Summer 2024

Gender Measurement Invariance of the Minnesota Multiphasic Personality Inventory - Adolescent - Restructured Form (MMPI-A-RF) Internalizing and Externalizing Specific Problem Scales

Thomas Jay Augustin
Old Dominion University, taugustin93@gmail.com

GENDER MEASUREMENT INVARIANCE OF THE MINNESOTA MULTIPHASIC PERSONALITY INVENTORY-ADOLESCENT-RESTRUCTURED FORM (MMPI-A-RF) INTERNALIZING AND EXTERNALIZING SPECIFIC PROBLEMS SCALES

by

Thomas Jay Augustin

B.S. December 2015, University of Nebraska Kearney
B.M. December 2015, University of Nebraska Kearney
M.S. July 2018, Fort Hays State University

A Dissertation Submitted to the Graduate Faculties of
Eastern Virginia Medical School
Norfolk State University
Old Dominion University
in Partial Fulfillment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

CLINICAL PSYCHOLOGY

VIRGINIA CONSORTIUM PROGRAM IN CLINICAL PSYCHOLOGY
August 2024

Approved by:

Richard Handel (Director)

Jennifer Flaherty (Member)

James Paulson (Member)

Kevin Waymire (Member)

ABSTRACT

GENDER MEASUREMENT INVARIANCE OF THE MINNESOTA MULTIPHASIC PERSONALITY INVENTORY-ADOLESCENT-RESTRUCTURED FORM (MMPI-A-RF) INTERNALIZING AND EXTERNALIZING SPECIFIC PROBLEMS SCALES

Thomas Jay Augustin
Virginia Consortium Program in Clinical Psychology, 2024
Director: Dr. Richard Handel

Due to social, psychological, biological, and cultural differences in adolescent development, it is important to evaluate measurement invariance in psychological measures. Although the Minnesota Multiphasic Personality Inventory (MMPI) assessments have a plethora of research to evaluate their utility in clinical practice, few published studies have examined the measurement invariance between males and females for any of the scales in the adolescent versions (MMPI-A/MMPI-A-RF). The present study examined the measurement invariance of the MMPI-A-RF Internalizing and Externalizing Specific Problem Scales between male and female adolescents. Data were obtained from an outpatient sample and from Pearson's mail-in service resulting in 1,622 valid protocols (811 boys and 811 girls) that were examined. Five of the nine Internalizing Scales (Anger Proneness, Anxiety, Self-Doubt, Specific Fears, and Stress/Worry) obtained full measurement invariance. The Behavior Restricting Fears Scale was dropped from further analyses due to a Heywood case in the girls' sample. For the remaining three Internalizing Scales, two (Inefficacy and Obsessions/Compulsions) met partial measurement invariance and the last scale (Helplessness/Hopelessness) only reached configural invariance. Three of the six Externalizing Scales (Antisocial Attitudes, Negative Peer Influence, and Negative School Attitudes) obtained full measurement invariance. For the remaining three Externalizing Scales, two (Aggression and Conduct Problems) only met configural invariance.

The final Externalizing Scale (Substance Abuse) did not meet configural invariance. Overall, results indicated that many of the existing MMPI-A-RF Externalizing and Internalizing Specific Problems Scales obtained measurement invariance in varying degrees. This study further identified several scales where test developers, researchers, and practitioners should be cognizant of the influence of noninvariant items (i.e., the HLP, NFC, OCS, AGG, and CNP Scales), as well as limitations of the BRF and SUB Scales.

ACKNOWLEDGMENTS

I wish to extend my deepest appreciation and admiration to my dissertation chair, Richard Handel, Ph.D., for assisting me through this process. Your guidance, knowledge, and encouragement have inspired my research and future endeavors.

A special thank you to the remaining members of my dissertation committee, Jennifer Flaherty, Ph.D., James Paulson, Ph.D., and Kevin Waymire, Ph.D. I am greatly appreciative to you all for your flexibility and feedback to make this project possible.

I want to recognize Bruce Cappo, Ph.D., ABPP who not only served as a mentor but also generously provided a plethora of archival data which made this research possible. I would like to recognize Pearson Clinical who supplemented Dr. Cappo's data set with the necessary protocols to obtain a stronger sample size.

I want to recognize two Virginia Consortium Program alumni and friends, Madison Smart-McCarthy, Ph.D., for helping with the entire data entering process and Cassidy Sandoval for giving me the encouragement and laughs I needed through this process.

I wish to extend a sincere thank you to all my friends who have supported me through this process including a special shout-out to Elaine Clifford for taking the time to read through this document and provide necessary feedback and Xiaonan Zhang for spending endless dissertation weekends with me.

Finally, I wish to acknowledge and offer my most sincere appreciation to my family and husband. You have all been my greatest support over these many years. You have helped me stay grounded when my mind would take flight, guided me when I was lost, and encouraged me when I was near defeat. I love you all! It is like you said dad, "The end is near, hang in there kiddo!"

NOMENCLATURE

χ^2	Chi – Squared
α	Internal Consistency (Alpha Coefficient)
$\Delta\chi^2$	Change in Chi – Squared
$\Delta\chi^2_{diff}$	Change in Chi – Squared Differential Test
Z	Z – Score
y*	Latent Continuous Response Variable
ϕ	Phi-coefficient

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

CHAPTER I

INTRODUCTION

The Minnesota Multiphasic Personality Inventory—Adolescent—Restructured Form (MMPI-A-RF; Archer et al., 2016) is an empirically-based 241 true-false item self-report measure of psychopathology and personality for adolescents. Development of the measure was informed by the Minnesota Multiphasic Personality Inventory—2—Restructured Form (MMPI-2-RF; Ben-Porath & Tellegen, 2008/2011). The MMPI-A-RF was developed from a sample of 15,128 adolescents (9,286 boys and 5,842 girls) derived from inpatient, outpatient, correctional, and school samples, and it includes separate samples for validation (Archer et al., 2016). The MMPI-A-RF is composed of 48 scales: six Validity, three Higher-Order, nine Restructured Clinical, twenty-five Specific Problems, and five PSY-5 Scales. The Specific Problems Scale set includes five Somatic/Cognitive Scales, nine Internalizing Scales, six Externalizing Scales, and five Interpersonal Scales.

