EXAMINING MEASUREMENT INVARIANCE OF THE MMPI-2 RESTRUCTURED FORM EXTERNALIZING SPECIFIC PROBLEMS SCALES USING AMERICAN AND KOREAN NORMATIVE AND CLINICAL SAMPLES

Jiebing Wang

A dissertation submitted in partial fulfillment of
the requirements for the degree of
Doctor of Philosophy

Department of Psychology

Central Michigan University
Mount Pleasant, Michigan
August 2017

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Dr. Kyunghee Han, for her continual encouragement, guidance, and support throughout this project; my committee members, Dr. Stephen Colarelli, Dr. Deb Poole, and Dr. Nathan Weed, for their insightful suggestions and feedback; Dr. Kyunghee Han and Dr. Yossef Ben-Porath, for their datasets; Robin Decker, Barb Houghton, and Betty Schulte, for their help and support. I am grateful for the continual support and prayers of my husband and my family. Last and mostly, I would like to thank my Heavenly Father!

ABSTRACT

EXAMINING MEASUREMENT INVARIANCE OF THE MMPI-2 RESTRUCTURED FORM EXTERNALIZING SPECIFIC PROBLEMS SCALES USING AMERICAN AND KOREAN NORMATIVE AND CLINICAL SAMPLES

by Jiebing Wang

This study examined the measurement invariance of the Minnesota Multiphasic Personality Inventory-2 Restructured Form (MMPI-2-RF; Ben-Porath & Tellegen et al., 2008) Externalizing Specific Problems Scales—Juvenile Conduct Problems (JCP), and Substance Abuse (SUB), Aggression (AGG), and Activation (ACT) Scales—across American and Korean normative and clinical samples. First, I tested the measurement invariance of the Scales across genders for each culture group to provide information on comparability of the construct across genders and found some gender noninvariant items. Then, I further examined the measurement invariance of the Scales across cultures (American clinical vs. Korean clinical, American normative vs. Korean normative) using a hybrid multi-group–MIMIC model by incorporating gender noninvariant items. Partial scalar invariance was achieved for all four Externalizing SP Scales, and some noninvariant items were identified across cultures. Compared to the results from measurement invariance without considering gender, there were fewer noninvariant items across cultures when incorporating gender noninvariant items in examining measurement invariance. It is possible that gender, a largely overlooked variable by researchers when conducting measurement invariance across cultures, may partial out some amount of variance appears across cultures.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

FIGURE PAGE

CHAPTER I

INTRODUCTION

The Minnesota Multiphasic Personality Inventory-2 (MMPI-2; Butcher et al., 2001) has been used extensively to measure personality and psychopathology. The MMPI-2 Clinical Scales were constructed using an empirical construction strategy, which caused the scales to become heterogeneous in item content and substantially correlated. To overcome these major shortcomings, Tellegen and his colleagues developed the Restructured Clinical Scales (MMPI-2 RC; Tellegen et al., 2003). Further, Ben-Porath and Tellegen (2008) added five additional sets of scales in addition to RC Scales and constructed a more concise MMPI-2 Restructured Form (MMPI-2 RF; Ben-Porath & Tellegen, 2008). Compared to the MMPI-2 of 567 items, the MMPI-2 RF contains 338 items representing nine Validity Scales, three High-Order Scales, nine RC Scales, two interest Scales, 23 Specific Problem (SP) Scales, and revised Personality Psychopathology Five (PSY-5) Scales.

Externalizing Specific Problems (SP) Scales

The development of Specific Problems (SP) Scales was intended to (a) cover some clinical constructs in the MMPI item pool that were not represented in the RC Scales (e.g., suicidality) and (b) "highlight important characteristics that are subsumed by or associated with one of the Restructured Clinical (RC) Scales but are not exclusively or directly targeted by that scale" (Ben-Porath & Tellegen, 2008, p. 50). Four sets of SP Scales were developed based on conceptual considerations and empirical analyses: five Somatic/Cognitive, nine Internalizing, four Externalizing, and five Interpersonal Functioning Scales. The four Externalizing Scales are related to two RC Scales: Juvenile Conduct Problems (JCP) and Substance Abuse (SUB) Scales measure two specific facets of RC4 (Antisocial Behavior) Scale; Aggression (AGG) and

Activation (ACT) measure two facets of RC9 (Hypomanic Activation) Scale. The four Externalizing Scales can be used to identify the elevations on RC4 and RC9 but should be interpreted independently from the related RC Scales (Ben-Porath & Tellegen, 2008).

The JCP Scale includes six items which describe a history of school problems, stealing, and negative peer pressure. An elevated score on the JCP Scale indicates juvenile delinquency history and current misbehavior. The SUB Scale consists of seven items describing past or current substance abuse, especially alcohol abuse. The AGG Scale includes nine items which describe physically aggressive behavior. An elevated AGG Scale score may be associated with interpersonal violence. The ACT Scale consists of eight items describing high energy level, lack of sleep, and mood swings. An elevated score on the ACT Scale may be associated with hypomanic or manic episodes; however, drug-induced activation needs to be taken into consideration. For the four externalizing SP Scales, the test-retest reliability coefficients ranged from .77 to .87 in a normative sample; the internal consistency coefficients ranged from .56 to .66 for a normative sample, .59 to .75 for an outpatient sample, and .71 to .77 for an inpatient sample, respectively (Tellegen & Ben-Porath, 2008).

Regarding predictive validities, the JCP Scale predicts premature termination of therapy and therapy absence when controlling other related MMPI-2 RF scales in a university-based psychology clinic (Anestis, Gorrfried, & Joiner, 2015). In a sample of individuals who are at risk for attrition from Drug Court treatment, JCP is one of the scales that are correlated with increased risk for failing to complete (Mattson, Powers, Halfaker, Akeson, & Ben-Porath, 2012). SUB and RC3 (Cynicism) predicted the substance related disorders in the *Diagnosis and Statistical Manual of Mental Disorder, Fourth Edition Text Revision (DSM-IV-TR)* in a university clinic sample (Haber, 2013). Substance abuse is one of the bariatric surgical contraindications, and significant classification accuracy is gained by lowering the interpretive cutoff on the SUB Scale among bariatric surgery candidates (Tarescavage et al., 2013). AGG was highly correlated with anger behaviors for both men and women in a college setting, and the effect was similar across gender (Forbey, Lee, & Handel, 2010). The ACT Scale was most useful in differentiating bipolar disorder from other diagnostic groups in a patient sample (Sellbom, Bagby, Kushner, Quilty, & Ayearst, 2012).

International Adaptation of MMPI-2 RF and Measurement Invariance Tests

The MMPI-2 RF has been translated into many languages including Korean. Psychologists who study personality and culture are interested in whether personality constructs are universal or culture-specific. When translating and importing personality inventories built based on Western culture into non-Western cultures, distinctions between universal and culture-specific constructs can be made by comparing scale means across cultures. However, the measurement invariance issue needs to be considered before comparing constructs cross-culturally to make sure that the inventory is measuring similar constructs across cultures, that is, the probability of item response is similar among different culture groups with the same underlying construct. Unless measurement is invariant across cultures, the mean differences among cultures can be interpret as genuine differences.

Measurement invariance is the measurement equivalence of the underlying construct in two or more groups and is most frequently tested at configural, metric, and scalar levels (Chen, 2008; Vandenberg & Lance, 2000). Group mean comparisons at the scale level are not valid until the assumption of the measurement invariance holds. Configural (or construct) invariance, the most basic level of invariance, is achieved when the number of factors underlying a construct is the same across groups. For example, configural invariance of the Externalizing SP Scale is reached if all items are best represented by a single factor model across groups. Metric (or factor loading) invariance indicates that the factor loadings, the magnitude of the relationships between items and the construct, are equivalent across groups. For example, metric invariance is achieved when the magnitude of the relationship between each item and its SP construct are similar across groups. Scalar (or intercept) invariance demonstrates that the intercept—point of origin—is equal across groups. For example, scalar invariance of the SP Scale indicates that the likelihood of endorsing "yes" to an SP item is similar across groups when both groups have the same SP level.

Two frequently used statistical methods for testing measurement invariance are multiple-group confirmatory factor analysis (MGCFA) and a multiple-indicators multiple-causes (MIMIC) model. In addition to testing measurement invariance, both methods identify noninvariant items—that is, items lacking measurement invariance at the item level. However, each method has its own strength and limitation. By using a series of hierarchical models, MGCFA enables researchers to test for equivalence of the factor structure, factor loadings, intercepts, residual variances, and other features of the construct across two or more groups (Vandenberg & Lance, 2000). Despite allowing researchers to examine the invariance of all model parameters, a weakness of the MGCFA is that invariance can be tested only across levels of a single categorical variable. Compared to the MGCFA method, the MIMIC method can be used to test across levels of both categorical and continuous variables, but it only allows researchers to test the invariance of two model parameters (factor means and thresholds) and assumes that all other model parameters are equal across groups (Brown, 2006).

The Korean MMPI-2 was initially translated through a process of multiple rounds of translation (English to Korean) and back-translation (Korean to English) by Han and pilot tested in a Korean college sample (Han, 1993, 1996). After revising items with minor translation problems, the final version of the Korean MMPI-2 was published in 2005 and standardized using a Korean adult normative sample (Kim, Han, Lim, Lee, Min, & Moon, 2005). The test-retest reliabilities and validity of the MMPI-2 in the Korean normative sample were comparable to those in the American normative sample. The Korean MMPI-2 was further validated by using a Korean psychiatric sample and showed robust validity by its correlation with SCL-90-R scales, behavioral correlates, and therapist ratings. Later, the Korean MMPI-2 RF was published in 2011 and standardized using the Korean MMPI-2 normative sample with minor modifications (Han, Moon, Lee, & Kim, 2011).

Only two studies have examined the measurement invariance of the MMPI-2/MMPI-2 RF across American and Korean samples. Ketterer (2010) examined the measurement invariance of the MMPI-2 Restructured Clinical (RC) Scales across American and Korean normative samples and found that most RC Scales showed poor fit in a one-factor model, which failed to meet the prerequisite for further testing measurement invariance. Ketterer suggested that the RC Scales might fit better in multi-factor models. Han and colleagues (2011) followed Ketterer's suggestion and identified underlying clusters for each of the RC Scales using rational judgment. Wang (2013) further tested the measurement invariance of the MMPI-2 RC4 Scale across American and Korean clinical samples and found that RC4 Scale achieved partial scalar invariance in a rationally-derived four-factor model (RC4 four-factor model; Han et al., 2011). The present study was designed to examine the measurement invariance of the Externalizing SP Scales across American and Korean normative and clinical samples in order to add knowledge to the literature of measurement invariance of the MMPI-2/MMPI-2 RF across cultures.

Gender and Externalization of Psychological Symptoms

Several epidemiological studies showed no overall gender difference in rates of psychopathology; however, women were higher than men in rates of internalizing disorders (e.g., depression and anxiety), and men were higher than women in externalizing disorders (e.g., substance abuse and antisocial disorders) (Rosenfiled, 1999). When using the *Diagnosistic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV)* for diagnosis, Hartung and Widiger (1998) found that 84% of the DSM-IV disorders reported having different diagnostic rates across gender. Men were more likely to be diagnosed with gender identity, impulsive, sexual, and substance-related disorders, whereas women were more likely to be diagnosed with anxiety, depression, dissociative, eating, and somatization disorders. Some of these gender differences in adulthood can be traced back to childhood. For example, internalizing problems (e.g., anxiety disorders) are more common in girls than boys, whereas externalizing problems (e.g., conduct disorders) are more common in boys (Crick & Zahn-Waxler, 2003). The literature review on gender and Externalizing Specific Problem Scale constructs is discussed below.

The DSM-IV TR reported that the lifetime prevalence rates of conduct disorder vary, ranging from 1 to 10% among studies, but that conduct disorder is more common among males than females, with prevalence rates ranging from 2.1 to 8.8% in males and from 0 to 1.4% in females (Doll, 1996; Maughan, Rowe, Messer, Goodman, & Metzer, 2004). There is a debate about whether the difference in prevalence rates across gender truly reflects gender differences or merely a sex bias in the diagnostic criteria. Moffitt et al. (2008) proposed that DSM-IV conduct disorder symptom criteria focus on behaviors that are more common in boys and may fail to detect conduct disorder among girls. For example, physical aggression behaviors are more common in boys than girls; however, behaviors such as relational aggression are more common in girls and are not included in the diagnostic symptom criteria.

Reviews have shown that males are more likely to have direct physically aggressive behaviors than females (e.g., Bettencourt & Miller, 1996; Eagly & Steffen, 1986; Frodi, McCauley, & Thome, 1977; Richardson & Hammock, 2007); however, some studies reported that females were not less aggressive than males when the aggression appears in an indirect form (e.g., Crick & Grotpeter, 1995; Österman, Björkqvist, & Lagerspetz, 1998). Direct aggression indicates behaviors that people use to harm others in direct encounters, including both verbal (e.g., curse the person) and physical (e.g., kick the person) forms. By contrast, indirect aggressive behaviors appear in indirect encounters and also include both verbal (e.g., spread gossips) and physical (e.g., take a belonging from the person) forms. Richardson and Green (1999) found that male-male interaction had the most frequent direct aggression, whereas females had similar levels of indirect aggression as their male counterparts.

Men are more likely than women to use all types of illegal drugs, including marijuana and prescribed drugs. Men have higher rates of use or dependence on illegal drug/alcohol than women for most age groups (SAMHSA, 2014). Wagner, Stempliuk, Zilberman, Barroso, and Guerra de Andrade (2007) examined gender differences and alcohol/drug usage among college students. The results showed that men had more experience with alcohol/drugs than women, whereas women had more prescribed medications than men. Regarding alcohol/drug consumption, college women tended to consume because of dissatisfaction with life, whereas men tended to consume during social interactions.

Psychologists who study the MMPI/MMPI-2/MMPI-2 RF have been interested in gender differences on the MMPI inventories. The majority of the studies tested the gender differences by simply comparing scale means across men and women. Among MMPI scales, Scale Masculinity-Femininity (Mf) showed a large effect in distinguishing gender, followed by Scales Hypochondriasis (Hs), Depression (D), and Hysteria (Hy), which showed small to medium effects across gender (Dahlstrom, Welsh, & Dahlstrom, 1975). In international MMPI research, Butcher and Pancheri (1976) also found that Scale Mf had the largest gender mean difference in each of the seven nations (Costa Rica, Israel, Italy, Japan, Mexico, Pakistan, and Switzerland). Compared to other nations, Costa Rica and Japan had larger gender mean differences on other scales, with Costa Rican women scoring higher than men on Scale Hs, Psychasthenia (Pt), and Schizophrenia (Sc) whereas Japanese men scored higher than women on Scale Hs, Hy, and Paranoia (Pa). Similar results were found in the international research of MMPI-2, with Scale Mf showing the greatest gender mean difference among American, Hong Kong, Japanese, Korean, and Mexican college students (Butcher, 1996). Substantial gender mean differences were also found for the Fears scale (FRS), Scale D, and Scale Hs. The FRS showed greatest gender mean differences for all cultural samples; the Scale D and Hs showed medium to large effect sizes across gender for American and Korean college samples.

However, there is a potential underlying problem when comparing scale means across genders. Mean comparisons across genders are based on an assumption that items are functioning the same for both genders; however, there are a lack of studies that support this assumption. On the contrary, two studies showed evidence of gender differences on the MMPI-A, MMPI-2, and MMPI-2 RF at the item level (Han, Park, Weed, Lim, Johnson, and Joles, 2013; Wang et. al., 2015). Han et al. (2013) found gender differences on the MMPI-A and MMPI-2 for American and Korean normative adolescent and adult samples. Wang et al. (2015) replicated Han et al.'s (2013) methodology and found gender differences on the MMPI-2 RF for American and Korean normative and clinical adult samples. The proportion of items with gender differences was higher in the American samples than in the Korean samples. Furthermore, the American normative sample had a higher proportion of items with gender differences than the American clinical sample, whereas the Korean normative and clinical samples had the same proportion of items. Across cultures and ages, items with gender differences reflected stereotyped gender behaviors, emotions, and interests.

