very helpful to gain additional insights into test reliability (p. 128).

McCrae, Costa, Del Pilar, Rolland, and Parker (1998) developed test-retest reliability data ranging from 0.89 to 0.92 for their cross-cultural administration of NEO PI-R. Other researchers (De Fruyt, De Bolle, McCrae, Terraciano, & Costa, 2009; Mottus, Pullman, & Allik, 2006) also found moderate to good reliability in test-retest situations, which hovered around 0.59 in the case of DeFruyt and colleagues, and 0.80 for Mottus et al.

Resulting form poor translations or low appropriateness of specific test items, item bias is the last prevalent form of test distortion that will be considered. Here, Van de Vijver and Tanzer (2004) suggest differential item functioning analysis (i.e. presenting culturally different individuals with similar test items and compare their responses) along with a documentation of 'spare items'. These items could be identified in the manual as holding similar meaning to active test items, thus could replace them without impact on the measurement of the construct (p. 128). De Fruyt, De Bolle, McCrae, Terraciano, and Costa (2009) found that item replacements in the NEO PI-R and NEO PI-3 on average improved the item-total correlations across languages. Hofstede and McCrae (2004) produced similar results suggesting an overall similar personality factor structure in a wide variety of cultures.

In summary, the meaningful cross-cultural application of NEO PI-R requires researchers to consider and control construct, method, and item biases resulting from indigenous characteristics. Since reliability of test instruments immediately impacts the quality of test scores and subsequent decision making, test administrators should be painfully familiar with the interpretation of the reliability construct as well as its implications on the quality of the information obtained. Furthermore, the methods of calculating reliability should be aligned with

the type of scoring norms and scales used in the test. For example, internal consistency should be used for assessing reliability of qualitative test items measuring agreement or disagreement, whereas reliability of dichotomous right or wrong items could be determined through test-retest assessment.

Standards and Principles relevant to a Cross-Cultural Administration of NEO PI-R

Section 9 of the American Psychological Association's (APA, 2002) ethical principles and code of conduct contains guidelines for administration, use, and interpretation of psychological testing. These guidelines require psychologists to unwaveringly maintain the respect for persons, provide informed consent to test takers, and administer tests only for overall beneficence. While respect for persons revolves around the idea of maintaining a person's dignity, the tenets of beneficence and justice safeguard that psychological testing is conducted only in the best interest of human beings, and research outcomes (i.e. benefits and risks) are being equally distributed among all segments of population (Turner, DeMers, Fox, & Reed, 2001). According to Turner and colleagues, APA's efforts to provide guidelines for the necessary knowledge, skills, abilities, and training aimed at establishing such a baseline of test taker protection.

Psychometric Testing Standards

The American Educational Research Association's (AERA, 2008) stringent standards for psychological testing add to the ethical requirements by more formally stressing the principles of justice and fairness in test use across cultures. At the forefront of consideration are thorough test

validation and reliability; both deemed critical to producing scientifically sound reasoning for inferences from and interpretations of test scores. Important standards for cross-cultural test use include 5.1 Standardization of test procedures, 11.2 Evidence of purposeful test validity and reliability, 11.3 in conjunction with 11.4 Proof of validity and reliability, 12.2 Bias, 12.3 Alignment of test selection with candidate background, and others (AERA, 2008). Particularly standards 11.3, 12.2, and 12.3 carry implications for assessing diverse or multicultural clients. For example, when applying the NEO PI-R personality test across 33 cultures, McCrae, Costa, Del Pilar, Rolland, and Parker (1998) found indigenous traits that significantly influenced particular facets within the personality dimensions of the Five Factor Model. While the actual dimensions remained valid and culturally universal, it is critical to acknowledge the difference in scoring for individual attributes in order to detect and avoid bias leading to ill-informed decision making (Van de Vijver, & Tanzer, 2004).

Ethical Principles relevant to Psychometric Assessments

Ethical principles in psychometric assessments are likewise ultimately determined by both validity and reliability. As discussed above, AERA standards (2008) clearly define the rights and responsibilities of both test users and test takers to ensure fairness in administration, scoring, and result interpretation. Ultimately, test takers and test users are looking for just conditions when engaging in testing processes. Balancing standardized tests and universal equality, however, may well become a goal conflict, from which potentially only an approximation of justice can arise. Since the validity (and lack thereof) of test results can have a significant impact on a person's current or future professional and personal life, testing experts need to be cognizant

to abide by established standards in order to avoid (cultural or gender) discrimination and potentially costly and reputation-damaging litigation.