When the original Minnesota Multiphasic Personality Inventory (MMPI; Hathaway & McKinley, 1943) was developed, raw scores were converted to T-score values through a linear transformation procedure which was computed separately by gender. That is, there were separate gendered norms due to the differences in raw scores between men and women (Graham, 2012). The use of separate norms was also used for the Minnesota Multiphasic Personality Inventory—2 (MMPI-2; Butcher et al., 1989/2001) and the Minnesota Multiphasic Personality Inventory—Adolescent (MMPI-A; Butcher et al., 1992). However, initiated by the Civil Rights Act of 1991, non-gendered norms were developed for the MMPI-2, and it was encouraged to use the non-gendered norms when the gendered norms were prohibited (Graham, 2012). Ben-Porath and Forbey (2003; as cited in Graham, 2012) examined the gendered and non-gendered scores for the

MMPI-2 scales and concluded that the T-scores were similar. Similarly, the scales on the MMPI-2-RF (Ben-Porath & Tellegen, 2008/2011) show a similar pattern to Ben-Porath and Forbey (2003), where none of the gender differences of the gendered norms reached or exceeded a clinically significant five T-score points with the exceptions of Behavioral/Externalizing Dysfunction (BXD) and Discontraint-Revised (DISC-r) where men scored higher than women, and Multiple Specific Fears (MSF) where women scored higher than men (Ben-Porath, 2012).

For the MMPI, Hathaway and Monachesi (1963) identified gender differences in item response patterns and frequency of code-types where boys endorsed 63 items with at least 25% or more endorsement rate compared to girls. An endorsement rate difference of nearly 10% was also seen in nearly 100 items in the MMPI-A normative sample when the data was analyzed by gender (Archer, 2017). Code types, which indicate the highest two point or three point endorsed Clinical Scales in a profile (Graham, 2012), have also demonstrated a differential pattern for boys and girls. For code type frequencies of the MMPI-A, it was found that the 4-9/9-4 and the 4-6/6-4 code types were more frequent in boys and the 1-3/3-1 and 2-3/3-2 code types were more frequent in girls (Archer, 2005).

In a few recent studies, gender differences were examined in the MMPI-2 and MMPI-A across cultural samples (American and Korean). Han et al. (2013) examined item endorsement frequency differences between genders in the MMPI-2 and MMPI-A items and content domains across cultures and found American adults had statistically significant higher gender-discriminating items compared to Korean adults. While not statistically significant, they also found American adolescents had a greater proportion of gender-discriminating items compared to Korean adolescents.

With a similar population, Wang et al. (2020) used multiple-group confirmatory analysis (MGCFA) to examine the measurement invariance of the MMPI-2-RF Externalizing Specific Problem Scales across American and Korean adult normative and clinical samples and found partial scalar invariance with some gender noninvariant items for the American clinical and normative samples.

The research examining the MMPI-A-RF and personality is limited. This is partially due to the very nature of adolescent development making it difficult to map personality characteristics. Additionally, there is limited published research that examines measurement invariance of the MMPI-A-RF. In Park's (2018) dissertation, he examined the measurement invariance of the Externalizing Scales of the MMPI-A-RF between Korean and American populations, and the results indicate a partial factorial invariance in four of the six Externalizing Scales.

In psychometric research, the utilization of MGCFA is advantageous as it allows the investigators more modeling flexibility to examine invariance in factor structure, factor loadings, thresholds, residuals, and latent means (Brown, 2015). Though some research explores comparisons using an analysis of variance (ANOVA), a confirmatory factor analysis (CFA) is far more superior as it does not assume an observed score reflects the latent construct in all groups, but rather compares them (Brown, 2015). Furthermore, the advantage of CFA over correlational and multiple regression analyses is that a CFA adjusts for measurement error when estimating relationships among variables (Brown, 2015).

The present study is the first to examine the measurement invariance of the MMPI-A-RF Externalizing and Internalizing Specific Problem Scales across gender using a multiple-group confirmatory factor analysis. Although gender is fluid and seen as a spectrum (genderqueer,

gender nonbinary, transgender, etc.) opposed to strictly binary (male/female; Altinay, 2020), when individuals complete the MMPI-A-RF, they indicate one of the two options (male/female). As such, for the purposes of this research, reference to gender will only be in reference to how the adolescent identified when they took the assessment (i.e., male/female).

The examination of measurement invariance is important as it would support construct comparability and validity between groups. Though performance differences across genders on the MMPI-A-RF would not indicate the invalidity of the instrument, it would instigate questions about the source of the differences. For example, are the differences due to the adolescent boys and adolescent girls interpreting the items differently, or are the items subjected to bias where they are aligned with societal norms for individuals who identify as male or as female?

To identify where the source of these differences arise is essential. If the differences are a result of the assessment, then there may be measurement bias. If the differences are a result of true group differences, then interpretation of scores needs to take into consideration the distinction. Given that the MMPI-A-RF uses nongendered norms, a multigroup measurement invariance strategy should be utilized to assess the invariance of the assessment's factors with the expectation that they perform identically between male and female groups.

CHAPTER II

LITERATURE REVIEW

MMPI History

Original MMPI

The original Minnesota Multiphasic Personality Inventory (MMPI) was first published in 1943. The authors, Starke Hathaway and J. Charnley McKinley, originally developed the measure because they wanted to have a psychometric instrument that could generate differential diagnoses (Ben-Porath & Archer, 2014; Graham, 2012; Groth-Marnat, 2009). They established their scale construction based on an empirical keying approach reliant on contemporaneous theories of psychopathology. This included the descriptive Kraepelinian nosology, existing surveys of psychiatric symptoms, and their own clinical experiences (Ben-Porath & Archer, 2014). From this, Hathaway and McKinley created an item pool based on the prevalent understandings of psychopathology and the psychometric knowledge of the time, with hopes it would continue to be an evolving instrument (Ben-Porath, 2012).

The construction of the MMPI thus began with an item pool of over 1000 statements from which Hathaway and McKinley selected 504 items. Their next step included the obtainment of different groups of "normal" and psychiatric patients. Of note, the "normal" reference group consisted of personal friends and relatives of patients at the University of Minnesota hospitals who were willing to complete the inventory. This reference group also consisted of recent high school graduates, work progress administration workers, and medical patients (Groth-Marnat, 2009). The second major reference group (clinical sample) included the patients at the psychiatric hospital, which in turn included all major psychiatric categories that were clinically diagnosed at that time. However, if a patient was diagnosed with more than one diagnosis or if

there was any doubt of the patient's diagnosis, that patient was excluded from the study (Graham, 2012).

After administration, item analyses were conducted with the original 504 test items to identify the items that significantly differentiated between the "normal" and clinical groups. Items identified through this procedure were then included in the MMPI scale for that clinical group (Graham, 2012; Groth-Marnat, 2009). Finally, an attempt was made to cross-validate the scales by selecting a new group of "normal" and comparing their responses with a different group of clinical patients. The remaining items that were statistically significantly different between the two groups were then selected for the scales (Groth-Marnat, 2009).

Shortly after the MMPI was put into clinical use, it was made evident that the instrument was not working as originally intended (Ben-Porath & Archer, 2014; Graham, 2012; Groth-Marnat, 2009). Rather than providing the user with a distinct diagnosis, the Clinical Scale profiles yielded multiple and at times contradictory patterns of elevation. However, patterns in code-types and correlates had emerged which prompted a shift from diagnosis to score patterns on the test (Ben-Porath & Archer, 2014).