Rational of Current Study and Hypotheses

Recent translations of the MMPI-2/MMPI-2 RF have drawn researchers' interests to personality trait differences across cultures. However, before examining personality constructs cross-culturally by simply comparing scale means across cultures, one should test the measurement invariance of scales to make sure that each construct is comparable across cultural groups and that each scale mean difference across groups truly reflects the group mean difference. The current study tests the measurement invariance of the MMPI-2 Externalizing Specific Problems (SP) Scales across American and Korean normative and clinical samples.

As previous studies have found a gender difference on the MMPI-2/MMPI-2 RF at the item level (Han et al., 2013; Wang et al., 2015), examining the measurement invariance of the MMPI-2 Externalizing SP Scales across cultures by simply combining both gender groups may overlook the possibility of the measurement noninvariance of the Scales across genders. For the present study, I first will test the measurement invariance of the Externalizing SP Scales across genders for each cultural sample using MGCFA to answer the question of whether the Scales function the same across genders. If the Scales fail to meet the measurement invariance, noninvariant items will be detected. Then, I will further examine the measurement invariance of the Scales across cultures using a hybrid multi-group–MIMIC model by incorporating gender noninvariant items (Figure 1; Marsh, Tracey, & Craven, 2006; Sideridis, Tsaousis, & Al-harbi, 2015). The basic form of the model is a multi-group CFA model across American and Korean sample; the MIMIC structure helps to incorporate gender noninvariant items when combining both gender groups together for each culture sample.

The hypotheses are as follows:

Given that the intention of creating SP Scales is to measure specific clinical constructs that are associated with one of the RC Scales (the JCP and SUB Scales measure two specific facets of RC4 Scale; the AGG and ACT Scales measure two facets of RC9 Scale; Ben-Porath & Tellegen, 2008), it is expected that the single factor model will produce a good fit in all samples.

Hypothesis 1: SP items will be well represented by a one-factor model for all eight samples (American normative male and female samples, American clinical male and female sample, Korean normative male and female samples, and Korean clinical male and female samples).

In a recent review of cross-cultural measurement invariance studies, Chen (2008) found that majority of the studies have reached configural invariance, and 74% of the studies showed a higher average factor loading in its source culture than in its imported culture. Since MMPI-2 RF is an inventory that was originally developed in the United States for measuring personality and psychopathology, it is expected that the average loading will be higher in the American samples than in the Korean samples, and higher in the clinical samples than in the normative samples.

Hypothesis 2: The average factor loading will be higher in the American samples than in the Korean samples, and higher in the clinical samples than in the normative samples.

As existing studies show that men are higher than women in externalizing disorders (e.g., substance abuse and antisocial disorders) (Rosenfiled, 1999), it is expected that there are gender noninvariant items in the Externalizing SP Scales. The lifetime prevalence rates of conduct disorder are reportedly more common among males than females (Maughan, Rowe, Messer, Goodman, & Metzer, 2004). For Juvenile Conduct Problems (JCP) Scale, all items except item 96 are expected to be gender noninvariant since all other items relate to behavior/school problems. In general, men have higher rates of use or dependence on illegal drug/alcohol than women (SAMHSA, 2014). For Substance Abuse (SUB) Scale, all items except item 237 are related to drug/alcohol problems and are expected to be gender noninvariant items. Males are more likely to have direct physical aggressive behaviors than females; however, females tend to be as aggressive as males, but in indirect form (e.g., spread gossips) (Crick & Grotpeter, 1995). For Aggression (AGG) Scale, item 23, 84, 312, 316, and 329 are related to physically aggressive behaviors and are expected to be gender noninvariant items. For the Activation (ACT) Scale, all items are related to emotional stability control and are expected to be gender noninvariant items. Given that men are more likely to have externalizing problems than women in general, it is expected that men will score higher in Externalizing SP Scale constructs than women in all four samples.

Hypothesis 3: The Externalizing SP Scales will have gender noninvariant items, and men will have higher scores on Externalizing SP Scale constructs than women in all four samples

In a cross-cultural studies review, Chen (2008) found that around 93% of the studies had achieved configural invariance, but 84% of the studies had at least 50% of the factor loadings higher in the source culture than in the imported culture, indicating that most studies had failed to reach to a more rigid form of measurement invariance after configural invariance. With binary data characteristics of MMPI-2 RF, factor loading and threshold are to be constrained at once so metric and invariance are tested at one step—scalar model (Muthén & Muthén, 2012). Two existing studies examined the measurement invariance of the MMPI-2 RC Scales. Partial scalar invariance was reached for RCd and RC8 in one-factor model across American and Korean normative samples (Ketterer, 2010). Further, partial scalar invariance was established for RC4 rationally-derived four-factor model (RC4 four-factor model; Han et al., 2011) across American and Korean clinical samples. Therefore, it is expected in the current study that partial scalar invariance will be achieved for the Externalizing SP Scales across American and Korean normative and clinical samples.

Hypothesis 4: Partial scalar invariance will be expected for the Externalizing SP Scales.

CHAPTER II

METHOD

Participants

American Normative Sample

The MMPI-2-RF American normative sample was adapted from the MMPI-2 American normative sample. It was created by combining 1,138 men from the MMPI-2 normative sample with 1,138 women selected randomly from the 1,462 women in the MMPI-2 normative sample (Ben-Porath & Tellegen, 2008). The MMPI-2-RF American normative sample (N = 2,276) included 1,138 men (M_{age} = 41.71, SD = 15.32) and 1,138 women (M_{age} = 40.44, SD = 15.25). The self-reported racial composition of the sample was 82% Caucasian American, 12% African American, and 6% "other." The MMPI-2 American normative sample was originally selected to be consistent with the 1980 Census in demographics (age and ethnicity) using proportional stratified sampling (Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989). The initial sample included 2,900 protocols, but 300 were excluded because of incomplete accompanying forms or MMPI-2 invalidity (F > 20, Back F > 20, or Cannot Say > 40 [out of 704 MMPI-2 experimental version items]), resulting in a final sample of 2,600 (1,138 men and 1,462 women).

Korean Normative Sample

The MMPI-2-RF Korean normative sample was created from the MMPI-2 Korean normative sample by combining 651 women from the MMPI-2 normative sample with 651 men selected randomly from the 701 men in the MMPI-2 normative sample (Han, Moon, Lee, & Kim, 2011). The MMPI-2-RF Korean normative sample (N = 1,302) included 651 men (M_{age} = 38.77, SD = 13.18) and 651 women (M_{age} = 40.26, SD = 13.73). The MMPI-2 Korean normative sample was selected to match the 2000 Census in demographics (geographical region, place of residence, age, and education ethnicity) using proportional stratified sampling (Kim, Han, Lim, Lee, Min, & Moon, 2005). The initial sample included 1,587 protocols, but 213 were eliminated due to incomplete accompanying forms or Korean MMPI-2 invalidity (Cannot Say > 9, VRIN > 17, TRIN <4 or TRIN > 15, KI > 21). Another 22 were deleted randomly to ensure sample representativeness, and then the final sample was formed (N = 1,352; 701 men and 651 women).

American Clinical Sample

The American clinical sample (N = 1,980) consisted of (a) an inpatient sample from Hennepin County Medical Center (HCMC) and a Veterans Administration (VA) Medical Center (Arbisi, Sellbom, & Ben-Porath, 2008) and (b) an outpatient sample of clients who were seeking services at a large urban Community Mental Health Center (CMHC) in northeast Ohio (Graham, Ben-Porath, & McNulty, 1999). The inpatient sample (N = 1,499; 76%) included 1,136 men (M_{age} = 43.02, SD = 14.30) and 363 women (M_{age} = 35.13, SD = 12.08). The outpatient sample (N = 481; 24%) included 179 men (M_{age} = 35.61, SD = 10.74) and 302 women (M_{age} = 33.37, SD = 9.81).

The initial sample was composed of 2,378 inpatient protocols and 1,219 outpatient protocols, but 161 outpatient cases were eliminated due to MMPI-2 invalidity criteria (Cannot Say ≥ 30, VRIN [T-score] ≥ 80, TRIN [T-score] ≥ 80, and Fp [T-score] > 100). In order to achieve correspondence with the Korean clinical sample in terms of primary diagnosis, we selected the American clinical protocols with DSM-III-R (or DSM-IV) primary diagnosis of one of the five Axis I disorders (Schizophrenia, Major Depressive Disorder, Bipolar Disorder, Anxiety Disorder, or Somatoform Disorder) to constitute the final sample (N = 1,983). Three cases were further deleted from the selected protocols due to missing values on the age variable.

The diagnostic breakdown of the finalized American clinical sample (N = 1,980) was as follows: Schizophrenia (n = 194; 10%), Depressive Disorder (n = 1,220; 62%), Bipolar Disorder (n = 246; 12%), Anxiety Disorder (n = 301; 15%), or Somatoform Disorder (n = 19; 1%).

The self-reported racial composition of the sample was 83% Caucasian American, 12% African American, and 5% "other." Most of the participants were single (40%) while 28% were divorced and 21% were married. Approximately 51% of individuals in the sample were unemployed; 26% were either self-employed or part-time employed, and 5% were retired.

Korean Clinical Sample

The Korean clinical sample (Han, Moon, Lee, & Kim, 2011a, 2011b) was composed of 395 patients utilizing inpatient and outpatient facilities at Samsung National Hospital in Seoul, Korea. The inpatient sample (N = 158; 40%) consisted of 64 men (M_{age} = 36.02, SD = 14.99) and 94 women (M_{age} = 40.07, SD = 15.39). The outpatient sample (N = 237; 60%) consisted of 112 men (M_{age} = 34.86, SD = 15.06) and 125 women (M_{age} = 41.19, SD = 14.97).

The initial sample included 400 protocols, but 5 cases were deleted based on MMPI-2 invalidity criteria (Cannot Say ≥ 30, VRIN [T-score] ≥ 80, TRIN [T-score] ≥ 80, and Fp [T-score] > 100). All patients were diagnosed using DSM-IV criteria except five who were diagnosed using the Mini International Neuropsychiatric Interview (MINI). Patients had a primary diagnosis of Schizophrenia (n = 71; 18%), Depressive Disorder (n = 85; 21%), Bipolar Disorder (n =89; 23%), Anxiety Disorder (n = 86; 22%), or Somatoform Disorder (n = 64; 16%).

Most of the participants were either single (49%) or married (44%). Approximately 25% of individuals in the sample were housewives, 23% were self-employed or part-time employed, 22% were students, and 22% were either unemployed, on leave, or retired.

Measures

MMPI-2-RF Externalizing Scales

The MMPI-2-RF (Tellegen & Ben-Porath, 2008) is a self-report instrument that assesses personality and psychopathology using 338 true/false items. There are four Externalizing Scales: Juvenile Conduct Problems (JCP) Scale, Substance Abuse (SUB) Scale, Aggression (AGG), and Activation (ACT). The JCP Scale includes six items (all keyed *true*) and its internal consistency reliability coefficients varies across gender and samples (α = .65 for men and α = .56 for women in American normative sample; α = .74 for men and α = .69 for women in American clinical sample). The SUB Scale includes seven items (six keyed *true* and one keyed *false*) and its internal consistency reliability coefficients varies across gender and samples (α = .62 for men and α = .62 for women in American normative sample; α = .74 for men and α = .71 for women in American clinical sample). The AGG Scale includes nine items (nine keyed *true*) and its internal consistency reliability coefficients varies across gender and samples (α = .66 for men and α = .58 for women in American normative sample; α = .77 for men and α = .73 for women in American clinical sample). The ACT Scale includes eight items (eight keyed *true*) and its internal consistency reliability coefficients varies across gender and samples (α = .60 for men and α = .60 for women in American normative sample; α = .70 for men and α = .70 for women in American clinical sample).

Korean MMPI-2-RF Externalizing Scales

The Korean MMPI-2-RF was published in 2011 (Han, Moon, Lee, & Kim, 2011). The internal consistency reliability coefficients of four Externalizing Scales are listed, by gender and samples, as follows: the JCP Scale, α = .51 for men and α = .52 for women in Korean normative sample; α = .58 for men and α = .51 for women in Korean clinical sample; the SUB Scale, α = .54 for men and α = .53 for women in Korean normative sample; α = .71 for men and α = .62 for women in Korean clinical sample; the AGG Scale, α = .59 for men and α = .60 for women in Korean normative sample; α = .73 for men and α = .71 for women in Korean clinical sample; the ACT Scale, α = .53 for men and α = .56 for women in Korean normative sample; α = .67 for men and α = .69 for women in Korean clinical sample.

Analytic Procedure

For all the MMPI-2-RF Externalizing Scales, a "true" response was coded "1," and a "false" response was coded "0." The only falsely keyed item in SUB Scale was reverse-coded to ensure point accrual in the direction of greater psychopathology. All analyses were conducted in Mplus 7.2 using weighted least square mean variance (WLSMV) estimation and Theta parameterization (Muthén & Muthén, 1998–2012).

First, a series of one-sample CFA (a total of 32 CFAs; gender [2] x within-cultural sample [2] x culture [2] x scales [4]) was performed to test the model fit of each of the four Externalizing SP scales (JCP, SUB, AGG, and ACT) for each gender group (American normative male and female, American clinical male and female, Korean normative male and female, and Korean clinical male and female samples). Second, a series of MGCFA (a total of four MGCFAs) was conducted to examine the measurement invariance of each of the four Externalizing SP scales across gender (American normative male vs. American normative female, American clinical male vs. American clinical female, Korean normative male vs. Korean normative female, Korean clinical male vs. Korean clinical female).

In the MGCFA, two nested models were tested accordingly: configural model (model 1) and scalar model (model 2) (Vandenberg & Lance, 2000). In the configural model (model 1), both factor loadings and thresholds are freely estimated, and the datasets of both gender groups were examined simultaneously to test whether the pattern of factor loadings was similar across groups; a good model fit would indicate that configural invariance is reached. Due to the binary nature of the data, Mplus places constraints on factor loadings and thresholds simultaneously (Muthén & Muthén, 1998-2012, p. 485), so the analysis of metric invariance is not permitted before the analysis of scalar invariance. The scalar model (model 2), in which both factor loadings and thresholds are constrained equally across gender groups, was compared to the configural model (model 1) in model fit; a nonsignificant model fit difference would indicate full scalar invariance. If the scale shows lack of full scalar invariance, noninvariant items are further examined. In the partial scalar model (e.g., model 2a), I successively detected the noninvariant items by constraining the equality constraints on the factor loading and threshold of one item at a time but keeping free on the remaining items. Then, I compared the model to the configural model (model 1) in model fit; a noninvariant item would result in a significant model fit difference. Third, a series of CFA was conducted to test the model fit of the each of the four Externalizing SP scales for each of the four culture groups (American normative, American clinical, Korean normative, and Korean clinical samples). For each of the samples, gender noninvariant items are incorporated into the model by using MIMIC model structure, that is, adding the direct effects of gender variable to the latent factor and gender noninvariant indicators (Figure 1). All the paths from gender variable to indicators are set as 0 whereas the path to the latent factor and gender noninvariant indicators are freely estimated. Fourth, MGCFA was conducted to examine the measurement invariance of the Scales across cultures (American normative vs. Korean normative, American clinical vs. Korean clinical). The procedure of examining measurement invariance and noninvariant items across cultures is the same as testing the measurement invariance across gender above.