Conclusion

In the field of cross-cultural I/O Psychology, psychometric testing can provide additional data points allowing for the interpretation and prediction of personality, attitudes, and behavior responses in the work environment. In order to ensure ethical test administration, scoring, and result interpretation psychologists need to be profoundly familiar with the different concepts of validity and reliability. Particularly in a multicultural context, appropriate steps need to be taken to ensure cultural universality. As psychometric tools are often used in predicting employee behavior on the job, their administration should only be performed if test users and their clients are completely informed about the tool's obvious limitations, as only then can score interpretations avoid the causing of harm of test takers.

References

American Educational Research Association (2008). *Standards for educational and psychological testing*. Washington, DC: AERA.

American Psychological Association. (2002). *Ethical principles of psychologists and code of conduct*. Retrieved May 26, 2010 from http://www.apa.org/ethics /code/index.aspx.

Aycan, Z. (2000). Cross-cultural industrial and organizational psychology: Contributions, past developments, and future directions. *Journal of Cross-Cultural Psychology*, *31*(1), 110-128.

Allik, J., & Realo, A. (2009). Personality and Culture - Editorial. *European Journal of Personality*, *23*, 149-152.

Cheung, F. M., Leung, K., Fan, R. M., Song, W. Z., Zhang, J. X., & Zhang, J.P. (1996). Development of the Chinese personality assessment inventory. *Journal of Cross-Cultural Psychology*, *27* (2), 181-199.

Costa, P., & McCrae, R. (1978). Revised NEO Personality Inventory (NEO PI-R). *Mental Measurements Yearbook 2004*, *12*, no page numbers.

De Fruyt, F., De Bolle, M., McCrae, R. R., Terraciano, A., & Costa, P. T. (2009). Assessing the universal structure of personality in early adolescence: The NEO-PI-R and NEO-PI-3 in 24 cultures. *Assessment*, *16*(3), 301-311.

Dickson, M. W., Den Hartog, D. N., & Mitchelson, J. K. (2003). Research on leadership in a cross-cultural context: Making progress and raising new questions. *The Leadership Quarterly*, *14*, 729-768

Harris, H., & Kumra, S. (2000). International manager development: Cross-cultural training in highly diverse environments. *Journal of Management Development*, *19*(7), 602-614.

Hofstede, G. (1980). Motivation, leadership, and organizations: Do American theories apply abroad? *Organizational Dynamics*, *9*(1), 42-63.

Hostede, G., & McCrae, R. R. (2004). Personality and culture revisited: Linking traits and dimensions of culture. *Cross-cultural Research*, *38*(1), 52-88.

Judge, T. A., Bono, J. E., Ilies, R. & Gerhardt, M. W. (2002). Personality and leadership: A qualitative and quantitative review. *Journal of Applied Psychology*, *87*(4), 765-780.

Kaplan, R. M., & Saccuzzo, D.P. (2009). *Psychological testing: Principles, applications, and issues* (7th ed.). Belmont, CA: Wadsworth Cengage.

Leedy, P. D., & Ormrod, J. E. (2010). *Practical research: Planning and design* (9th ed.). Upper Saddle River, NJ: Pearson.

Maynard, D. C., Joseph, T. A., & Maynard, A. M. (2006). Underemployment, job attitudes, and turnover intentions. *Journal of Organizational Behavior*, *27*, 509-536.

McCrae, R. R., Costa, P. T., Del Pilar, G. H., Rolland, J. P., & Parker, W. D. (1998). Cross-cultural assessment of the five factor model: The revised NEO personality inventory. *Journal of Cross-Cultural Psychology*, *29*(1), 171-188.

McCrae, R. R., & John, O. P. (1992). An introduction to the five-factor model and its applications. *Journal of Personality*, *60*(2), 175–215.

Mottus, R., Pullmann, H., & Allik, J. (2006). Toward more readable big five personality inventories. *European Journal of Psychological Assessment*, *22*(3), 149-157.

Piedmont, R. L., & Chae, J. H. (1997). Cross-cultural generalizability of the five-factor model of personality: Development and validation of the NEO PI-R for Koreans. *Journal of Cross-Cultural Psychology*, *28*(2), 131-155.

Singelis, T. M. (2009). Some thoughts on the future of cross-cultural social psychology. *Journal of Cross-Cultural Psychology*, *31*(1), 76-91.

Turner, S., DeMers, S., Fox, H., & Reed, G. (2001). APA's guidelines for test user qualifications: An executive summary. *American Psychologist*, *56*(12), 1099-1113.

Van de Vijver, F. J. R., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment, 13*(1), 29-37.

Van de Vijver, F., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *Revue europeenne de psychologie appliquee', 54*, 119-135.