MMPI-2

Over the years, many criticisms, an abundance of research, and changing times prompted the restandardization project for the MMPI. From the original MMPI to its revision (MMPI-2 in 1989), it was decided to maintain continuity of the Clinical Scales to ensure the relevancy of the large research base that had accumulated since the test's publication (Ben-Porath, 2012; Graham, 2012). Aims for the revised version were summarized into two goals: (1) improve the test and (2) maintain as much continuity as possible (Ben-Porath, 2012; Graham, 2012). Ways to improve the test included goals such as updating the normative sample, replacing nonworking original

MMPI items with newer ones to assess then-contemporary issues, rewriting awkwardly phrased or problematic items, and generating new items to expand the content dimensions of the item pool (Ben-Porath, 2012; Ben-Porath & Archer, 2014; Graham, 2012).

For the new norms, census data from 1980 was used to guide the revision. The project obtained approximately 2,900 participants to complete the test and of those, 2,600 (1,462 women and 1,138 men) had valid and complete protocols (Ben-Porath, 2012). The normative data racial composition for men included: Caucasian, 82%; African American, 11%; Hispanic, 3%; Native American, 3%; and Asian American, less than 1%. For women: Caucasian, 81%; African American, 13%; Hispanic, 3%; Native American, 3%; and Asian American, 1% (Butcher et al., 2001; Graham, 2012). Participants' age ranged from 18-85 years old and education from less than high school graduate to post-graduate (Butcher et al., 2001). This new normative sample was a greater representation of the general population, though higher educational levels were overrepresented (Graham, 2012). However, research indicated a negligible relationship between scores on the Validity and Clinical Scales compared to educational level of the MMPI-2 normative sample (Long et al., 1994).

From the revision project, 567 items were selected for inclusion in the MMPI-2 test booklet; 372 of the 383 items scored on the 13 basic Validity and Clinical Scales of the original MMPI were retained; 11 items were deleted; 64 of the 82 reworded items were included on the MMPI-2 (Ben-Porath, 2012). Validity and Clinical Scales of the MMPI-2 were nearly identical to those of the MMPI. A new way of calculating MMPI-2 standard scores, new Validity Scales, and the MMPI-2 Content Scales were then obtained (Ben-Porath, 2012).

To address issues related to Clinical Scale heterogeneity, the pervasive influence of a general distress factor, and item overlap, Tellegen et al. (2003) developed the Restructured

Clinical (RC) Scales. These RC Scales were designed to address the heterogeneity problem of the MMPI-2 and to facilitate access to clinically significant information. Construction of the RC Scales included a four-step process. The first step was to develop a measure of Demoralization, the common non-specific factor that contributes shared variance to all of the Clinical Scales and inflates correlations between measures. The second step was to conduct separate item principal component analyses of each of the original Clinical Scales combined with the Demoralization items. The third step was the construction of a set of seed scales representing the 12 identified Clinical Scale components from step two. The fourth step was to develop the nine final RC Scales representing Demoralization and those eight Clinical Scales that represent or are related to major recognized psychopathologies: Hs, D, Hy, Pd, Pa, Pt, Sc, and Ma (Ben-Porath & Tellegen, 2008/2011).

MMPI-2-RF

This development of the RC Scales was the first step in developing the MMPI-2-RF, which is a revised, 338-item version of the MMPI-2. "The overall objective of this revision was to represent the clinically significant substance of the MMPI-2 item pool with a comprehensive set of psychometrically adequate measures" (Tellegen & Ben-Porath, 2008/2011, p. 1). Building upon the statistical techniques that produced the RC Scales, the test developers conducted factor analyses, created seed scales, and added items to the MMPI-2 item pool (Ben-Porath, 2012). The resulting measure was theory-based and empirically informed with scales that demonstrated strong psychometric properties. Of note, the development of the MMPI-2-RF was intended to provide a valuable alternative to the MMPI-2, not to replace it (Graham, 2012). The resulting MMPI-2-RF is a concise measure with 338 items consisting of a total 51 scales: 9 Validity Scales and 42 Substantive Scales.

MMPI-3

The Minnesota Multiphasic Personality Inventory—3 (MMPI-3) was developed, in part, to negate potential issues observed in previous iterations (MMPI-2/MMPI-2-RF) of the measure by expanding the item pool, enhancing content, and updating norms (Ben-Porath & Tellegen, 2020). In the process of updating the MMPI-2-RF, the developers examined response format and whether a True/False format should be maintained or whether a gradated response format should be adapted. Given that the validity was not altered, the True/False format was maintained (Pearson Assessments, Ben-Porath, 2021). The test developers updated 24 items and created five new scales: Combined Response Inconsistency (CRIN), Eating Concerns (EAT), Compulsivity (CMP), Impulsivity (IMP), and Self-Importance (SFI). In addition to adding scales, three were modified: Anxiety (AXY) was modified to Anxiety Related Experiences (ARX), Stress/Worry (STW) was modified to two separate scales Stress (STR) and Worry (WRY), and Interpersonal Passivity (IPP) was modified to Dominance (DOM). Finally, many scales were dropped: Gastrointestinal Complaints (GIC), Head Pain Complaints (HPC), Multiple Specific Fears (MSF), Aesthetic/Literary Interests (AES), and Mechanical/Physical Interests (MEC; Ben-Porath & Tellegen, 2020). After an extensive process, the MMPI-3 was developed in 2020 with 335 items.

A primary goal of the MMPI-3 was to update the normative sample given that the MMPI-2 norms were collected during the 1980s. Over the course of the 35 years, much of the population had become more diverse. For example, since the norms of the 1980s the internet and social media had become more prevalent in day-to-day lives. Therefore, it was important to update the norms to represent the population more accurately. Consequently, the English-speaking norms were updated and were compared to the 2020 projected census data showing a

slight underrepresentation of the Hispanic population (Ben-Porath & Tellegen, 2020). However, the normative sample is believed to be ethnically consistent with the projected 2020 census. The MMPI-3 normative sample does, however, underrepresent individuals age 60+ and individuals with less than a high school education. It also slightly overrepresents individuals with college or graduate education. Overall, however, all groups are an improvement over the normative sample used with the MMPI-2/MMPI-2-RF.

MMPI-A

Though the MMPI was intended for the assessment of adults, it was standardized on a normative sample of individuals ages 16 and older. Therefore, the original MMPI was also used with adolescents. In perhaps one of the first studies with the MMPI and adolescents, Dora Capwell (1945) identified each scale, with the exception of the Hysteria (Hy) Scale, had clear differences between delinquent and nondelinquent adolescent girls with the greatest statistically significant difference in scores on the Psychopathic Deviate (Pd) Scale.