The chi-square statistic (χ^2), comparative fit index (CFI; with acceptable values above .90), and root mean square error of approximation (RMSEA; with acceptable values below .08) were used to evaluate the model fit (Kline, 2011). The configural model, scalar model, and partial scalar model were nested; a chi-square difference test, $\Delta\chi^2$, was used to compare nested models. For the binary data with WLSMV estimator, the DIFFTEST was used to acquire the chi-square difference instead of the regular chi-square difference calculation (Muthén & Muthén, 1998-2012, p. 625).

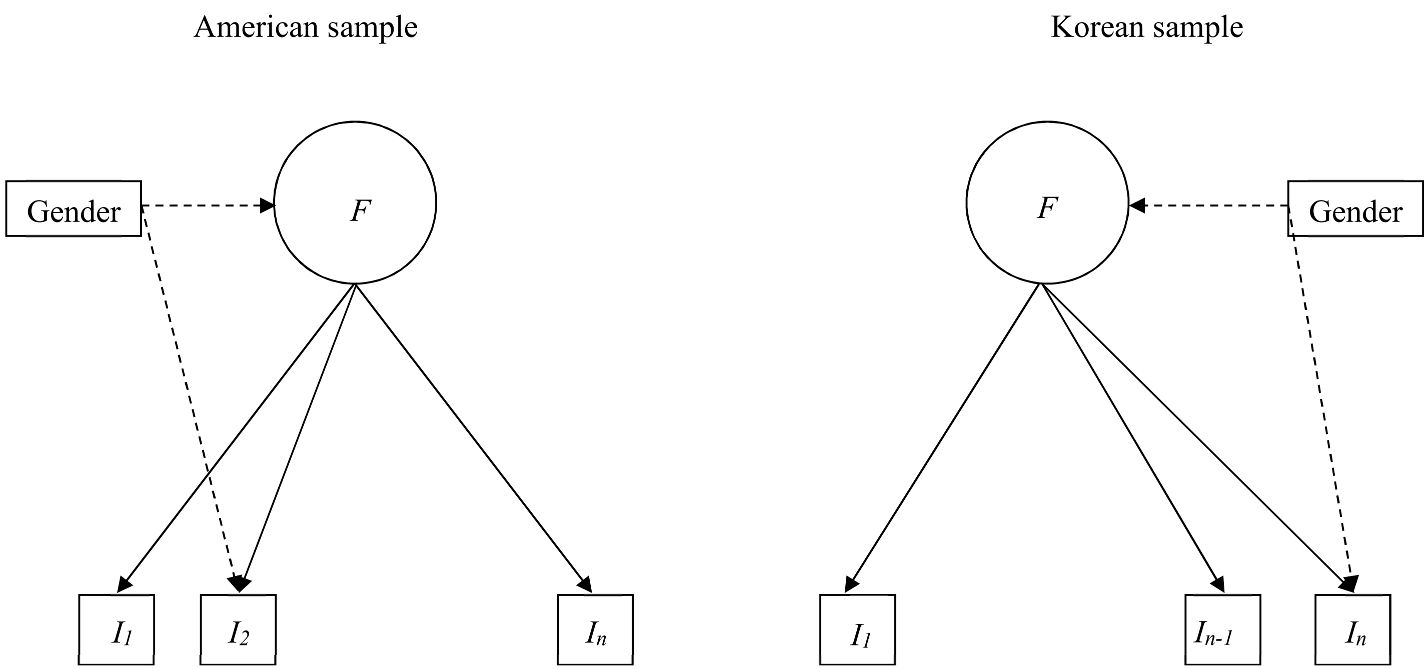

Figure 1. *The Hybrid Multi-group–MIMIC Model Across Cultures (American Sample vs. Korean Sample) Incorporating Gender Noninvariant Items*

Solid lines denote CFA model components; dashed lines denote MIMIC model components. I_n = item *n* of factor *F*. In Figure 1, I_2 is a noninvariant item for American sample, and I_n is a noninvariant item for Korean sample.

CHAPTER III

RESULTS

Descriptive Statistics

Table 1 presented the endorsement percentages in the keyed direction on Externalizing SP items for each gender group. A high value indicated an elevated level of psychopathology. Regarding the gender difference within each sample, the America rmative male sample had an average higher endorsement percentage than the American normative female sample on most of the scales (3 out of 4) except for the ACT scale (see Table 2). This trend also applied to American clinical and Korean clinical samples. For the Korean normative sample, the male sample had an average higher endorsement percentage than the female sample on all four scales, with male a slightly higher than female on the ACT. Across cultures, the American normative male sample had an average higher endorsement percentage than the Korean normative male sample on most of the scales (3 out of 4) except for the ACT, and the same trend also applied to the normative female samples. The American clinical male sample had an average lower endorsement percentage than the Korean clinical male sample on most of the scales (3 out of 4) except for the AGG, and the clinical female samples shared the same trend. Across settings, the American clinical samples had an average higher endorsement percentage than the American normative samples on all four scales for both male and female samples. However, the Korean clinical male sample had an average higher endorsement percentage than its normative counterpart on the JCP and the AGG scales. The Korean clinical female sample had an average higher endorsement percentage on the AGG and the ACT scales.

Table 1. *Percentage Responding to Externalizing SP Scale Items in Keyed Direction*

	American normative			American clinical			Korean normative			Korean clinical		
	M[1]	F[2]	Tot[3]	M[4]	F[5]	Tot[6]	M[7]	F[8]	Tot[9]	M[10]	F[11]	Tot[12]
JCP												
21	58	36	47	67	55	63	35	23	29	45	25	34
66	31	10	20	51	35	46	28	18	23	24	12	18
96	19	19	19	43	53	46	30	23	26	32	27	29
205	16	10	13	38	41	39	31	24	27	34	29	31
223	17	7	12	39	27	35	15	8	11	11	5	8
253	9	3	6	28	17	24	13	8	10	13	6	9
M	25.00	14.17	19.50	44.33	38.00	42.17	25.33	17.33	21.00	26.50	17.33	21.50
SD	17.67	11.92	14.40	13.38	14.79	13.08	9.09	7.53	8.37	13.10	10.93	11.43
α	.65	.56	.63	.74	.69	.73	.51	.52	.53	.58	.51	.57
SUB												
49	34	25	30	39	31	36	8	7	7	4	1	3
86	5	3	4	13	12	13	29	21	25	15	13	14
141	45	24	34	58	40	52	32	17	25	26	11	18
192	14	8	11	20	14	18	29	16	23	13	7	10
237(r)	30	25	27	39	40	40	39	38	38	38	36	37
266	7	3	5	32	21	28	15	12	14	13	7	9
297	20	9	14	22	19	21	26	15	21	18	8	12
M	22.14	13.86	17.86	31.86	25.29	29.71	25.43	18.00	21.86	18.14	11.86	14.71
SD	14.83	10.37	12.32	15.16	11.74	13.77	10.53	9.83	9.70	10.95	11.29	10.86
α	.62	.62	.63	.74	.71	.73	.54	.53	.55	.71	.62	.68
AGG												
23	39	39	39	52	55	53	38	36	37	49	43	46
26	27	13	20	36	26	33	40	36	38	51	37	43
41	5	5	5	9	9	9	16	20	18	14	18	17
84	16	8	12	31	27	30	52	36	44	50	37	43
231	8	5	6	14	10	12	20	17	19	13	16	14
312	8	4	6	27	17	24	14	8	11	9	12	11
316	51	39	45	57	46	53	63	58	60	67	62	64
329	18	8	13	31	21	28	17	13	15	17	12	14

Table 1. *Percentage Responding to Externalizing SP Scale It[illegible] in Keyed Direction (Continued)*

	American normative			American clinical			Korean normative			Korean clinical		
	M[1]	F[2]	Tot[3]	M[4]	F[5]	Tot[6]	M[7]	F[8]	Tot[9]	M[10]	F[11]	Tot[12]
337	43	57	50	62	70	65	50	54	52	60	70	66
M	23.89	19.78	21.78	35.44	31.22	34.11	34.44	30.89	32.67	36.67	34.11	35.33
SD	16.94	19.79	17.98	18.42	21.16	19.19	18.30	17.69	17.61	22.98	21.58	21.89
α	.66	.58	.63	.77	.73	.76	.59	.60	.59	.73	.71	.72
ACT												
72	81	82	82	77	78	77	48	44	46	57	49	53
81	6	19	12	34	51	40	28	40	34	40	53	47
166	37	41	39	32	30	31	30	27	28	17	17	17
181	35	36	35	38	34	37	22	15	18	26	31	29
207	64	70	67	48	48	48	63	62	63	41	52	47
219	48	55	52	58	55	57	31	34	32	43	53	48
267	15	11	13	34	28	32	28	19	23	17	20	19
285	19	25	22	31	40	34	19	21	20	21	29	25
M	38.13	42.38	40.25	44.00	45.50	44.50	33.63	32.75	33.00	32.75	38.00	35.63
SD	25.49	24.98	25.35	16.26	16.44	15.81	14.65	15.64	15.07	14.61	15.41	14.61
α	.60	.60	.60	.70	.70	.70	.53	.56	.54	.67	.69	.68

Notes: [1]*Ns* ranged from 1123 (α for AGG) to 1131 (α for JCP and for SUB) using listwise deletion for reliability coefficients. [2]*Ns* ranged from 1124 (α for AGG) to 1137 (α for JCP) using listwise deletion for reliability coefficients. [3]*Ns* ranged from 2247 (α for AGG) to 2268 (α for JCP) using listwise deletion for reliability coefficients. [4]*Ns* ranged from 660 (α for AGG) to 663 (α for JCP) using listwise deletion for reliability coefficients. [5]*Ns* ranged from 1305 (α for AGG) to 1315 (α for JCP) using listwise deletion for reliability coefficients. [6]*Ns* ranged from 1965 (α for AGG) to 1978 (α for JCP) using listwise deletion for reliability coefficients. [7]*N* = 651. [8]*N* = 651. [9]*N* = 1302. [10]*N* = 176. [11]*N* = 219. [12]*N* = 359. Item numbers correspond to the MMPI-2-RF item booklet. (r) indicates that the item was reverse keyed, a “false” endorsement indicative of greater psychopathology. These items were recoded to be keyed the same direction.

Table *2. Gender and Cultural Differences on Average Percentage Responding to Externalizing SP Scale Items in Keyed Direction*

			JCP	SUB	AGG	ACT
Gender difference	America	Normative	10.83	8.28	4.11	-4.25
		Clinical	6.33	6.57	4.22	-1.50
	Korea	Normative	8.00	7.43	3.55	.88
		Clinical	9.17	6.28	2.56	-5.25
Cultural difference	Men	Normative	-.33	-3.29	-10.55	4.50
		Clinical	17.83	13.72	-1.23	11.25
	Women	Normative	-3.16	-4.14	-11.11	9.63
		Clinical	20.67	13.43	-2.89	7.50
Setting difference	America	Men	19.33	9.72	11.55	5.87
		Women	23.83	11.43	11.44	3.12
	Korea	Men	1.17	-7.29	2.23	-.88
		Women	0	-6.14	3.22	5.25

Note: A negative value indicates a higher mean percentage for women than for men, for the Korean sample than for the American sample, or for the normative sample than the clinical sample.

Confirmatory Factor Analysis

Table 3 presented the model fit of the one-factor model. The one-factor model had a good model fit based on CFI and RMSEA for all samples, except for the Korean normative male (*CFI* = .798) and female (*CFI* = .806) samples on the AGG and the American normative male sample on the SUB (*CFI* = .886, *RMSEA* = .097). Table 4 further presented the standard factor loadings and thresholds of the items.

For the JCP scale, the American normative samples (λ_{mean} = .68 for male; λ_{mean} = .67 for female) had average higher standardized factor loadings than the Korean normative samples (λ_{mean} = .53 for male; λ_{mean} = .56 for female); the American clinical samples (λ_{mean} = .72 for male; λ_{mean} = .68 for female) had average higher factor loadings than the Korean clinical samples (λ_{mean} = .62 for male; λ_{mean} = .60 for female); the American clinical samples had slightly higher average factor

Table 3. *Confirmatory Factor Analysis of Externalizing SP Scales Across Genders*

Model	χ^2	*df*	*p*	CFI	RMSEA [90% CI]
JCP					
American normative male[1]	32.838	9	<.001	.980	.048 [.031, .066]
American normative female[2]	12.223	9	.201	.995	.018 [.000, .040]
Korean normative male[3]	18.559	9	.029	.965	.040 [.012, .067]
Korean normative female[4]	11.589	9	.238	.989	.021 [.000, .052]
American clinical male[5]	49.050	9	<.001	.990	.058 [.043, .075]
American clinical female[6]	17.351	9	.044	.996	.037 [.006, .064]
Korean clinical male[7]	6.430	9	.696	1.000	.000 [.000, .066]
Korean clinical female[8]	5.064	9	.829	1.000	.000 [.000, .046]
SUB					
American normative male[1]	162.638	14	<.000	.886	.097 [.084, .110]
American normative female[2]	116.799	14	<.000	.902	.080 [.067, .094]
Korean normative male[3]	55.082	14	<.000	.917	.067 [.049, .086]
Korean normative female[4]	31.902	14	.004	.946	.044 [.024, .065]
American clinical male[5]	251.957	14	<.001	.929	.114 [.102, .126]
American clinical female[6]	81.360	14	<.001	.953	.085 [.068, .103]
Korean clinical male[7]	14.640	14	.403	.999	.016 [.000, .076]
Korean clinical female[8]	24.201	14	.043	.972	.058 [.010, .095]
AGG					
American normative male[1]	74.511	27	<.001	.967	.039 [.029, .050]
American normative female[2]	103.214	27	<.001	.913	.050 [.040, .060]
Korean normative male[3]	138.820	27	<.001	.798	.080 [.067, .093]
Korean normative female[4]	136.671	27	<.001	.806	.079 [.066, .092]
American clinical male[5]	137.475	27	<.001	.972	.056 [.047, .065]
American clinical female[6]	50.652	27	.004	.983	.036 [.020, .051]
Korean clinical male[7]	49.794	27	.005	.952	.069 [.038, .099]
Korean clinical female[8]	37.325	27	.089	.975	.042 [.000, .072]
ACT					
American normative male[1]	57.361	20	<.001	.962	.041 [.028, .053]

Table 3. *Confirmatory Factor Analysis of Externalizing SP Scales Across Gender (Continued)*

Model	χ^2	*df*	*p*	CFI	RMSEA [90% CI]
American normative female[2]	95.404	20	<.001	.919	.058 [.046, .069]
Korean normative male[3]	39.993	20	.005	.931	.039 [.021, .057]
Korean normative female[4]	48.557	20	<.001	.923	.047 [.030, .064]
American clinical male[5]	138.370	20	<.001	.948	.067 [.057, .078]
American clinical female[6]	123.064	20	<.001	.908	.088 [.073, .103]
Korean clinical male[7]	24.444	20	.224	.978	.036 [.000, .078]
Korean clinical female[8]	31.139	20	.053	.963	.050 [.000, .083]

Note. [1]*N* = 1,138. [2]*N* = 1,138. [3]*N* = 651. [4]*N* = 651. [5]*N* =1,315. [6]*N* = 665. [7]*N* = 176. [8]*N* = 219.
CFI = comparative fit index. TLI = Tucker-Lewis index. RMSEA = root mean square error of approximation.
CI = confidence interval.

Table 4. *Standardized Factor Loadings and Thresholds of Externalizing SP Scale Items*

	American normative[1]				Korean normative[2]				American clinical[3]				Korean clinical[4]			
	Male		Female		Male		Female		Male		Female		Male		Female	
	FL	TH	FL	TH	FL	TH	FL	TH	FL	TH	FL	TH	FL	TH	FL	TH
JCP																
21	.46	-.20	.47	.36	.31	.39	.41	.74	.54	-.44	.43	-.11	.28	.13	.40	.67
66	.84	.50	.86	1.30	.81	.57	.69	.93	.94	-.03	.99	.38	1.02	.69	.90	1.18
96	.60	.87	.53	.90	.45	.54	.67	.74	.54	.18	.46	-.08	.49	.46	.45	.62
205	.65	1.00	.70	1.27	.43	.50	.58	.72	.66	.30	.56	.22	.45	.41	.55	.56
223	.80	.94	.74	1.45	.68	1.05	.55	1.42	.90	.29	.89	.63	.72	1.21	.56	1.69
253	.70	1.37	.70	1.88	.49	1.13	.45	1.43	.71	.59	.74	.96	.74	1.15	.72	1.52
M	.68	.75	.67	1.19	.53	.70	.56	1.00	.72	.15	.68	.33	.62	.68	.60	1.04
SD	.14	.54	.14	.52	.18	.31	.11	.34	.17	.35	.23	.42	.26	.43	.18	.49
SUB																
49	.73	.41	.72	.67	.41	1.43	.59	1.51	.61	.28	.52	.49	.23	1.75	.27	2.21
86	.34	1.68	.44	1.90	.36	.56	.36	.79	.47	1.13	.43	1.20	.62	1.02	.41	1.12
141	.53	.14	.67	.71	.63	.46	.67	.94	.77	-.20	.73	.25	.90	.66	.88	1.23
192	.48	1.07	.59	1.39	.65	.56	.65	.98	.80	.85	.85	1.08	.92	1.12	.81	1.45
237	.72	.53	.70	.67	.15	.28	.27	.32	.83	.48	.85	.80	.16	.30	.44	.37
266	.67	1.50	.61	1.85	.63	1.03	.66	1.17	.85	.76	.87	.86	.91	1.15	.80	1.49
297	.86	.86	.87	1.33	.85	.64	.72	1.05	.62	.28	.58	.25	.96	.91	1.01	1.42
M	.62	.88	.66	1.22	.53	.71	.56	.97	.71	.51	.69	.70	.67	.99	.66	1.33
SD	1.77	.57	.13	.54	.23	.39	.17	.36	.14	.44	.18	.38	.34	.45	.28	.55
AGG																
23	.64	.27	.60	.29	.64	.31	.64	.37	.76	-.06	.70	-.13	.91	.01	.87	.17
231	.60	1.43	.67	1.65	.49	.83	.56	.95	.71	1.10	.68	1.29	.72	1.15	.53	1.01
312	.60	1.44	.55	1.80	.29	1.07	.33	1.38	.61	.61	.67	.94	.57	1.34	.50	1.16

Table 4. *Standardized Factor Loadings and Thresholds of Externalizing SP Scale Items (Continued)*

	American normative[1]				Korean normative[2]				American clinical[3]				Korean clinical[4]			
	Male		Female		Male		Female		Male		Female		Male		Female	
	FL	TH	FL	TH	FL	TH	FL	TH	FL	TH	FL	TH	FL	TH	FL	TH
316	.50	-.03	.47	.28	.51	-.32	.37	-.20	.62	-.16	.58	.10	.60	-.43	.45	-.30
329	.79	.91	.67	1.44	.48	.95	.49	1.15	.72	.48	.72	.81	.61	.95	.67	1.18
337	.65	.17	.63	-.18	.63	.01	.53	-.11	.81	-.29	.74	-.53	.76	-.26	.74	.53
M	.60	.83	.59	1.05	.49	.45	.51	.57	.69	.43	.67	.57	.67	.42	.63	.59
SD	.11	.61	.12	.73	.12	.52	.11	.55	.10	.55	.08	.64	.14	.69	.14	.51
ACT																
72	.41	-.87	.44	-.93	.37	.04	.52	.16	.45	-.72	.46	-.76	.63	-.19	.66	.02
81	.48	1.59	.40	.89	.41	.59	.50	.26	.44	.41	.43	-.04	.59	.25	.71	-.06
166	.71	.33	.69	.23	.48	.54	.48	.61	.79	.47	.70	.52	.65	.95	.47	.96
181	.63	.40	.59	.36	.60	.79	.61	1.03	.63	.30	.68	.42	.51	.64	.74	.51
207	.77	-.36	.67	-.52	.45	-.33	.31	-.31	.76	.06	.71	.04	.67	.23	.50	-.05
219	.47	.04	.47	-.12	.46	.50	.46	.42	.61	-.20	.70	-.12	.60	.19	.64	-.07
267	.41	1.02	.53	1.25	.46	.58	.58	.89	.63	.43	.66	.60	.44	.98	.62	.84
285	.45	.88	.54	.68	.54	.87	.56	.81	.58	.49	.55	.26	.73	.81	.52	.56
M	.54	.38	.54	.23	.47	.45	.50	.48	.61	.16	.61	.12	.60	.48	.61	.34
SD	.14	.79	.10	.73	.07	.40	.09	.44	.13	.43	.11	.44	.09	.42	.10	.43

Notes. [1]*N* = 2,276. [2]*N* = 1,302. [3]*N* = 1,980. [4]*N* = 395. Item numbers correspond to the MMPI-2-RF item booklet. (r) indicates that the item was reverse keyed, a "false" endorsement indicative of greater psychopathology. FL = Factor loading. TH = Threshold.

loadings than the American normative samples; the Korean clinical samples had average higher factor loadings than the Korean normative samples.

For the SUB scale, the American normative samples (λ_{mean} = .62 for male; λ_{mean} = .66 for female) had average higher standardized factor loadings than the Korean normative samples (λ_{mean} = .53 for male; λ_{mean} = .56 for female); the American clinical samples (λ_{mean} = .71 for male; λ_{mean} = .69 for female) had slightly higher average factor loadings than the Korean clinical samples (λ_{mean} = .67 for male; λ_{mean} = .66 for female); the American clinical samples had average higher factor loadings than the American normative samples; the Korean clinical samples had average higher factor loadings than the Korean normative samples. When confirmatory factor analyses were conducted combining gender samples together for each culture, the one-factor model showed a good model fit for all culture samples except for the Korean normative sample on the AGG scale (CFI = .808) (Table 5).

For the AGG scale, the American normative samples (λ_{mean} = .60 for male; λ_{mean} = .59 for female) had average higher standardized factor loadings than the Korean normative samples (λ_{mean} = .49 for male; λ_{mean} = .51 for female); the American clinical samples (λ_{mean} = .69 for male; λ_{mean} = .67 for female) had average slightly higher factor loadings than the Korean clinical samples (λ_{mean} = .67 for male; λ_{mean} = .63 for female); the American clinical samples had average higher factor loadings than the American normative samples; the Korean clinical samples had average higher factor loadings than the Korean normative samples.

For the ACT scale, the American normative samples (λ_{mean} = .54 for male; λ_{mean} = .54 for female) had average higher standardized factor loadings than the Korean normative samples (λ_{mean} = .47 for male; λ_{mean} = .50 for female); the American clinical samples (λ_{mean} = .61 for male; λ_{mean} = .61 for female) had similar average factor loadings as the Korean clinical samples (λ_{mean}

= .60 for male; λ_{mean} = .61 for female); the American clinical samples had average higher factor loadings than the American normative samples; the Korean clinical samples had average higher factor loadings than the Korean normative samples.

Table 5. *Confirmatory Factor Analysis of Externalizing SP Scales*

Model	χ^2	*df*	*p*	CFI	RMSEA [90% CI]
JCP					
America normative[1]	39.978	9	<.001	.986	.039 [.027, .052]
Korean normative[2]	18.306	9	.032	.983	.028 [.008, .047]
American clinical[3]	58.801	9	<.001	.992	.053 [.040, .066]
Korean clnical[4]	7.764	9	.558	1.000	.000 [.000, .051]
SUB					
America normative[1]	276.874	14	<.001	.897	.091 [.082, .100]
Korean normative[2]	69.836	14	<.001	.939	.055 [.043, .069]
American clinical[3]	326.310	14	<.001	.936	.106 [.096, .116]
Korean clnical[4]	19.748	14	.138	.995	.032 [.000, .063]
AGG					
America normative[1]	174.307	27	<.001	.939	.049 [.042, .056]
Korean normative[2]	241.937	27	<.001	.808	.078 [.069, .087]
American clinical[3]	173.892	27	<.001	.972	.052 [.045, .060]
Korean clnical[4]	52.662	27	.002	.970	.049 [.029, .069]
ACT					
America normative[1]	134.804	20	<.001	.940	.050 [.042, .058]
Korean normative[2]	65.590	20	<.001	.929	.042 [.031, .053]
American clinical[3]	229.640	20	<.001	.937	.073 [.064, .081]
Korean clnical[4]	35.365	20	.018	.969	.044 [.018, .067]

Note. [1]*N* = 2,276. [2]*N* = 1,302. [3]*N* = 1,980. [4]*N* = 395. CFI = comparative fit index. TLI = Tucker-Lewis index. RMSEA = root mean square error of approximation. CI = confidence interval.

Measurement Invariance Tests

Measurement invariance tests of the one-factor model were conducted across genders (Table 6). First, the configural model was tested, and the model fit was acceptable for all samples, except that the model fit of the AGG scale was not good for the Korean normative sample. Next, equally constraining both factor loadings and thresholds across the two gender samples resulted in a nonsignificant model fit difference between the scalar model and the configural model for the Korean clinical samples on the ACT scale, the Korean clinical samples on the AGG, the Korean normative and clinical samples on the JCP, and the Korean normative and clinical samples on the SUB; the results support full scalar invariance. By contrast, other samples showed a significant model fit difference between the scalar model and the configural model for the scales, which did not support full scalar invariance. As these samples which lacked of full scalar invariance for the scales, noninvariant items were further detected. The partial scalar model, equally constraining the factor loading and threshold of one item at a time while freeing the remaining items, was compared to the configural model in model fit; models with significant model fit differences were identified as noninvariant items.

For the JCP scale, items 21, 66, and 233 were identified as noninvariant items for the American normative sample; items 21, 66, 96, 223, and 253 were identified as noninvariant items for the American clinical sample; $ps < .001$. The following noninvariant items were identified for the SUB scale: items 141 and 192 for the American normative sample, and item 141 for the American clinical sample; $ps < .001$. The following noninvariant items were identified for the AGG scale: item 26, 84, 316, 329, and 337 for the American normative sample; item 84 for the Korean normative sample; and items 312, 316, and 329 for the American clinical sample; $ps < .001$. For the ACT Scale, item 81 was identified as a noninvariant item for the American normative sample; items 81 and 267 for the Korean normative sample; and item 81 for American clinical sample; $ps < .001$.

Table 6. *Measurement Invariance Tests Across Genders*

Model	χ^2	*df*	*p*	CFI	RMSEA[90% CI]	Contrast	$\Delta\chi^2$	Δdf	*p*
JCP									
America normative[1]									
Configural model[5]	45.184	18	<.001	.985	.036 [.023, .050]				
Scalar model[6]	123.083	28	<.001	.948	.055 [.045, .065]	6 vs. 5	68.000	10	<.001
21	199.513	20	<.001	.901	.089 [.078, .100]	21 vs. 5	113.668	2	<.001
66	117.870	20	<.001	.946	.066 [.054, .077]	66 vs. 5	50.077	2	<.001
96	41.848	20	.003	.988	.031 [.018, .044]	96 vs. 5	1.387	2	.500
205	58.319	20	<.001	.979	.041 [.029, .053]	205 vs. 5	11.010	2	.004
223	67.725	20	<.001	.974	.046 [.034, .058]	223 vs. 5	16.732	2	<.001
253	62.842	20	<.001	.976	.043 [.032, .056]	253 vs. 5	13.367	2	.001
Korean normative[2]									
Configural model[5]	30.600	18	.032	.975	.033 [.010, .052]				
Scalar imodel[6]	48.300	28	.010	.959	.033 [.016, .049]	6 vs. 5	18.146	10	.053
American clinical[3]									
Configural model[5]	66.051	18	<.001	.992	.052 [.039, .066]				
Scalar model[6]	169.764	28	<.001	.976	.071 [.061, .082]	6 vs. 5	90.119	10	<.001
21	118.557	20	<.001	.984	.070 [.059, .083]	21 vs. 5	43.447	2	<.001
66	126.370	20	<.001	.982	.073 [.061, .086]	66 vs. 5	48.787	2	<.001
96	98.482	20	<.001	.987	.063 [.051, .076]	96 vs. 5	27.864	2	<.001
205	73.015	20	<.001	.991	.052 [.039, .065]	205 vs. 5	7.761	2	.021
223	111.061	20	<.001	.985	.068 [.056, .080]	223 vs. 5	39.366	2	<.001
253	108.795	20	<.001	.985	.067 [.055, .080]	253 vs. 5	35.998	2	<.001
Korean clinical[4]									
Configural model[5]	11.968	18	.849	1.000	.000 [.000, .036]				
Scalar model[6]	24.521	28	.654	1.000	.000 [.000, .046]	6 vs. 5	11.531	10	.318
SUB									
America normative[1]									
Configural model[7]	279.313	28	<.001	.893	.089 [.080, .098]				
Scalar model[8]	289.617	40	<.001	.893	.074 [.066, .082]	8 vs. 7	44.881	12	<.001

Table 6. *Measurement Invariance Tests Across Genders (Continued)*

Model	χ^2	*df*	*p*	CFI	RMSEA[90% CI]	Contrast	$\Delta\chi^2$	Δdf	*p*
49	275.643	30	<.001	.895	.085 [.076, .094]	49 vs. 7	13.293	2	.001
86	260.572	30	<.001	.902	.082 [.073, .092]	86 vs. 7	4.750	2	.093
141	401.151	30	<.001	.841	.104 [.095, .113]	141 vs. 7	99.618	2	<.001
192	289.555	30	<.001	.889	.087 [.078, .096]	192 vs. 7	18.469	2	<.001
237	279.124	30	<.001	.894	.085 [.076, .095]	237 vs. 7	4.722	2	.094
266	270.967	30	<.001	.897	.084 [.075, .093]	266 vs. 7	5.548	2	.062
297	272.945	30	<.001	.896	.084 [.075, .094]	297 vs. 7	5.004	2	.082
Korean normative[2]									
Configural model[7]	87.621	28	<.001	.928	.057 [.044, .071]				
Scalar model[8]	100.829	40	<.001	.926	.048 [.037, .060]	8 vs. 7	19.708	12	.073
American clinical[3]									
Configrual model[7]	331.203	28	<.001	.936	.105 [.095, .115]				
Scalar model[8]	296.273	40	<.001	.946	.080 [.072, .089]	8 vs. 7	35.493	12	<.001
49	289.864	30	<.001	.945	.094 [.084, .104]	49 vs. 7	6.343	2	.042
86	277.321	30	<.001	.948	.091 [.082, .101]	86 vs. 7	1.152	2	.562
141	372.709	30	<.001	.928	.107 [.098, .117]	141 vs. 7	40.742	2	<.001
192	306.095	30	<.001	.942	.096 [.087, .106]	192 vs. 7	6.742	2	.034
237	322.206	30	<.001	.938	.099 [.090, .109]	237 vs. 7	11.617	2	.003
266	304.870	30	<.001	.942	.096 [.087, .106]	266 vs. 7	2.335	2	.311
297	314.922	30	<.001	.940	.098 [.088, .108]	297 vs. 7	2.963	2	.227
Korean clnical[4]									
Configural model[7]	40.472	28	.060	.990	.047 [.000, .078]				
Scalar model[8]	51.787	40	.100	.990	.039 [.000, .066]	8 vs. 7	14.445	12	.273
AGG									
America normative[1]									
Configural model[9]	178.583	54	<.001	.946	.045 [.038, .052]				
Scalar model[10]	340.297	70	<.001	.884	.058 [.052, .065]	10 vs. 9	139.269	16	<.001
23	174.361	56	<.001	.949	.043 [.036, .050]	23 vs. 9	.678	2	.713
26	240.238	56	<.001	.921	.054 [.047, .061]	26 vs. 9	43.770	2	<.001