In the 1940s and 1950s, Hathaway and Monachesi began collecting MMPI data of ninth graders from Minnesota communities in an attempt to establish relationships between MMPI findings and delinquent behaviors (Archer, 2005). In their attempt, Hathaway and Monachesi had a combined sample of nearly 15,000 adolescents which they used for their book, *Adolescent Personality and Behavior: MMPI Patterns of Normal, Delinquent, Dropout, and Other Outcomes.* From their data, they concluded the probability of an adolescent engaging in antisocial behaviors would increase if they had elevations on the Psychopathic Deviate (Pd), Schizophrenia (Sc), and Mania (Ma) Scales, or what they termed “Excitatory Scales”. Additionally, on what Hathaway and Monachesi termed “Inhibitory” Scales, elevations in Depression (2), Masculinity-Femininity (5), and Social Introversion-Extroversion (0) decreased

the probability of antisocial behavior (Archer, 2005). Moreover, the research conducted with this data established that the MMPI could provide information on adolescent behavior and depicted how adolescents endorsed test items differently than adults (Archer, 2005).

Recognizing that the MMPI produced different scale elevations when used with adolescents compared to adults, Marks and Briggs developed more age-appropriate adolescent norms and first published them in Dahlstrom et al. (1972). Gathering responses from 720 of Hathaway and Monachesi's state-wide sample and combining them with the 1,046 adolescent sample obtained from six other states, Marks and Briggs developed adolescent norms for ages 17, 16, 15, and 14 and under (Archer et al., 2016; Marks & Briggs, 1972). The new adolescent norms, set forth by the Marks and Briggs approach, did not include the K-correction procedure that was employed with the adult norms of the original MMPI, as the K-correction reduced rather than increased the correlations of the external criterion and adolescent MMPI scale scores (Alperin et al., 1996; Archer, 1987; Archer, 2005; Archer et al., 2016).

In the 1970s, Marks et al. (1974) summarized their findings to produce the first personality correlates for a set of 29 MMPI code types (highest elevated T-score patterns of Clinical Scales) based on roughly 1,250 adolescents who received psychiatric services between 1965 and 1973 (Archer, 2005; Archer et al., 2016). This line of research became the first descriptive statements necessary to interpret adolescent code-type patterns based solely on adolescent data.

Despite the unofficial norms developed for using the MMPI in evaluating adolescents, rising concerns instigated the University of Minnesota Press to institute the MMPI Adolescent Project with the goal to create a standardized MMPI assessment specifically for adolescents (Archer et al., 2002; Archer et al., 2016). Most notably, the adolescent version was needed to

address the concerns of the appropriateness and applicability of using the original MMPI with the adolescent population. Research indicated that using the adult MMPI norms would over-pathologize adolescents (Archer, 1984; Klinge et al., 1978). However, some studies suggested that the unofficial adolescent norms under-pathologized adolescents in clinical settings (Archer, 1984; Klinge & Strauss, 1976). The project also sought to shorten the number of test items, update and change the outdated or inappropriate language, include items and scales that relate to adolescent experiences, adolescent development, as well as adolescent psychopathology, and finally, to establish age-appropriate norms (Archer et al., 2002; Archer, 2005). In addition to these goals, the developers wished to maintain compatibility with the MMPI Validity and Clinical Scales, thereby retaining the criterion-keying method utilized by Hathaway and McKinley (Archer et al., 2016). Thus, in 1992, the MMPI-A was published for adolescents ages 14 through 18 (Butcher et al., 1992).

The MMPI-A normative sample consisted of 805 boys and 815 girls between the ages of 14 to 18 from eight states (Archer, 2005). The ethnic origin of the adolescents in the normative sample were a reasonable match against the U.S. Census at the time; however, the parents of the adolescents used in the normative sample had higher educational levels compared to the 1980 U.S. Census data (Archer, 2005). Similar to the MMPI, there was one set of norms for boys and another for girls.

The MMPI-A is very similar to the original MMPI and MMPI-2, which means it also carries many of the strengths and limitations of the original instrument (Archer et al., 2016). The MMPI-A consisted of 478 items, which were a combination of the original Standard Scales, revised and reworded items, and new item content that is relevant to adolescent concerns. The new instrument included the original ten Clinical Scales (Hypochondriasis, Depression, Hysteria,

Psychopathic Deviate, Masculinity-Femininity, Paranoia, Psychasthenia, Schizophrenia, Hypomania, and Social Introversion) and three of the Validity Scales (L, F, and K) which fulfilled one of the goals of maintaining continuity with the original MMPI and MMPI-2. However, it also included four new Validity Scales (F_1, F_2, VRIN, and TRIN), 15 Content Scales (Anxiety, Obsessiveness, Depression, Health Concerns, Alienation, Bizarre Mentation, Anger, Cynicism, Conduct Problems, Low Self-Esteem, Low Aspiration, Social Discomfort, Family Problems, School Problems, and Negative Treatment Indicators), and six Supplementary Scales (MacAndrew Alcoholism Scale- Revised, Alcohol/Drug Problem Acknowledgment, Alcohol/Drug Problem Proneness, Immaturity, Anxiety, and Repression).

In comparison to the adolescent norms developed for the MMPI, the MMPI-A norms revealed statistically significant lower T-scores for most of the Clinical Scales (Graham, 2012). Additionally, research indicated the difference in scores for adolescent psychiatric patients was greater than the five T-score point difference that was used as a criterion (Janus et al., 1996). Having generally lower T-scores on the MMPI-A, as compared to the MMPI adolescent norms, Butcher et al. (1992) recommended that scores of 65 or greater on the Clinical Scales be considered clinically significant while scores between 60 – 64 be interpreted as high scores.

Though the MMPI-A became the most popular objective, self-report assessment of personality for adolescents (Archer & Newsom, 2000), it had limitations. First, the MMPI-A possibly under-pathologized adolescents who presented with clinical difficulties (Hilts & Moore, 2003). Additionally, the MMPI-A—despite being a valid and reliable instrument—faced psychometric limitations due to the criterion-keying method utilized in the MMPI. These limitations include the inter-related problems of multidimensionality, content heterogeneity, and extensive item overlap between the scales which resulted in excessive intercorrelations and

limited discriminant validity (Archer et al., 2016). Furthermore, the length of the MMPI-A (478 items), though shorter than the original MMPI (566 items), was still a considerable length for the concentration and attention span of some adolescents (Archer, 2005; Archer, 2017).

MMPI-A-RF

The MMPI-A-RF development project arose in late 2007 with the goals of addressing the psychometric limitations of the MMPI-A (Archer & Handel, 2019). As described in Archer et al. (2016), to address the psychometric limitation of heterogeneity, the developers built upon the approach used to develop the MMPI-2-RF:

1) Develop a measure of demoralization for adolescents;
2) Using exploratory factor analyses, identify the distinct components of the Clinical Scales, separate from the demoralization factor;
3) Develop additional Substantive Scales that address other areas represented in the MMPI-A item pool;
4) Develop MMPI-A-RF Validity Scales for over-reporting, under-reporting, and non-content-based responding; and
5) Revise the PSY-5 Scales using the item pool from the MMPI-A-RF

The developers sought to shorten the length of the MMPI-A measure with the goal of having an instrument with roughly 250 items. They also wanted to develop an adolescent self-report measure comparable to the MMPI-2-RF but adapted to also include components related specifically to adolescent psychopathology (Archer & Handel, 2019). In doing so, the clinician could transition between the MMPI-A-RF and the MMPI-2-RF with greater ease. Although the MMPI-A-RF and the MMPI-2-RF share similar scale names and measure similar constructs, they do not contain the exact same items (Pearson Assessment US, Handel, 2021). In addition to

the goals mentioned above, the test developers also wanted to link the measure to more contemporary models of psychopathology and personality.