Table 6. *Measurement Invariance Tests Across Genders (Continued)*

Model	χ^2	*df*	*p*	CFI	RMSEA[90% CI]	Contrast	$\Delta\chi^2$	Δdf	*p*
41	173.720	56	<.001	.949	.043 [.036, .050]	41 vs. 9	1.001	2	.606
84	203.959	56	<.001	.936	.048 [.041, .055]	84 vs. 9	21.958	2	<.001
231	181.810	56	<.001	.946	.044 [.037, .052]	231 vs. 9	5.970	2	.051
312	182.441	56	<.001	.946	.045 [.037, .052]	312 vs. 9	6.646	2	.036
316	214.550	56	<.001	.932	.050 [.043, .057]	316 vs. 9	27.523	2	<.001
329	201.750	56	<.001	.937	.048 [.041, .055]	329 vs. 9	20.116	2	<.001
337	227.031	56	<.001	.926	.052 [.045, .059]	337 vs. 9	37.336	2	<.001
Korean normative[2]									
Configural model[9]	275.493	54	<.001	.802	.079 [.070, .089]				
Scalar model[10]	309.592	70	<.001	.786	.073 [.064, .081]	10 vs. 9	50.908	16	<.001
23	272.372	56	<.001	.807	.077 [.068, .086]	23 vs. 9	1.796	2	.408
26	275.025	56	<.001	.805	.078 [.068, .087]	26 vs. 9	3.704	2	.157
41	277.343	56	<.001	.802	.078 [.069, .087]	41 vs. 9	4.961	2	.084
84	312.762	56	<.001	.771	.084 [.075, .093]	84 vs. 9	34.613	2	<.001
231	274.690	56	<.001	.805	.077 [.068, .087]	231 vs. 9	3.435	2	.180
312	278.403	56	<.001	.801	.078 [.069, .087]	312 vs. 9	8.019	2	.018
316	277.553	56	<.001	.802	.078 [.069, .087]	316 vs. 9	5.700	2	.058
329	275.342	56	<.001	.804	.078 [.069, .087]	329 vs. 9	4.448	2	.108
337	276.846	56	<.001	.803	.078 [.069, .087]	337 vs. 9	4.190	2	.123
American clinical[3]									
Configural model[9]	184.647	54	<.001	.975	.049 [.042, .057]				
Scalar model[10]	264.184	70	<.001	.963	.053 [.046, .060]	10 vs. 9	78.277	16	<.001
23	181.539	56	<.001	.976	.048 [.040, .055]	23 vs. 9	2.273	2	.321
26	200.761	56	<.001	.973	.051 [.044, .059]	26 vs. 9	13.905	2	.001
41	185.705	56	<.001	.976	.048 [.041, .056]	41 vs. 9	5.931	2	.052
84	185.231	56	<.001	.976	.048 [.041, .056]	84 vs. 9	3.879	2	.144
231	179.271	56	<.001	.977	.047 [.039, .055]	231 vs. 9	2.042	2	.360
312	213.301	56	<.001	.970	.053 [.046, .061]	312 vs. 9	21.431	2	<.001
316	205.449	56	<.001	.972	.052 [.044, .060]	316 vs. 9	16.509	2	<.001

Table 6. *Measurement Invariance Tests Across Genders (Continued)*

Model	χ^2	*df*	*p*	CFI	RMSEA[90% CI]	Contrast	$\Delta\chi^2$	Δdf	*p*
329	205.142	56	<.001	.972	.052 [.044, .060]	329 vs. 9	16.891	2	<.001
337	192.191	56	<.001	.974	.050 [.042, .057]	337 vs. 9	8.766	2	.013
Korean clinical[4]									
Configural model[9]	86.894	54	.003	.963	.056 [.033, .076]				
Scalar model[10]	117.350	70	<.001	.947	.059 [.039, .077]	10 vs. 9	31.880	16	.010
ACT									
America normative[1]									
Configural model[11]	153.043	40	<.001	.941	.050 [.042, .058]				
Scalar model[12]	254.689	54	<.001	.894	.057 [.050, .064]	12 vs. 11	94.632	14	<.001
72	148.407	42	<.001	.944	.047 [.039, .056]	72 vs. 11	2.127	2	.345
81	244.073	42	<.001	.894	.065 [.057, .073]	81 vs. 11	58.461	2	<.001
166	153.461	42	<.001	.941	.048 [.040, .057]	166 vs. 11	3.785	2	.151
181	150.292	42	<.001	.943	.048 [.040, .056]	181 vs. 11	1.763	2	.414
207	158.428	42	<.001	.939	.049 [.041, .058]	207 vs. 11	7.271	2	.026
219	159.889	42	<.001	.938	.050 [.042, .058]	219 vs. 11	8.348	2	.015
267	165.536	42	<.001	.935	.051 [.043, .059]	267 vs. 11	12.192	2	.002
285	156.799	42	<.001	.940	.049 [.041, .057]	285 vs. 11	6.532	2	.038
Korean normative[2]									
Configural model[11]	88.320	40	<.001	.927	.043 [.031, .055]				
Scalar model[12]	144.969	54	<.001	.862	.051 [.041, .061]	12 vs. 11	53.018	14	<.001
72	90.026	42	<.001	.927	.042 [.030, .054]	72 vs. 11	3.332	2	.189
81	106.322	42	<.001	.903	.049 [.037, .060]	81 vs. 11	14.731	2	<.001
166	90.640	42	<.001	.926	.042 [.030, .054]	166 vs. 11	3.764	2	.152
181	102.836	42	<.001	.908	.047 [.036, .059]	181 vs. 11	12.448	2	.002
207	95.906	42	<.001	.918	.044 [.033, .056]	207 vs. 11	7.468	2	.024
219	89.392	42	<.001	.928	.042 [.030, .054]	219 vs. 11	2.877	2	.237
267	111.344	42	<.001	.895	.050 [.039, .062]	267 vs. 11	18.198	2	<.001
285	88.019	42	<.001	.930	.041 [.029, .053]	285 vs. 11	1.904	2	.386
American clinical[3]									

Table 6. *Measurement Invariance Tests Across Genders (Continued)*

Model	χ^2	*df*	*p*	CFI	RMSEA[90% CI]	Contrast	$\Delta\chi^2$	Δdf	*p*
Configural model[11]	260.940	40	<.001	.935	.075 [.066, .083]				
Scalar model[12]	325.086	54	<.001	.920	.071 [.064, .079]	12 vs. 11	78.794	14	<.001
72	247.127	42	<.001	.940	.070 [.062, .079]	72 vs. 11	.875	2	.646
81	315.962	42	<.001	.919	.081 [.073, .090]	81 vs. 11	39.274	2	<.001
166	250.781	42	<.001	.938	.071 [.062, .079]	166 vs. 11	2.218	2	.330
181	254.734	42	<.001	.937	.071 [.063, .080]	181 vs. 11	4.413	2	.110
207	251.446	42	<.001	.938	.071 [.063, .079]	207 vs. 11	1.463	2	.481
219	253.416	42	<.001	.938	.071 [.063, .080]	219 vs. 11	2.636	2	.268
267	257.392	42	<.001	.937	.072 [.064, .080]	267 vs. 11	6.231	2	.044
285	265.342	42	<.001	.934	.073 [.065, .082]	285 vs. 11	10.746	2	.005
Korean clinical[4]									
Configural model[11]	55.687	40	.051	.969	.045 [.000, .071]				
Scalar model[12]	83.744	54	.006	.941	.053 [.029, .074]	12 vs. 11	27.179	14	.018

Note. [1]*N* = 2,276. [2]*N* = 1,302. [3]*N* = 1,980. [4]*N* = 395. CFI = comparative fit index. TLI = Tucker-Lewis index. RMSEA = root mean square error of approximation. CI = confidence interval. Items with bold face are noninvariant items. Model 5, 7, 9, 11: Configural model. Model 6, 8, 10, 12: Scalar model. Item numbers correspond to the MMPI-2-RF item booklet.

After conducting measurement invariance tests across genders for each of the culture samples, measurement invariance across cultures was explored (Table 7, Figure 2 and 3). Measurement invariance across cultures was tested in models with and without gender noninvariant items, separately. The configural model fit was acceptable for all samples. The model fit difference between the scalar model and the configural model was significant for all samples except for clinical samples for the JCP scale with gender noninvariance, indicating only full scalar invariance across culture was reached for clinical samples. Then, noninvariant items were further identified.

For the JCP scale, one noninvariant items (item 21) was identified across normative samples with incorporating gender noninvariance, and four items (item 21, 96, 205, and 253) were identified without; no noninvariant item was identified across clinical samples with incorporating gender noninvariance, and five items (item 21, 66, 96, 223, and 253) were identified without. For the SUB scale, five noninvariant items (item 49, 86, 141, 192, and 237) were identified across normative samples with incorporating gender noninvariance, and all seven items (item 49, 86, 141, 192, 237, 266, and 297) were identified without; no noninvariant item was identified across clinical samples with incorporating gender noninvariance, and five items (item 49, 141, 192, 266, and 297) were identified without. For the AGG scale, three noninvariant items (item 84, 312, and 329) were identified across normative samples with incorporating gender noninvariance, and seven items (item 26, 41, 84, 231, 312, 316, and 329) were identified without; three noninvariant items (item 26, 312, and 329) were identified across clinical samples with incorporating gender noninvariance, and six items (item 23, 26, 84, 312,

Table 7. *Measurement Invariance Tests Across Cultures*

Model	χ^2	*df*	*p*	CFI	RMSEA[90% CI]	Contrast	$\Delta\chi^2$	Δdf	*p*
JCP									
Normative[1]									
Configural model[5]	79.776	25	<.001	.980	.035 [.027, .044]				
Scalar model[6]	155.057	35	<.001	.957	.044 [.037, .051]	6 vs. 5	76.131	10	<.001
21	117.216	27	<.001	.968	.043 [.035, .051]	21 vs. 5	33.082	2	<.001
66	93.152	27	<.001	.976	.037 [.029, .045]	66 vs. 5	13.014	2	.002
96	90.286	27	<.001	.977	.036 [.028, .045]	96 vs. 5	11.122	2	.004
205	91.324	27	<.001	.977	.036 [.028, .045]	205 vs. 5	11.911	2	.003
223	78.034	27	<.001	.982	.033 [.024, .041]	223 vs. 5	2.198	2	.333
253	79.175	27	<.001	.981	.033 [.025, .041]	253 vs. 5	3.761	2	.153
Normative[2]									
Configural model[5]	55.870	18	<.001	.986	.034 [.024, .045]				
Scalar model[6]	389.959	28	<.001	.869	.085 [.078, .093]	6 vs. 5	279.210	10	<.001
21	191.905	20	<.001	.938	.069 [.061, .078]	21 vs. 5	106.949	2	<.001
66	73.233	20	<.001	.981	.039 [.029, .048]	66 vs. 5	15.254	2	.001
96	93.132	20	<.001	.974	.045 [.036, .055]	96 vs. 5	26.749	2	<.001
205	206.509	20	<.001	.933	.072 [.063, .081]	205 vs. 5	102.039	2	<.001
223	59.317	20	<.001	.986	.033 [.024, .043]	223 vs. 5	5.489	2	.064
253	112.793	20	<.001	.966	.051 [.042, .060]	253 vs. 5	39.789	2	<.001
Clinical[3]									
Configural model[5]	70.819	23	<.001	.993	.042 [.031, .053]				
Scalar model[6]	79.214	33	<.001	.993	.034 [.025, .044]	6 vs. 5	18.528	10	.047
Clinical[4]									
Configural model[5]	62.512	18	<.001	.993	.046 [.034, .058]				
Scalar model[6]	111.947	28	<.001	.987	.050 [.041, .060]	6 vs. 5	49.224	10	<.001
21	224.456	20	<.001	.969	.093 [.082, .104]	21 vs. 5	140.698	2	<.001
66	219.049	20	<.001	.970	.091 [.081, .103]	66 vs. 5	139.745	2	<.001
96	107.437	20	<.001	.987	.061 [.050, .072]	96 vs. 5	30.679	2	<.001
205	64.471	20	<.001	.993	.043 [.032, .055]	205 vs. 5	5.722	2	.057

Table 7. *Measurement Invariance Tests Across Cultures (Continued)*

Model	χ^2	*df*	*p*	CFI	RMSEA[90% CI]	Contrast	$\Delta\chi^2$	Δdf	*p*
223	96.757	20	<.001	.988	.057 [.046, .068]	223 vs. 5	24.517	2	<.001
253	118.907	20	<.001	.985	.064 [.054, .076]	253 vs. 5	40.855	2	<.001
SUB									
Normative[1]									
Configural model[7]	364.544	38	<.001	.908	.069 [.063, .076]				
Scalar model[8]	565.548	50	<.001	.855	.076 [.070, .082]	8 vs. 7	221.360	12	<.001
49	397.811	40	<.001	.899	.071 [.064, .077]	49 vs. 7	36.035	2	<.001
86	397.490	40	<.001	.899	.071 [.064, .077]	86 vs. 7	37.253	2	<.001
141	387.949	40	<.001	.902	.070 [.063, .076]	141 vs. 7	26.683	2	<.001
192	405.723	40	<.001	.897	.071 [.065, .078]	192 vs. 7	38.725	2	<.001
237	369.125	40	<.001	.907	.068 [.062, .074]	237 vs. 7	16.037	2	<.001
266	368.071	40	<.001	.908	.068 [.061, .074]	266 vs. 7	14.066	2	.001
297	370.332	40	<.001	.907	.068 [.062, .074]	297 vs. 7	12.635	2	.002
Normative[2]									
Configural model[7]	348.438	28	<.001	.909	.080 [.073, .088]				
Scalar model[8]	1315.342	40	<.001	.639	.133 [.127, .140]	8 vs. 7	823.277	12	<.001
49	478.469	30	<.001	.873	.091 [.084, .099]	49 vs. 7	98.382	2	<.001
86	730.189	30	<.001	.802	.114 [.107, .121]	86 vs. 7	257.583	2	<.001
141	383.240	30	<.001	.900	.081 [.074, .088]	141 vs. 7	35.494	2	<.001
192	440.467	30	<.001	.884	.087 [.080, .095]	192 vs. 7	74.249	2	<.001
237	450.554	30	<.001	.881	.089 [.081, .096]	237 vs. 7	78.825	2	<.001
266	438.481	30	<.001	.884	.087 [.080, .095]	266 vs. 7	72.659	2	<.001
297	360.038	30	<.001	.906	.078 [.071, .086]	297 vs. 7	17.288	2	<.001
Clinical[3]									
Configrual model[7]	345.948	39	<.001	.948	.081 [.074, .089]				
Scalar model[8]	343.605	51	<.001	.950	.069 [.063, .077]	8 vs. 7	69.225	12	<.001
49	316.667	41	<.001	.953	.075 [.068, .083]	49 vs. 7	9.830	2	.007
86	300.466	41	<.001	.956	.073 [.065, .081]	86 vs. 7	1.401	2	.496
141	329.513	41	<.001	.951	.077 [.069, .085]	141 vs. 7	13.797	2	.001