The MMPI-A-RF development sample consisted of archival MMPI-A data from Pearson Assessments. It included 15,128 adolescents (9,286 boys and 5,842 girls) from inpatient (*n* = 419), outpatient (*n* = 11,699), correctional (*n* = 1,756), and school settings (*n* = 1,254) with a mean age of 15.61 (Archer et al., 2016). Due to developmental factors, the sample was further subdivided into four developmental subsamples by age: younger boys (14- 15), younger girls (14 – 15), older boys (16 – 18), and older girls (16 – 18) (Archer et al., 2016).

After multiple exploratory factor analyses to identify demoralization as a separate construct and multiple correlational analyses to identify additional items with optimal convergent and discriminant correlations, the Restructured Clinical Scales were developed (Handel, 2016). The next step was to develop additional Substantive Scales (Specific Problem Scales) to address adolescent problem areas in the MMPI-A item pool that were not clearly addressed by the RC Scales (Archer, 2016). Following a similar pattern to the creation and identification of the RC Scales, the test developers identified 25 Specific Problem Scales that do not contain overlapping items and which have adequate standard errors of measurement; five of the scales are unique to the MMPI-A-RF (Handel, 2016). The Personality Psychopathology Five (PSY-5) Scales were developed by McNulty and Harkness based on their five-factor personality model and using similar methodology that Harkness et al. (1995) and McNulty et al. (1997) used when they developed the PSY-5 Scales for the MMPI-2 and MMPI-A (Archer, 2016).

Upon completion of the development process, the MMPI-A-RF resulted in a 241-item measurement with 48 scales: six Validity, three Higher-Order, nine Restructured Clinical, 25 Specific Problems, and five PSY-5 Scales.

MMPI-A-RF Norms

The normative sample consists of a subset of the MMPI-A normative sample and includes 1,610 adolescents (805 boys and 805 girls) ages 14 – 18, inclusive, with a mean age of 15.56 (*SD* = 1.18; Archer et al., 2016). The age percentages are as follows: 14-years-old (22.7%), 15-years-old (27.2%), 16-years-old (26.5%), 17-years-old (18.1%), and 18-years-old (5.4%). The ethnicity was composed of 1,229 Whites (76.3%), 199 Blacks (12.4%), 46 Asians (2.9%), 46 Native Americans (2.9%), 33 Hispanics (2.0%), 41 Other (2.5%), and 16 who did not report ethnicity (1.0%; Archer et al., 2016).

Scales for Analyses

Internalizing Scales

The MMPI-A-RF contains nine Internalizing Scales which measure aspects of three RC Scales: Demoralization (RCd), Low Positive Emotions (RC2), and Dysfunctional Negative Emotions (RC7). Demoralization and Low Positive Emotionality are assessed by the Helplessness/Hopelessness (HLP), Self-Doubt (SFD), and Inefficacy (NFC) Scales (Archer et al., 2016). Dysfunctional Negative Emotions is assessed by the Obsessions/Compulsions (OCS), Stress/Worry (STW), Anxiety (AXY), Anger Proneness (ANP), Behavior Restricting Fears (BRF), and Specific Fears (SPF). Due to the development of the MMPI-A-RF, the Internalizing Scales can also be interpreted if the affiliated RC Scale is not elevated.

Outlined in the MMPI-A-RF manual (Archer et al., 2016), the Demoralization (RCd) Scale contains eighteen items that assess for low morale and unhappiness. A lower RCd score ($T \leq 40$) indicates a higher level of morale and life satisfaction while elevated RCd scores ($T \geq 60$) is associated with feelings of general unhappiness, hopelessness/helplessness, lack of self-confidence, and an inability to effectively cope with difficulties. The Low Positive Emotions

(RC2) Scale is a ten-item scale that assesses an individual's lack of positive emotional experiences. A lower RC2 score ($T \leq 40$) indicates a higher level of psychological well-being, positive emotions, and social engagement, while elevated RC2 scores ($T \geq 60$) are associated with feeling socially isolated, ineffective, and unsuccessful (Archer et al., 2016). The Dysfunctional Negative Emotions (RC7) Scale is an eleven-item scale that assesses various negative emotional experiences such as anxiety, irritability, apprehensiveness, embarrassment, and impatience. A lower RC7 score ($T \leq 40$) indicates a lower degree of negative emotional experiences while elevated RC7 scores ($T \geq 60$) are associated with an increased risk for anxiety-related forms of psychopathology (Archer et al., 2016).

The Helplessness/Hopelessness (HLP) Scale consists of ten-items that describe the individual as hopeless and pessimistic, or that they are unable to succeed in life. Low scores indicate a low level of hopelessness or helplessness (Archer et al., 2016). Elevated scores are associated with an increase in depression, low self-esteem, suicidal ideation, and other feelings of hopelessness. The Self-Doubt (SFD) Scale consists of five items describing lack of self-confidence, low self-esteem, and feelings of uselessness. Low scores indicate a low level of self-doubt. Elevated scores can be associated with inferiority feelings, self-doubt, and self-disparagement (Archer et al., 2016). The Inefficacy (NFC) Scale consists of four items that describe feeling incapable of dealing with difficult situations. Elevated scores are associated with being indecisive and ineffective in coping with difficult situations (Archer et al., 2016). The Obsessions/Compulsions (OCS) Scale consists of four items that describe obsessive and compulsive behaviors. Low scores indicate no reports of obsessions or compulsions. Elevated scores are associated with rumination, feeling anxious, compulsiveness, and obsessiveness. The Stress/Worry (STW) Scale consists of seven items that describe a preoccupation with worries,

losing sleep when stressed, and being prone to take things too hard (Archer et al., 2016). Low scores indicate no reports of stress-related symptoms. Elevated scores are associated with feelings of anxiety, difficulties with concentration, and complaining of sleeplessness (Archer et al., 2016).

The Anxiety (AXY) Scale consists of four items describing feelings of uneasiness and dread. Elevated scores are associated with anxious feelings, difficulties concentrating, and many specific fears (Archer et al., 2016). The Anger Proneness (ANP) Scale consists of five items describing a tendency to experience and express anger. Low scores indicate no reporting of anger problems. Elevated scores are associated with interpersonal difficulties and feeling angry and irritable. The Behavior-Restricting Fears (BRF) Scale consists of three items that describe fears that prevent normal activities in or out of the home. Elevated scores are associated with having fears that impede normal daily activities. Finally, the Specific Fears (SPF) Scale consists of four items that describe multiple fears. Low scores indicate few fears reported. Elevated scores are associated with many fears and/or phobias (Archer et al., 2016).