Table 7. *Measurement Invariance Tests Across Cultures (Continued)*

Model	χ^2	*df*	*p*	CFI	RMSEA[90% CI]	Contrast	$\Delta\chi^2$	Δdf	*p*
192	320.591	41	<.001	.953	.076 [.068, .084]	192 vs. 7	6.764	2	.034
237	302.622	41	<.001	.956	.073 [.066, .081]	237 vs. 7	1.097	2	.578
266	316.904	41	<.001	.953	.075 [.068, .083]	266 vs. 7	6.552	2	.038
297	315.712	41	<.001	.953	.075 [.067, .083]	297 vs. 7	4.038	2	.133
Clinical[4]									
Configural model[7]	338.886	28	<.001	.947	.097 [.088, .106]				
Scalar model[8]	463.240	40	<.001	.928	.094 [.087, .102]	8 vs. 7	163.595	12	<.001
49	452.082	30	<.001	.928	.109 [.100, .118]	49 vs. 7	82.006	2	<.001
86	281.057	30	<.001	.957	.084 [.075, .093]	86 vs. 7	1.794	2	.408
141	543.667	30	<.001	.913	.120 [.111, .129]	141 vs. 7	155.952	2	<.001
192	438.631	30	<.001	.931	.107 [.098, .116]	192 vs. 7	92.607	2	<.001
237	282.058	30	<.001	.957	.084 [.075, .093]	237 vs. 7	.843	2	.656
266	322.830	30	<.001	.950	.091 [.082, .100]	266 vs. 7	20.922	2	<.001
297	326.678	30	<.001	.950	.091 [.082, .100]	297 vs. 7	19.519	2	<.001
AGG									
Normative[1]									
Configural model[9]	447.343	64	<.001	.899	.058 [.053, .063]				
Scalar model[10]	482.080	80	<.001	.894	.053 [.048, .058]	10 vs. 9	78.371	16	<.001
23	442.320	66	<.001	.901	.056 [.052, .062]	23 vs. 9	4.464	2	.107
26	442.456	66	<.001	.901	.056 [.052, .062]	26 vs. 9	4.463	2	.107
41	442.414	66	<.001	.901	.056 [.052, .062]	41 vs. 9	6.422	2	.040
84	505.398	66	<.001	.884	.061 [.056, .066]	84 vs. 9	50.763	2	<.001
231	448.937	66	<.001	.899	.057 [.052, .062]	231 vs. 9	10.558	2	.005
312	457.796	66	<.001	.897	.058 [.053, .063]	312 vs. 9	18.249	2	<.001
316	438.813	66	<.001	.902	.056 [.051, .061]	316 vs. 9	2.178	2	.337
329	464.196	66	<.001	.895	.058 [.053, .063]	329 vs. 9	20.150	2	<.001
337	451.506	66	<.001	.898	.057 [.052, .062]	337 vs. 9	9.521	2	.009
Normative[2]									
Configural model[9]	418.811	54	<.001	.896	.061 [.056, .067]				

Table 7. *Measurement Invariance Tests Across Cultures (Continued)*

Model	χ^2	*df*	*p*	CFI	RMSEA[90% CI]	Contrast	$\Delta\chi^2$	Δdf	*p*
Scalar model[10]	871.963	70	<.001	.772	.080 [.075, .085]	10 vs. 9	382.603	16	<.001
23	418.452	56	<.001	.897	.060 [.055, .066]	23 vs. 9	6.721	2	.035
26	583.633	56	<.001	.850	.073 [.067, .078]	26 vs. 9	121.183	2	<.001
41	545.037	56	<.001	.861	.070 [.065, .075]	41 vs. 9	94.158	2	<.001
84	743.558	56	<.001	.805	.083 [.078, .088]	84 vs. 9	238.780	2	<.001
231	532.200	56	<.001	.865	.069 [.064, .074]	231 vs. 9	86.571	2	<.001
312	485.948	56	<.001	.878	.066 [.060, .071]	312 vs. 9	49.458	2	<.001
316	501.268	56	<.001	.874	.067 [.061, .072]	316 vs. 9	64.238	2	<.001
329	448.791	56	<.001	.888	.063 [.057, .068]	329 vs. 9	28.082	2	<.001
337	413.469	56	<.001	.899	.060 [.054, .065]	337 vs. 9	1.633	2	.442
Clinical[3]									
Configural model[9]	294.594	67	<.001	.964	.053 [.047, .060]				
Scalar model[10]	309.246	83	<.001	.965	.048 [.042, .054]	10 vs. 9	49.065	16	<.001
23	304.715	69	<.001	.963	.054 [.048, .060]	23 vs. 9	12.068	2	.002
26	307.965	69	<.001	.963	.054 [.048, .060]	26 vs. 9	15.435	2	<.001
41	292.918	69	<.001	.965	.052 [.046, .058]	41 vs. 9	5.023	2	.081
84	306.262	69	<.001	.963	.054 [.048, .060]	84 vs. 9	13.501	2	.001
231	304.181	69	<.001	.963	.054 [.047, .060]	231 vs. 9	13.263	2	.001
312	322.958	69	<.001	.960	.056 [.050, .062]	312 vs. 9	26.539	2	<.001
316	304.397	69	<.001	.963	.054 [.047, .060]	316 vs. 9	12.628	2	.002
329	306.143	69	<.001	.963	.054 [.048, .060]	329 vs. 9	14.763	2	<.001
337	300.315	69	<.001	.964	.053 [.047, .059]	337 vs. 9	8.887	2	.012
Clnical[4]									
Configural model[9]	214.216	54	<.001	.974	.050 [.043, .057]				
Scalar model[10]	370.253	70	<.001	.951	.060 [.054, .066]	10 vs. 9	138.644	16	<.001
23	235.288	56	<.001	.971	.052 [.045, .059]	23 vs. 9	18.482	2	<.001
26	243.259	56	<.001	.970	.053 [.046, .060]	26 vs. 9	24.093	2	<.001
41	225.177	56	<.001	.972	.050 [.044, .057]	41 vs. 9	11.817	2	.003
84	236.273	56	<.001	.971	.052 [.045, .059]	84 vs. 9	20.673	2	<.001

Table 7. *Measurement Invariance Tests Across Cultures (Continued)*

Model	χ^2	*df*	*p*	CFI	RMSEA[90% CI]	Contrast	$\Delta\chi^2$	Δdf	*p*
231	225.692	56	<.001	.972	.050 [.044, .057]	231 vs. 9	12.099	2	.002
312	271.472	56	<.001	.965	.057 [.050, .064]	312 vs. 9	42.471	2	<.001
316	248.697	56	<.001	.969	.054 [.047, .061]	316 vs. 9	28.055	2	<.001
329	270.383	56	<.001	.965	.057 [.050, .064]	329 vs. 9	45.865	2	<.001
337	224.552	56	<.001	.973	.050 [.044, .057]	337 vs. 9	10.933	2	.004
ACT									
Normative[1]									
Configural model[11]	239.894	51	<.001	.931	.046 [.040, .051]				
Scalar model[12]	387.928	65	<.001	.882	.053 [.048, .058]	12 vs. 11	150.677	14	<.001
72	259.655	53	<.001	.925	.047 [.041, .052]	72 vs. 11	19.853	2	<.001
81	303.039	53	<.001	.909	.051 [.046, .057]	81 vs. 11	45.988	2	<.001
166	246.797	53	<.001	.929	.045 [.040, .051]	166 vs. 11	10.721	2	.005
181	233.822	53	<.001	.934	.044 [.038, .049]	181 vs. 11	1.202	2	.548
207	267.485	53	<.001	.922	.048 [.042, .053]	207 vs. 11	25.787	2	<.001
219	240.684	53	<.001	.931	.044 [.039, .050]	219 vs. 11	6.754	2	.034
267	244.859	53	<.001	.930	.045 [.039, .051]	267 vs. 11	8.793	2	.012
285	232.495	53	<.001	.934	.044 [.038, .049]	285 vs. 11	1.150	2	.563
Normative[2]									
Configural model[11]	199.115	40	<.001	.938	.047 [.041, .054]				
Scalar model[12]	1213.710	54	<.001	.551	.110 [.104, .115]	12 vs. 11	811.712	14	<.001
72	718.801	42	<.001	.738	.095 [.089, .101]	72 vs. 11	335.094	2	<.001
81	433.981	42	<.001	.848	.072 [.066, .078]	81 vs. 11	155.299	2	<.001
166	241.666	42	<.001	.923	.052 [.045, .058]	166 vs. 11	33.602	2	<.001
181	324.887	42	<.001	.890	.061 [.055, .068]	181 vs. 11	92.565	2	<.001
207	246.411	42	<.001	.921	.052 [.046, .059]	207 vs. 11	37.482	2	<.001
219	343.897	42	<.001	.883	.063 [.057, .070]	219 vs. 11	102.910	2	<.001
267	247.279	42	<.001	.920	.052 [.046, .059]	267 vs. 11	36.477	2	<.001
285	195.722	42	<.001	.940	.045 [.039, .052]	285 vs. 11	2.427	2	.297
Clinical[3]									

Table 7. *Measurement Invariance Tests Across Cultures (Continued)*

Model	χ^2	*df*	*p*	CFI	RMSEA[90% CI]	Contrast	$\Delta\chi^2$	Δdf	*p*
Configural model[11]	305.050	53	<.001	.936	.063 [.056, .070]				
Scalar model[12]	320.010	67	<.001	.936	.056 [.050, .063]	12 vs. 11	53.907	14	<.001
72	297.492	55	<.001	.939	.061 [.054, .068]	72 vs. 11	1.903	2	.386
81	298.070	55	<.001	.938	.061 [.054, .068]	81 vs. 11	1.825	2	.402
166	321.786	55	<.001	.932	.064 [.057, .071]	166 vs. 11	19.702	2	<.001
181	308.326	55	<.001	.936	.062 [.056, .069]	181 vs. 11	10.225	2	.006
207	323.043	55	<.001	.932	.064 [.057, .071]	207 vs. 11	20.023	2	<.001
219	313.620	55	<.001	.935	.063 [.056, .070]	219 vs. 11	13.246	2	.001
267	316.102	55	<.001	.934	.063 [.057, .070]	267 vs. 11	16.237	2	<.001
285	299.408	55	<.001	.938	.061 [.054, .068]	285 vs. 11	4.777	2	.092
Clinical[4]									
Configural model[11]	262.323	40	<.001	.942	.068 [.061, .076]				
Scalar model[12]	386.061	54	<.001	.913	.072 [.065, .079]	12 vs. 11	124.738	14	<.001
72	330.421	42	<.001	.924	.076 [.068, .084]	72 vs. 11	53.029	2	<.001
81	262.011	42	<.001	.942	.066 [.059, .074]	81 vs. 11	6.043	2	.049
166	309.529	42	<.001	.930	.073 [.066, .081]	166 vs. 11	35.981	2	<.001
181	271.729	42	<.001	.940	.068 [.060, .076]	181 vs. 11	13.138	2	.001
207	275.401	42	<.001	.939	.068 [.061, .076]	207 vs. 11	15.253	2	.001
219	268.948	42	<.001	.940	.067 [.060, .075]	219 vs. 11	10.910	2	.004
267	296.113	42	<.001	.933	.071 [.064, .079]	267 vs. 11	27.507	2	<.001
285	272.858	42	<.001	.939	.068 [.060, .076]	285 vs. 11	13.962	2	.001

Notes. *N* = 2,276 for the American normative. *N* = 1,302 for the Korean normative. *N* = 1,980 for the American clinical. *N* = 395 for the Korean clinical. Normative[1]: measurement invariance tests (American normative vs. Korean normative) with the gender incorporated. Normative[2]: measurement invariance tests (American normative vs. Korean normative) without the gender incorporated. Clinical[3]: measurement invariance tests (American clinical vs. Korean clinical) with the gender incorporated. Clinical[4]: measurement invariance tests (American clinical vs. Korean clinical) without the gender incorporated. Model 5, 7, 9, 11: Configural model. Model 6, 8, 10, 12: Scalar model. CFI = comparative fit index. TLI = Tucker-Lewis index. RMSEA = root mean square error of approximation. CI = confidence interval. Items with bold face are noninvariant items. Item numbers match the MMPI-2-RF

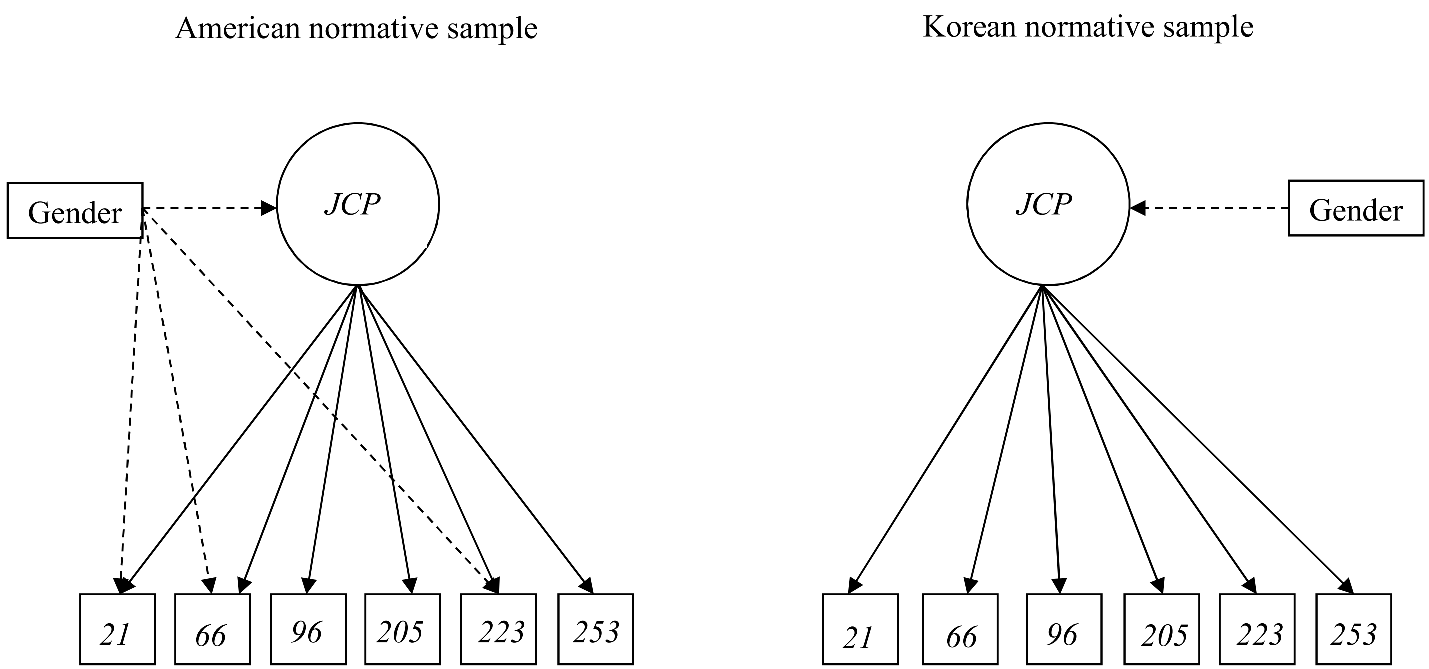

Figure 2. *The Hybrid Multi-group–MIMIC Model of JCP Scale Across Cultures (American Normative Sample vs. Korean Normative Sample) Incorporating Gender Noninvariant Items*

In Figure 2, *21, 66, 223* are noninvariant items for the American normative sample, and there are no noninvariant items for the Korean normative sample. Item numbers correspond to the MMPI-2-RF item booklet.

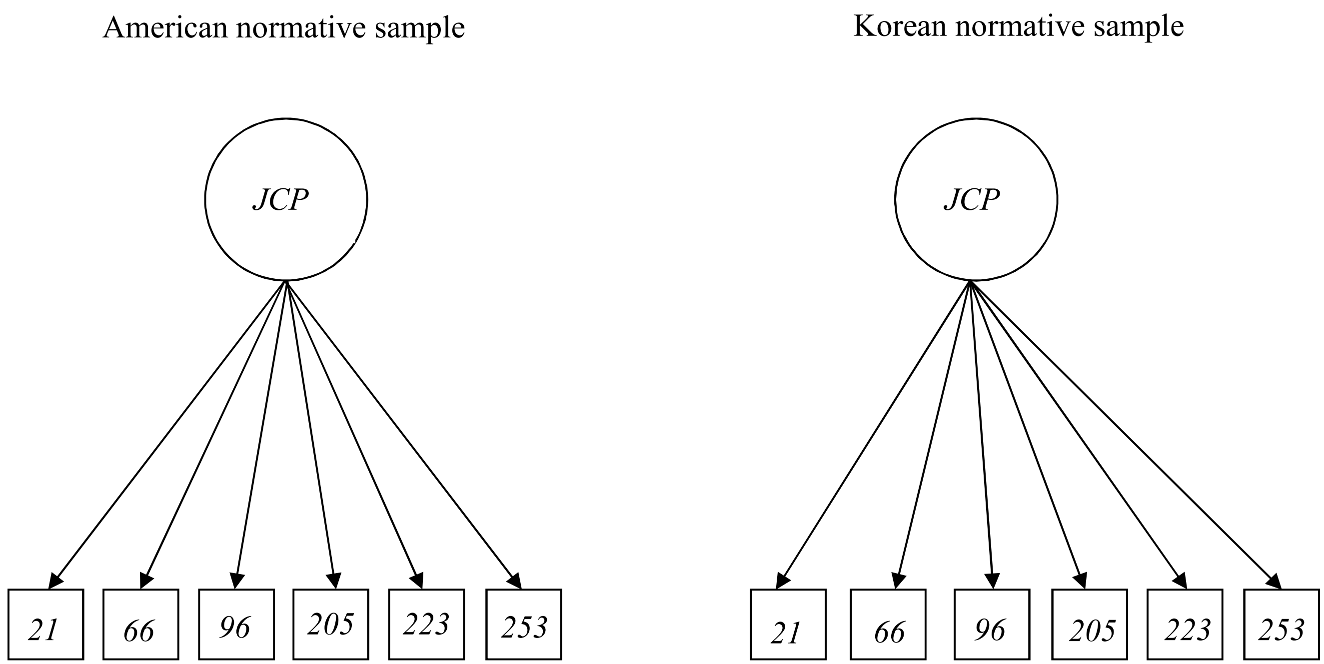

Figure 3. *The MI Model of JCP Scale Across Cultures (American Normative Sample vs. Korean Normative Sample) Without Incorporating Gender Noninvariant Items*
Item numbers correspond to the MMPI-2-RF item booklet.

316, and 329) were identified without. For the ACT scale, three noninvariant items (item 72, 81, and 207) were identified across American and Korean normative samples when incorporating gender noninvariant items, and seven items (item 72, 81, 166, 181, 207, 219, and 267) were identified without; three items (item 166, 207, and 267) were identified across American and Korean clinical samples with incorporating gender noninvariance, and three items (item 72, 166, and 267) were identified without.

For scales that reached partial scalar invariance across cultures, I followed Byrne et al.'s (1989) suggestion to simultaneously free the equality constraints on the factor loadings and thresholds of noninvariant items while keeping constraints of other items in the scalar model, and then I compared the latent means on each of the four Externalizing SP constructs across cultures for normative and clinical samples. In measurement invariance analyses, the Korean normative and clinical samples were set as the anchor groups, and the latent means of these samples were set as zero for identification purposes. Then, the latent means of the American normative and clinical samples were estimated, and latent means were further compared across cultures. The American normative sample had significant lower latent means than the Korean normative sample on the AGG ($Z = -.32$, $p < .01$) but similar latent means on the JCP, the SUB, and the ACT ($Z = -.10$, $p = .64$ for the JCP; $Z = -.25$, $p = .09$ for the SUB; $Z = -.03$, $p = .82$ for the ACT). The American clinical sample had a significantly higher latent mean on the ACT ($Z = .38$, $p = .01$) but a similar latent mean as the Korean clinical sample on the AGG ($Z = -.23$, $p = .36$). For the JCP and the SUB scales that have reached scalar invariance across Korean and American clinical samples, latent means were compared in the scalar models without freeing any equality constraint. The American clinical sample had a significantly higher latent mean on the JCP than Korean clinical sample ($Z = .46$, $p = .01$) but a similar latent mean on the SUB ($Z = .30$, $p = .11$).

CHAPTER IV

DISCUSSION

This study examined the measurement invariance of the Minnesota Multiphasic Personality Inventory-2 Restructured Form (MMPI-2-RF; Ben-Porath & Tellegen et al., 2008) Externalizing Specific Problems scales across American and Korean normative and clinical samples. The first hypothesis was partially supported: All SP items are well represented by a one-factor model except the Korean normative sample on the AGG Scale, which partially fulfilled the scale construction goal of measuring specific clinical constructs. Regard to the second hypothesis, results showed that the American samples had an average higher standardized factor loading than the Korean samples, and the clinical samples had an average higher factor loading than the normative samples. These results suggest that scale items are better indicators of Activation, Aggression, Juvenile conduct problems, and Substance abuse for the American samples than for the Korean samples and better indicators for the clinical samples than for the normative samples. This is in accordance with Chen's (2008) cross-cultural measurement invariance review, in which the majority of the scales had an average higher factor loading in its source culture than in its imported culture. As the goal of the MMPI-2 RF is to assess personality and psychopathology, it is reasonable that these items would be better indicators for clinical samples than for normative samples.

Measurement Invariance Across Genders

Before testing the measurement invariance across cultures, I tested the measurement invariance of the scales across genders for each culture group. Externalizing SP scales exhibited partial scalar invariance with some gender noninvariant items for all American clinical and normative samples. For the Korean samples, by contrast, most of the Externalizing SP scales showed measurement invariance, except that the ACT and the AGG scales only reached partial scalar invariance for Korean normative sample. The ACT had two gender noninvariant items and the AGG had one gender noninvariant item.

The American samples have twenty-one gender noninvariant items, and the Korean samples have three noninvariant items. The American samples have more gender noninvariant items than the Korean samples in general. Before combining two gender groups into one culture group in conducting measurement invariance, special attention needs to be paid to the noninvariant items—individuals with the same latent mean on the constructs, from different gender groups, responded differently to the items—as well as the possible explanations for the bias before applying noninvariant items. Explanations for some of these noninvariant items are offered in the following passages.

For the JCP scale, the majority of the noninvariant items for American samples (item 21, 66, 233, 253) are measuring behaviors related to externalizing problems. This may be explained by the debate about sex bias in conduct disorder diagnostic criteria. The prevalence rate of conduct disorder is much higher among males than females, with 2.1 to 8.8% in males and 0 to 1.4% in females; however, Moffitt et al. (2008) argued that the conduct disorder symptom criteria focus on behaviors more relevant to males and may fail to detect among females. No noninvariant item was found in Korean samples. For the SUB scale, the gender noninvariant item 141 and item 192, as found in the American samples, are related to alcohol drinking. No noninvariant item was found in the Korean samples.

For the AGG scale, most of the gender noninvariant items for the American normative sample (84, 316, 329) are related to direct physical aggression, that is, physical behaviors that people use to harm others in direct encounters. This is consistent with the existing literature on gender difference in aggression, which has concluded that males are more likely to interpret aggression in a more direct physical way, and females are more likely to interpret aggression in a more indirect verbal way (e.g., Bettencourt & Miller, 1996; Eagly & Steffen, 1986; Frodi, McCauley, & Thome, 1977; Richardson & Hammock, 2007).

The noninvariant item 337 is related to anger. Men and women view their experience of aggression differently. In general, women view aggression as a loss of self-restraints and express guilt about it, whereas men view it as a socially useful source of control over other people and felt less guilty (Campbell & Muncer, 1987). Serotonin (5-HT) is identified as the transmitter responsible for constraining control over aggression. Researchers have found a gender difference in serotonin in the frontal cortex in that females are higher than males in serotonin uptake, which results in females having better inhibitory control.

For the ACT scale, item 81 was shown to be a gender noninvariant item in the American normative and clinical samples and the Korean normative samples. In general, women tend to experience and express more intense emotions than men, and they are more likely to seek out emotional support than men (Kring & Gordon, 1998; Grossman & Wood, 1993; Balswick & Avertt, 1977). Regarding emotional regulation, men are reported to use more disengagement strategies than women to lower their emotional intensity (Davis, Greenberger, Charles, Chen, Zhao, & Dong, 2012).

Measurement Invariance Across Cultures

After incorporating gender noninvariant items in conducting measurement invariance of the scales across cultures, most of the scales reached partial scalar invariance except that JCP reached full scalar invariance for the clinical samples. As the prerequisite for further cross-cultural comparisons—partial scalar invariance—was met, I further compared means across two culture samples. For the normative samples, the JCP, the SUB, and the ACT were similar between the American and Korean samples, whereas the AGG was different, with the American sample scoring lower than the Korean sample on the AGG. For the clinical samples, the SUB and the AGG were similar, whereas the JCP and the ACT were different, with the American sample scoring higher than the Korean sample on the JCP and the ACT. However, before interpreting these Externalizing Specific Problems constructs cross-culturally, special attention needed to be paid to the non-invariant items—individuals with the same latent mean on the constructs, from distinct cultural groups, responded differently to the items—as well as the possible explanations for the bias before applying non-invariant items cross-culturally.

For Juvenile Conduct Problems construct, the Korean normative sample yielded a lower endorsement on item 21 than the American normative sample at the same levels of the JCP score; however, clinical samples function the same on this item. One of the possibilities was that Korean children were taught social acceptable manners at home and expected to behave properly when they were young. In Korea, children's manners reflect their parents' parenting. A well-behaved child brings honor to the family whereas an ill- behaved child brings shame and criticisms to parents that they are not teaching their child properly (Yang, 2009). The latent mean comparisons showed that normative samples were similar on the JCP, but American clinical sample had a significant higher latent mean than Korean clinical sample on the JCP. Majority of the JCP items (4 out of 6) are related to school and classroom behaviors. Items of the JCP and the SUB overlap significantly with two MMPI-2 RC4 clusters (Substance Abuse and School Problems; Wang, 2014). Among the JCP's six items, four of them are in five-item RC4

School Problems cluster, whereas five of the seven items on SUB are included in five-item RC4 Substance Abuse cluster. This is consistent with the finding from Wang (2014) that the Korean clinical sample had a lower latent means on RC4 School Problems factor than the American clinical sample.

For the Substance Abuse construct, noninvariant item 141, item 86, and item 192 are related to drinking, that at the same level of the Substance Abuse latent score, Korean sample has a higher endorsement than American sample on the items. This may be related to the Korean drinking culture. Korean society has a great tolerance towards heavy drinking, even other forms of drugs use are prohibited. Men are encouraged to drink and their drinking capacities sometimes indicate social competence. Drinking together is commonly seen as a way of deepening friendships (Kwon-Ahn, 2001). Korean and American normative samples had a similar latent mean on the SUB after freely estimating these five noninvariant items across cultures. As majority of the items (five out of eight) were detected as noninvariant items, interpretation of the latent mean comparisons need to be cautious. There are five noninvariant items when examined measurement invariance across American and Korean clinical samples without gender noninvariant items, and no noninvariant item after incorporating gender noninvariant items. Two clinical samples also had a similar latent mean on the SUB. This is similar with the finding from Wang (2014) that American and Korean clinical sample had a similar latent mean on RC4 Substance Abuse factor.

There is limited existing research on aggression across cultures. One study examined fourth and eighth graders' aggressive behavior through school principals' reports across 62 countries, and found it was related to countries' individualistic versus collectivistic value (Bergmüller, 2013). After controlling for school and country characteristics, individualistic versus collectivistic value was found to be a strong predictor of principal-reported student physical and verbal aggression. Compared to the school principals that from countries of more individualistic culture, school principals from more collectivistic cultures reported less aggressive behaviors of students. Based on Hofstede et al. (2010), the individualistic index score for Korea is 18, which is much lower than that of United States at 91. Principals reported less student physical and verbal aggression in Korea than in the United States. Another study examined the measurement structure of Aggression Questionnaire (AQ; Buss & Perry, 1992) and Conditional Reasoning Test for Aggression (CRT-A; James et al., 2005) using American and Korean college samples, and found that the aggression measurement structures were not invariant across cultures. In the present study, one third of the items in the scale (3 out of 9) showed to be noninvariant across American and Korean cultures, and Korean normative sample had a higher latent mean than American normative sample on the aggression construct.

For the Activation construct, noninvariance in some of the items (item 81, 166, 207) may be due to the cultural difference in emotional arousal level between Korean and United States. Lim (2016) reviewed the literature on cultural differences in emotion and found that arousal level of emotions was related to individualistic versus collectivistic aspects of the culture. The United States is considered an individualistic culture, and it values high arousal emotions, whereas Korea is more of a collectivistic culture which emphasizes low arousal emotions (Hofstede et al., 2010). This may further explain the higher latent mean that the American clinical sample had compared to the Korean sample on the ACT scale.

Strengths, Limitations, and Future Study

Due to increasing globalization, researchers are increasingly interested in cross-cultural studies. The conventional research method is to compare scale means of the construct across cultural groups. Now, more researchers have realized the importance of measurement invariance to see whether scales function the same across cultural groups and then to compare latent means across cultures. However, most of the studies which test the measurement invariance across cultures fail to examine the influence of the gender variable to the instrument. It is likely that most researchers assumed that items function the same for males and females within the culture, so they proceed the research question directly to the culture level to test whether these items function similarly across cultures. This assumption is contradicted with the existing knowledge that gender difference is existed in various of aspects, such as cognition, personality, and social behavior (Costa, Terracciano, & McCrae, 2001; Giudice, 2015; Maylor et al., 2007). Unlike previous studies which only tested measurement invariance of the MMPI-2 RC across American and Korean samples (Ketterer, 2010; Wang, 2013), one of the strengths of the present study is that it tested measurement invariance across gender and cultures.

One way to test measurement invariance across genders and cultures is to test the measurement invariance across genders first, and then test measurement invariance across cultures combining the two gender groups (or ignoring gender). A recent study used this method to test measurement invariance of the Violent Behavior Checklist—Modified across genders and ethnicities (Cotter, Evans, & Smokowski, 2017). Because this method ignores the gender variance when testing measurement invariance across ethnicities, it is possible that gender variance may balance out the ethnicity variance and that, therefore, the method will provide incomplete information on measurement invariance across ethnicities. To overcome this methodological weakness, I used a hybrid multi-group—MIMIC model to incorporate gender noninvariant items when testing measurement invariance across cultures. As the result, I found fewer noninvariant items across cultures after corporating gender noninvariant items in examining measurement invariance across cultures. It is concluded that considering gender, a largely overlooked variable by researchers when conducting measurement invariance across cultures, could eliminate some degree of variance across cultures. It is recommended that future research incorporate gender when testing measurement invariance across cultures. Another strength of the study is using both normative and clinical samples. There are two existing studies on measurement invariance of MMPI-2/MMPI-2 RF across American and Korean samples (Ketterer, 2010; Wang, 2013). One study only used American and Korean normative samples, and the other one only used American and Korean clinical samples. The present study is the first one using both normative and clinical samples.