Externalizing Scales

The MMPI-A-RF also contains six Externalizing Scales that measure aspects of two RC Scales: Antisocial Behavior (RC4) and Hypomanic Activation (RC9). Antisocial Behavior (RC4) is assessed by the Negative School Attitudes (NSA), Antisocial Attitudes (ASA), Conduct Problems (CNP), Substance Abuse (SUB), and Negative Peer Influence (NPI) Scales (Archer et al., 2016). Hypomanic Activation (RC9) is assessed by the Aggression (AGG) Scale. The results of the Externalizing Scales can also be used to interpret the Cynicism (RC3) Scale. Due to the development of the MMPI-A-RF, the Externalizing Scales can also be interpreted if the affiliated RC Scale is not elevated (Archer et al., 2016).

As outlined in the MMPI-A-RF manual (Archer et al., 2016), the Antisocial Behavior (RC4) Scale contains twenty items that assess various aspects of disordered and antisocial conduct. A lower RC4 score ($T \leq 40$) indicates a reduced risk of disorderly conduct, while an elevated RC4 score ($T \geq 60$) is associated with difficulties at home or school, issues with substances, and a tendency to affiliate oneself with socially undesirable peer groups. The Hypomanic Activation (RC9) Scale is an eight-item scale that assesses an individual's aggression, impulsivity, need for excitement, and high levels of psychomotor energy. A lower RC9 score ($T \leq 40$) indicates a lower level of activation, while elevated RC9 scores ($T \geq 60$) are associated with risk-taking behaviors, aggressive behaviors, and a history of conduct problems (Archer et al., 2016).

As indicated in the MMPI-A-RF manual (Archer et al., 2016), the Negative School Attitudes (NSA) Scale consists of six items describing attitudes and beliefs of school being unproductive and aversive to attend. Low scores indicate the adolescent reports a favorable attitude of school. Elevated scores are associated with a higher endorsement rate of rule-breaking behavior, dislike of school, thinking school is boring or a waste of time, and school avoidance. The Antisocial Attitudes (ASA) Scale consists of six items describing antisocial beliefs and attitudes, evading rules, and dishonesty. Low scores indicate a prosocial attitude. Elevated scores are associated with oppositional behaviors, fighting, conduct problems, rule breaking, juvenile detention, suspension, and substance use (Archer et al., 2016). The Conduct Problems (CNP) Scale consists of seven items describing a history of conduct problems at school and home. Low scores indicate a history of "good" behavior. Elevated scores are associated with criminal charges, running away from home, stealing, fighting, suspensions, poor academic performance, and other behavioral problems within the school and home environment (Archer et al., 2016).

The Substance Abuse (SUB) Scale consists of four items describing drug and alcohol use. Elevated scores are associated with problematic use of drugs and/or alcohol (Archer et al., 2016). The Negative Peer Influence (NPI) Scale consists of five items describing an association with peers who encourage and support antisocial behaviors. Elevated scores are associated with an affiliation with a societally negative or undesirable peer group that engages in rule-breaking behaviors or is oppositional (Archer et al., 2016). The Aggression (AGG) Scale consists of eight items describing aggressive behaviors and aggressive attitudes. Low scores indicate low levels of aggression. Elevated scores are associated with physically aggressive behaviors, violent behaviors, fighting, or enjoyment from intimidating others (Archer et al., 2016).

Adolescent Development

Adolescent gender development is a process of physical, neurological, cognitive, and emotional growth with significant variations based upon gender, race, and environmental and social influences (Curtis, 2015). Adolescent girls generally experience the onset of puberty between ages 7 and 13, with boys typically experiencing onset between ages of 9 and 13.5 (with the average of the two being 11 years of age; Curtis, 2015). This significant growth and change in appearance varies by age, but generally is completed around age 17 to 19 for girls and around age 20 for boys (Christie & Viner, 2005). Complimentary to the physical growth, sexual maturation is present in adolescent development and can negatively impact an adolescents emotional, social, and psychopathological development. Boys with early maturation are likely to be involved in more high-risk behaviors (American Psychiatric Association, 2013), whereas early maturation in girls is associated with a higher risk of depression, substance abuse, eating disorders, and behavioral concerns (Ge et al., 2001). Boys with late maturation are at an

increased risk for depression, increased conflict, and difficulty in peer relationships (Graber et al., 1997).

In addition to physical and sexual maturation, there are also differences in adolescents' neurological and cognitive (Berenbaum et al., 2008), as well as the emotional development. Due in part to societal and environmental norms, boys and girls often differ in the challenges they encounter in their emotional development and in terms of how they address their identity development. Girls tend to have a decreased self-esteem and may not express anger or assertiveness in an adaptive way, whereas boys may have difficulty expressing their internalized emotions.

Another change in adolescent development is the shift from familial relationships to the development of relative independence and involvement with peers. With respect to the latter, adolescents vary in their peer relationships, as boys tend to engage in more activity-based functions, while girls place a greater value on their friendships and social support (American Psychiatric Association, 2013).

Adolescent Psychopathology between Genders

Given the variations in adolescent development, societal and environmental influences, and gender stereotypes, gender-based differences in psychopathology and symptomology have emerged. For example, there have been reports of adolescent boys frequently experiencing more externalizing behaviors (American Psychiatric Association, 2013), such as aggression (Lahey et al., 2000), conduct disorders (Zahn-Waxler et al., 2008), hostility, and hyperactivity (Maras et al., 2003). Given that conduct disorder is comorbid with other externalizing disorders, boys are at higher risk of exhibiting externalizing disorders and exhibit higher rates than do girls (American Psychiatric Association, 2013).

Adolescent girls tend to have the more stereotypical feminine traits and are often described as more nurturant, emotional, passive, and dependent, which can increase symptoms of internalizing disorders, such as anxiety and depression (Kazdin, 2000; Perry & Pauletti, 2011; Zahn-Waxler, et al., 2008). This increase in depression is further elevated during adolescence, as reflected by higher rates of depression in girls than boys (McLaughlin & King, 2015; Zahn-Waxler et al., 2008). Depression is also comorbid with decreased feelings of self-worth and anxiety, which may possibly lead to girls having a higher risk of those internalizing symptoms.

Multiple studies, including Romano et al. (2001), identify adolescent girls as reporting higher rates of internalizing disorders and adolescent boys as reporting higher rates of externalizing disorders. However, are these gender-based differences true differences, and can they be captured in psychological assessments?

Gender in MMPI Instruments

When the original MMPI was developed, raw scores were converted to T-score values through a linear transformation procedure, which was computed separately by gender. That is, there were separate gendered norms due to the differences in raw scores between men and women (Graham, 2012). The use of separate norms was also used for the MMPI-2 and the MMPI-A. However, with the Civil Rights Act of 1991 explicitly prohibiting consideration of race, color, religion, national origin, or sex in employment practice, the use of gendered norms in employment screening became a violation of the prohibition. Therefore, non-gendered norms were developed for the MMPI-2, and it was encouraged to use the non-gendered norms in employment screening or when the gendered norms were prohibited (Graham, 2012). Ben-Porath and Forbey (2003; as cited in Graham, 2012) examined the gendered and non-gendered scores for the MMPI-2 scales and concluded that the T-scores, based on gendered versus non-gendered

norms, were similar. Likewise, the MMPI-2-RF shows consistent results of Ben-Porath and Forbey (2003), where none of the gender differences reached or exceeded five T-score points with the exceptions of Behavioral/Externalizing Dysfunction (BXD) and Discontraint-Revised (DISC-r) where men scored higher than women, and Multiple Specific Fears (MSF), where women scored higher than men (Ben-Porath, 2012).

For the MMPI, Hathaway and Monachesi (1963) identified gender differences in item response patterns and frequency of code-types where boys endorsed 63 items with at least 25% or more endorsement rate compared to girls. An endorsement rate difference of nearly 10% was also seen in nearly 100 items in the MMPI-A normative sample when the data was analyzed by gender (Archer, 2017). For code type frequencies of the MMPI-A, it was found that the 4-9/9-4 and the 4-6/6-4 code types were more frequent in boys and the 1-3/3-1 and 2-3/3-2 code types were more frequent in girls (Archer, 2005).

It is often difficult to untangle the knot of any mean score differences by gender, as it may be complicated to gather whether the difference in scores is a result of gender differences in response style, opposed to true differences in psychopathology (Krishnamurthy, 2016). Furthermore, in using gender-specific norms, true gender differences may be masked because the gender-related variance is eliminated (Handel, 2016; Krishnamurthy, 2016). Therefore, it is preferable to use nongendered norms.

In limited, recent studies, gender differences were examined in the MMPI-2 and MMPI-A across cultural samples (American and Korean). Han et al. (2013) examined gender differences in the MMPI-2 and MMPI-A items and content domains across cultures and found American adults had statistically significant higher gender-discriminating items compared to

Korean adults. While not statistically significant, they also found American adolescents had a greater proportion of gender-discriminating items compared to Korean adolescents.

With a similar population, Wang et al. (2020) used multiple-group CFA to examine the measurement invariance of the MMPI-2-RF Externalizing Specific Problem Scales across American and Korean adult normative and clinical samples and found partial scalar invariance with some gender noninvariant items for the American clinical and normative samples.

Most recently Bryant et al. (2021) examined how individuals who identified as transgender scored on the MMPI-2-RF scales; the results indicate that individuals who identify as transgender and are not in treatment score statistically significantly higher on 31 of the MMPI-2-RF Substantive Scales. Additionally, individuals who identified as transgender and who were not in treatment had higher elevations in scales pertaining to the internalized disorders. Though a difference in mean scores does not necessarily indicate measurement bias, the difference in mean score patterns may suggest the need for an exploration of the MMPI instruments with alternative statistical analyses to determine if there is bias across groups, factor structures, or differential item functioning.

Test Bias

While assessment measures are an important tool in psychology, they are far from perfect. Reliability can be compromised by measurement error and validity can be compromised by responses biases. In turn, this could have many implications for individuals, including those relative to misdiagnosis, incorrect placements, admissions, and employment, to name a few. Bias, as outlined in the *Standards* (AERA et al., 2014), is centered around the context of fairness where there is an underrepresentation or construct-irrelevant components of test scores that may impact the test scores of different groups of individuals, consequently impacting the reliability

and validity of interpretations. Additionally, bias can be a systematic error in the test score that, when applied to different groups, may underestimate or overestimate the construct domain that the test is designed to measure (AERA et al., 2014). If the bias is due to a nominal cultural variable, such as gender, the test can result in cultural bias.

In general, there are two important types of test bias. The first reflects the biases in the meaning of a test. This is referred to as construct bias (also known as internal bias or measurement bias). Construct bias is when the test has different meanings for two groups, and it concerns the relationship of the true score to the observed scores (Furr, 2018). The second reflects the biases in the use of the test. This is referred to as predictive bias (also known as external bias or differential validity). Predictive bias is when a test's use has different implications for two groups, and it further concerns the relationship between scores on two different tests (Furr, 2018). It is important to note that these two types of bias are independent. Indeed, where one test may have strong construct bias, it may lack in predictive bias, or vice versa.

In detecting test bias, early research examined score differences between groups. However, group differences in mean scores do not always have a direct implication for test bias (Furr, 2018; Reynolds et al., 2021). The differences may in fact be an estimate of the true group difference.

Construct Bias

As mentioned above, construct bias concerns the relation to the meaning of test scores and occurs when the measurement has a different meaning between two groups (Furr, 2018). If the relationship of the observed score to the true score is systematically different for two groups, then one may conclude that the test is biased. For example, say a group of students took a math

test and on average, the boys scored higher than the girls. The test may overestimate the true math ability of boys or underestimate the true math ability of girls. Given both groups have the same math ability, it is probable that the math test is biased and yields greater observed scores for boys. To help evaluate this internal type of bias, individual items on a test are often examined. An item on a test is said to be biased under two conditions: 1) if individuals in different groups respond differently to the item and 2) the different responses are not related to group differences in the construct measured by the test (Furr, 2018).

There are five primary methods used to aid in the detection of construct bias: reliability, rank order, item discrimination index, factor analysis, and differential item functioning analyses. Estimating reliability for each group through internal consistency provides further insight into the internal structure. When group differences in reliability are present, it would suggest the test is more reliable in one group compared to the other group(s). While early methods utilized coefficient alpha, confirmatory factor analysis may utilize indices such as omega (Raykov, 2002). Rank order may also provide a way to estimate construct bias. If the rank of items' difficulty differs across groups, construct bias may exist. Construct bias may be examined by separately computing the item discrimination indices between two groups. After computing the discrimination index for two groups, they can be compared. If the index values are approximately equal, the item is considered to reflect the same construct for both groups. If the values are not approximately equal, the item would not equally reflect the construct in the same way for both groups. Factor analysis is another method used to estimate construct bias. While it can be examined using exploratory factor analysis, construct bias is more quantifiable through the confirmatory factor analysis. Further detail about confirmatory factor analysis and measurement invariance will follow. It is important to note that measurement invariance is

currently the most common sophisticated approach in examining construct bias (Furr, 2018). The final most common method is the differential item functioning analysis which is developed within the context of Item Response Theory (Furr, 2018). Item-Response Theory assumes that trait levels are directly pulled from test data and that the true scores for psychological attributes are being measured. Therefore, if there are estimates of trait levels for two groups, the responses can be matched to determine if they are similar for both groups.