Three limitations of the study should be noted. When researchers collected the American and Korean normative samples, they used a proportional stratified sampling method to control sample bias by matching the sample to census data demographics. The American normative sample was selected to match to U.S. 1980 Census in geographic representation, age, and ethnicity. The Korean normative sample was selected to match Korean 2000 Census in geographical region, place of residence, age, and education. Even though sample bias was controlled by collecting data that was matched to the Census, one limitation of the present study is that it may introduce cohort effects due to the 20-year gap between the year of collection of the American and Korean normative samples. It is possible that externalizing specific problem constructs measured in the United States are not the same as the ones measured in Korea 20 years later. The same problem exists for the clinical samples because the U.S. outpatient sample was collected in the 1990s whereas the Korean clinical sample was collected in 2009–2011 (Graham, Ben-Porath, & McNulty, 1999; Han et al., 2011a). It is recommended that future research replicate the study by using more recent datasets. Another limitation of the study is that I did not include some demographic variables of participants that were related to Externalizing Specific Problems (e.g., drinking habit, substance abuse history) due to the inconsistent demographic information in these datasets. In the previous RC4 study, age and gender were both controlled as covariates (Wang, 2013). In the present study, I did not include the age variable. It is recommended that future research incorporate age. A third limitation of this study is that no further information on effect size in measurement invariance testing was available. Even noninvariant items were identified after some of the scales failed to reach full scalar invariance, the magnitude of noninvariance among these noninvariant items was unknown due to limitation of the MGCFA methodology. It is recommended that future study use a different methodology for item analyses.

In the previous discussion, I proposed some possible cultural explanations for the noninvariant items in the Korean samples. However, it is possible that noninvariance in some of the items not only come from differences in culture but also from difference in response style. Response style refers the tendency a group has to respond systematically when compared to the other group despite item content (Baumgartner & Steenkamp, 2001). For example, Chen, Lee, and Stevenson (1995) found in their study that Japanese and Chinese college students used a more middle response style (MRS) than their Canadian and American counterparts in answering rating scales, and American students used the most extreme response style (ERS). They further explained that usage of MRS may be due to the emphasis on modesty in Asian cultures. The similar ERS trend also found in Chun, Campbell, and Yoon's (1974) and Lee's (2012) study that

American college students used more ERS than Korean college students. Also, Lee (2012) found that Koreans were not different than Americans in middle response style (MRS) when answering Aggression Questionnaire (Buss & Perry, 1992). I recommend that future research further detect response style in Korean samples to test whether noninvariance in some of the items is due to response style.

In general, the findings of this study indicate that Externalizing Specific Problem constructs are similar across American and Korean samples, implying that cross-cultural comparisons on MMPI-2 Externalizing SP scales could be done with considerations of noninvariant items for Korean sample. One of the interesting findings is that the Korean normative sample has a similar latent mean as the American normative sample on the majority of the scales (three out of four). Han, Moon, Lee and Kim, (2011) reported that the Korean normative sample had raw elevated scale scores on most of the MMPI-2 Restructured Clinical (RC) and Specific Problems (SP) scales when compared to American norm. It is possible that this elevation is not due to the true latent mean difference but to the biased items. It is recommended that future research extend to other MMPI-2 RC and SP scales to test whether elevated scale scores in Korean normative sample truly reflect latent mean differences. As MMPI-2 RF was developed in the United States using American culture, it is possible that it ignores some of the Externalizing symptoms specific to Korean culture. Therefore, future research should use an emit-etic approach to develop cultural-specific Externalizing SP scales (Cheung, van de Vijver, & Leong, 2011).

REFERENCES

Arbisi, P. A., Sellbom, M., & Ben-Porath, Y. S. (2008). Empirical correlates of the MMPI-2 Restructured Clinical (RC) scales in psychiatric inpatients. *Journal of Personality Assessment, 90*, 122–128.

Anestis, J. C., Gottfried, E. D., & Joiner, T. E. (2015). The utility of MMPI-2-RF substantive scales in prediction of negative treatment outcomes in a community mental health center. *Assessment, 22*, 23–35.

Balswick, J., & Avertt, C. P. (1977). Differences in expressiveness: Gender, interpersonal orientation, and perceived parental expressiveness as contributing factors. *Journal of Marriage and Family, 39*, 121–127.

Baumgartner, H., & Steenkamp, J. B. (2001). Response style in marketing research: A cross-national investigation. *Journal of Marketing Research, 38*, 143–156.

Ben-Porath, Y. S., & Tellegen, A. (2008). *MMPI-2 Restructured Form: Manual for administration, scoring, and interpretation*. Minneapolis, MN: University of Minnesota Press.

Bergmüller, S. (2013). The relationship between cultural individualism-collectivism and student aggression across 62 countries. *Aggressive Behavior, 39*, 182–200.

Betttencourt, B. A., & Miller, N. (1996). Gender differences in aggression as a function of provocation: A meta-analysis. *Psychological Bulletin, 119*, 422–447.

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.

Bryne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*, 456–466.

Buss, A. H., & Perry, M. P. (1992). The aggression questionnaire. *Journal of Personality and Social Psychology, 63*, 452–459.

Butcher, J. N. (1996). Translation and adaptation of the MMPI-2 for international use. In J. N. Butcher (Ed.), *International adaptations of the MMPI-2: Research and clinical applications* (pp. 26–43). Minneapolis, MN: University of Minnesota Press.

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., tellegen, A., & Kaemmer, B. (1989). *Minnesota Multiphasic Personality Inventory-2 (MMPI-2): Manual for administration, scoring, and interpretation*. Minneapolis, MN: University of Minnesota Press.

Butcher, J. N., Graham, J. R., Ben-Porath, Y. S., Tellegen, A., Dahlstrom, W. G., & Kaemmer, B. (2001). *MMPI-2 (Minnesota Multiphasic Personality Inventory-2): Manual for administration, scoring, and interpretation, revised edition*. Minneapolis, MN: University of Minnesota Press.

Butcher, J. N., & Pancheri, P. (1976). *A handbook of cross-national MMPI research*. Minneapolis, MN: University of Minnesota Press.

Campbell, A., & Muncer, S. (1987). Models of anger and aggression in the social talk of women and men. *Journal for the Theory of Social Behavior, 17*, 489-512.

Chen, C., Lee, S. L., & Stevenson, H. W. (1995). Response style and cross-cultural comparisons of rating scales among East Asian and North American students. *Psychological Science, 6*, 170–175.

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology, 95*, 1005–1018.

Cheung, F. M., van de Vijver, F. J. R., & Leong, F. T. L. (2011). Towards a new approach to the study of personality in culture. *American Psychologist, 66*, 593–603.

Chun, K. T., Campbell, J. B., & Yoon, J. H. (1974). Extreme response style in cross-cultural research. *Journal of Cross-Cultural Psychology, 5*, 465–480.

Crick, N., & Crotpeter, J, (1995). Relational aggression, gender, and social-psychological adjustment. *Child Development, 66*, 710–722.

Crick, N. R., & Zahn-Waxler, C. (2003). The development of psychopathology in females and males: Current progress and future challenges. *Development and Psychopathology, 15*, 719–742.

Dahlstrom, W. G., Welsh, G. S., & Dahlstrom, L. E. (1975). *An MMPI handbook: Volume II. Research applications*. Minneapolis, MN: University of Minnesota Press.

Davis, E., Greenberger, E., Charles, S., Chen, C., Zhao, L., & Dong, Q. (2012). Emotion experience and regulation in China and the United States: How do culture and gender shape emotion responding? *International Journal of Psychology, 47*, 230–239.

Doll, B. (1996). Prevalence of psychiatric disorders in children and youth: An agenda for advocacy by school psychology. *School Psychology Quarterly, 11*, 20–-47.

Eagly, A. H., & Steffen, V. J. (1986). Gender and aggressive behavior: A meta-analytic review of social psychological literature. *Psychological Bulletin, 100*, 309–330.

Forbey, J. D., Lee, T. T. C., & Handel, R. W. (2010). Correlates of the MMPI-2-RF in a college setting. *Psychological Assessment*, *22*, 737–744.

Frodi, A., Macaulay, J., & Thome, P. (1977). Are women always less aggressive than men? A review of the experimental literature. *Psychological Bulletin, 84*, 634-660.

Giudice, M. D. (2015). Gender differences in personality and social behavior. In J. D. Wright (Ed.), *International encyclopedia of the social and behavioral sciences* (2nd ed.) (pp. 750-756). New York, NY: Elsevier.

Graham, J. R., Ben-Porath, Y. S., and McNulty, J. L. (1999). *MMPI-2 correlates for outpatient community mental health settings*. Minneapolis, MN: University of Minnesota Press.

Grossman, M. & Wood, W. (1993). Sex differences in intensity of emotional experience: A social role interpretation. *Journal of Personality and Social Psychology, 65*, 1010–1022.

Haber, J. C. (2013). *MMPI-2-RF scales as predictors of psychiatric diagnoses*. (Doctoral dissertation). Retrieved from ProQuest. (UMI Number: 3579641)

Han, K. (1993). *The use of the MMPI-2 in Korea: Inventory adaptation, equivalence evaluation, and initial validation*. (Doctoral dissertation). Retrieved from ProQuest. (UMI Number: 9328342)

Han, K. (1996). The Korean MMPI-2. In J. Butcher (Ed.). *International Adaptation of the MMPI-2: Research and clinical applications* (pp.88–136). Minneapolis, MN: University of Minnesota Press.

Han, K., Moon, K., Lee, J., & Kim, J. (2011a). *Minnesota Multiphasic Personality Invery-2: Revised Manual.* Seoul, Korea: Maumsarang.

Han, K., Moon, K., Lee, J., & Kim, J. (2011b). *Minnesota Multiphasic Personality Inventory-2 Restructured Form Manual.* Seoul, Korea: Maumsarang.

Han, K., Park, H. I., Weed, N. C., Lim, J., Johnson, A., & Joles, C. (2013). Gender differences on the MMPI across American and Korean adult and adolescent normative samples. *Journal of Personality Assessment, 95*, 197–206.

Hartung, C. M., & Widiger, T. A. (1998). Gender differences in the diagnosis of mental disorders: Conclusions and controversies of the DSM-IV. *Psychological Bulletin,123*, 260–278.

Hathaway, S. R., & McKinley, J. C. (1943). *The Minnesota Multiphasic Personality Inventory*. Minneapolis, MN: University of Minnesota Press.

James, L. R., McIntyre, M. D., Glisson, C. A., Green, P. D., Patton, T. W., LeBreton, J. M.,...Williams, L. J. (2005). A conditional reasoning measure for aggression. *Organizational Research Methods, 8*, 69–99.

Kim, J., Han, K., Lim, J., Lee, J., Min, B., & Moon, K. (2005). *Korean MMPI-2 user manual*. Seoul, Korea: Maumsarang.

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford Press.

Kring, A. M. & Gordon, A. H. (1998). Sex differences in emotion: Expression, experience, and physiology. *Journal of Personality and Social Psychology, 74*, 686–703.

Lee, H. J. (2012). Equivalence and faking issues of the aggression questionnaire and the conditional reasoning test for aggression in Korean and American samples (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3571294)

Lim, N. (2016). Cultural differences in emotion: Differences in emotional arousal level between the East and the West. *Integrative Medicine Research, 5*, 105–109.

Marsh, H. W., Tracey, D. K., & Craven, R. G. (2006). Multidimensional self-concept structure for preadolescents with mild intellectual disabilities: A hybrid multigroup-MIMIC approach to factorial invariance and latent mean differences. *Educational and Psychological Measurement, 66*, 795–818.

Mattson, C., Powers, B., Halfaker, D., Akeson, S., & Ben-Porath, Y. (2012). Predicting drug court treatment completion using the MMPI-2-RF. *Psychological Assessment, 24*, 937–943.

Maughan, B., Rowe, R., Messer, J., Goodman, R., & Metzer, H. (2004). Conduct disorder and oppositional defiant disorder in a national sample a developmental epidemiology. *Journal of Child Psychology and Psychiatry, 45*, 609–621.

Maylor, E. A., Reimers, S., Choi, J., Collaer, M. L., Peters, M., & Silverman, I. (2007). Gender and sexual orientation differences in cognition across adulthood: Age is kinder to women than to men regardless of sexual orientation. *Archives of Sexual Behavior. 36*, 235–249.

Moffitt, T. E., Arsenneault, L., Jaffee, S. R., Kim-Cohen, J., Koenen, K. C., Odgers, C. L.,…Viding, E. (2008). Research review: DSM-V conduct disorder: research needs for an evidence base. *Journal of Child Psychology and Psychiatry, 49*, 3–33.

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Nye, C. D., Roberts, B. W., Saucier, G., & Zhou, X. (2008). Testing the measurement equivalence of personality adjective items across cultures. *Journal of Research in Personality, 42*, 1524–1536.

Österman, K., Björkqvist, K., & Lagerspetz, K. M. J. (1998). Cross-cultural evidence of female indirect aggression. *Aggressive Behavior, 24*, 1–8.

Richardson, D. S., & Hammock, G. S. (2007). Social context of human aggression: Are we paying too much attention to gender? *Aggression and Violent Behavior, 12*, 417–426.

Richardson, D. R., & Green, L. R. (1999). Social sanction and threat explanations of gender effects on direct and indirect aggression. *Aggressive Behavior, 25*, 425–434.

Rosenfild, S. (1999). Gender and mental health: Do women have more psychopathology, men more, or both the same (and why)? In A. V. Horwitz & T. L. Scheid (Eds.), *A handbook for the study of mental health: Social contexts, theories, and systems* (pp. 348–360). New York, NY: Cambridge University Press.

Sellbom, M., Bagby, R. M., Kushner, S., Quilty, L. C., & Ayearst, L. E. (2012). Diagnostic construct validity of MMPI-2 Restructured Form (MMPI-2-RF) scale scores. *Assessment, 19*, 176–186.

Sideridis, G. D., Tsaousis, I., & Al-harbi, K. A. (2015). Multi-population invariance with dichotomous measures: Combining multi-group and MIMIC methodologies in evaluating the general aptitude test in the Arabic language. *Journal of Psychoeducational Assessment, 33*, 568–584.

Substance Abuse and Mental Health Services Administration. (2014, April 3). *The TEDS report*. Retrieved from http://www.samhsa.gov/data/sites/default/files/sr077-gender-differences-2014.pdf

Tarescavage, A. M., Windover, A., Ben-Porath, Y. S., Boutacoff, L. I., Marek, R. J., Ashton, K., … Heinberg, L. J. (2013). Use of the MMPI-2-RF Suicidal/death Ideation and Substance Abuse scales in screening bariatric surgery candidates. *Psychological Assessment, 25*, 1384–1389.

Tellegen, A., Ben-Porath, Y. S., McNulty, J. L., Arbisi, P. A., Graham, J. R., & Kaemmer, B. (2003). *MMPI-2 restructured clinical (RC) scales: Development, validation, and interpretation*. Minneapolis, MN: University of Minnesota Press.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practice, and recommendations for organizational research. *Organizational Research Methods, 3*, 4–70.

Wagner, G. A., Stempliuk, V. A., Zilberman, M. L., Barroso, L. P., & Guerra de Andrade, A. (2007). Alcohol and drug use among university students: gender differences. *Revista Brasileira de Psiquiatria*, 29, 123–129.

Wang, J. (2014). *Establishing measurement invariance of the MMPI-2 Restructured Clinical Scale 4 (Antisocial Behavior) Using American and Korean clinical samples* (Unpublished master thesis). Central Michigan University, Mount Pleasant, MI.

Wang, J., Pudumjee, S. B., Courrégé, S. C., Weed, N. C., Han, K., Ben-Porath, Y. S., Moo, K., & Kim, J. (2015, June). *Comparing gender differences on the MMPI-2 RF across American and Korean normative and clinical samples*. Poster presented at the 50th Annual MMPI Symposium, Minneapolis, MN.

Yang, S. (2009). Cane of love: Parental attitudes towards corporal punishment in Korea. *British Journal of Social Work, 39*, 1540–1555.