Prediction Bias

The second type of test bias, predictive bias, reflects the biases in the use of the test and occurs when a test's use has different implications for two groups (Furr, 2018). When a test instrument is not capable of predicting outcomes equally for different groups on a given psychological criterion, then the instrument is considered biased. There are two primary forms of prediction bias, intercept and slope bias. Intercept bias involves the direction of the intercept for each group and whether the predictor under- or overestimates the criterion variable of the group differences (Wicherts & Dolan, 2010). Intercept bias is often studied using moderated multiple regression and examining the change in R^2 (Mattern & Patterson, 2013). Slope bias involves examining the slope of the regression line between the criterion and predictor variables of the differing groups (Anastasi & Urbina, 1997; Furr, 2018; Reynolds et al., 2021).

Establishing Measurement Invariance

In test development, it is important to determine how well measurement models can be generalized across groups of individuals or across time. In psychometrics, measurement invariance is applied to determine whether the test items relate to a factor similarly across groups or time and it allows for making valid comparisons across groups (Furr, 2018; Millsap, 2011). A test that lacks invariance is a test in which the internal structure is different across groups. This

would suggest construct bias or that test items are related to each other in a different way across the groups (Brown, 2015; Furr, 2018). If a test does have invariance, it suggests component items are being measured in the same way across the groups and provides evidence against construct bias. Questions of measurement invariance can be addressed via Confirmatory Factor Analysis (CFA) by Multiple Indicators, Multiple Causes (MIMIC) models, or Multiple-Group Confirmatory Factor Analysis (MGCFA; Brown, 2015).

Confirmatory Factor Analysis

Confirmatory Factor Analysis (CFA) is a hypothesis-driven type of structural equation modeling (SEM) that examines the relationship between indicators (test items/test scores) and latent variable structures or factors (Brown, 2015). In addition to having a theory and hypothesis, there are at least three preliminary steps before conducting a CFA (Furr, 2018). The first preliminary step is to clarify the construct and develop the test items. The second preliminary step is the collection of a large sample with recommendations ranging from a minimum of 50 people to 400 people or more (Furr, 2018). Other recommendations are based upon the ratio of respondents to items with recommendations ranging from five respondents per item to 20 or more (Furr, 2018). The third preliminary step is to reverse score any items that are negatively keyed.

After the preliminary steps, a CFA has four additional steps, with the first requiring the researcher to articulate and evaluate the measurement model based upon past evidence and theory. Within the specification of the measurement model, the researcher must specify the number of dimensions, factors, or latent variables that underlie the test items. They must also specify the links between the items and the factors and potential associations between factors (Furr, 2018). Step two of a CFA is the computation step which involves four phases: (1) variance

and covariance of the items, (2) parameters estimates (and inferential tests), (3) implied variance and covariance, (4) indices of model fit.

Step three is the interpretation and reporting of the output. There are multiple statistical issues and a variety of psychometric questions that can be addressed with the output from a CFA. However, the results obtained will influence the outcome of the next steps. It may be that further analyses will be warranted, conclusion of the analyses and reporting the findings, or modification of the hypothesized model and rerunning the analysis (Furr, 2018). When interpreting the output, there are primarily two sets of results that are of interest: fit indices, and parameter estimates and significance tests. Both will be discussed further below. If the fit indices indicate the model fits well, the analyses are complete. However, if the fit indices indicate the model fits poorly, the examiner moves to step four: model modification and reanalysis. Over the course of the CFA process, the parameter estimates are aimed at maximizing the probability that the sample and predicted variance/covariance matrix is not statistically different from one another while the goodness of fit indices are examined to evaluate the fit of the model based on the observed variance and covariance.

CFA Model Parameters

Within a CFA framework, the parameters of the model can be free, constrained, or fixed when they are estimated. Freely estimated parameters allow the researcher to find the optimal values of the parameter that reduces the differences between the observed and predicted variance/covariance matrix. Fixed parameters are when researchers assign specific values. Fixed parameters are often used to provide scaling of latent variables (Brown, 2015). Constrained parameters are when the researcher places other restrictions on the values but does not specify

the exact value of the parameter. The most common constrained parameters are equality constraints where unstandardized parameters are restricted to be equal in value (Brown, 2015).

CFA model parameter estimates are generally completed with unstandardized forms and contain factor loadings, unique variances, and factor variances (Brown 2015). However, completely standardized solutions and partially standardized solutions may also be completed. Furthermore, and if desired, the error covariance (correlated residual or correlated errors) and factor covariance (factor correlation) can be specified in a model. Factor loadings are the regression slopes for predicting the indicators from the latent variable. Unique variance, often referred to as measurement error, is the variance in the indicator that is not accounted for by the latent variables. Factor variance, in an unstandardized solution, expresses the sample variability of the factor.

Given the CFA analyzes the variance-covariance structures, the factor loadings error variances/covariances, and factor variances/covariances are estimated to reproduce the input variance/covariance matrix. CFA models may also include an analysis of mean structures. As seen in a multiple group CFA model, the CFA parameters are expanded to reproduce the observed sample means of the indicator intercept and latent variables within the input variance-covariance matrix and analyzed to ascertain how the groups differ (Brown, 2015).

Goodness-of-fit Indices

The goodness-of-fit indices are the comparison of the implied variance/covariance with the actual variance/covariance and represent how well the measurement models fit or reflect the actual pattern of the responses. If a measurement model has a “good fit,” the hypothesized measurement model adequately reflects the actual pattern of responses, thus supporting the validity of the model. If a measurement model has a “poor fit,” the hypothesized measurement

model does not adequately reflect the actual response pattern, thus not supporting the dimensionality of the measure. There are multiple goodness-of-fit indices that can be used in a CFA. For the purpose of this dissertation project, the most generally used indices are discussed: chi-square statistic (χ^2), standardized root mean square residual (SRMR), root mean square error of approximation (RMSEA), comparative fit index (CFI), and Tucker-Lewis index (TLI; Brown, 2015; Furr, 2018).

The chi-square (χ^2), which statistically indicates the poorness of fit of the model, is the primary and most common fit index. A statistically significant χ^2 indicates the model estimates do not reproduce the variance-covariance matrix and thus, is evidence of poor fit. Conversely, a non-significant χ^2 indicates the model does reproduce the variance-covariance matrix and indicates good fit (Furr; 2018). Therefore, in CFAs, researchers want to find a test with a non-significant χ^2. However, the χ^2 should not be the only test of model fit used because of its limitations. The χ^2 is influenced by the sample size, where a large sample will produce large χ^2 values, leading to statistical significance and poor fit. Given sample size must be large to conduct a CFA—so reliable parameter estimates can be obtained—additional fit indices are reported that do not include a formal test of statistical significance.

The SRMR is a goodness-of-fit statistic that measures the average discrepancy between the observed and the predicted correlations by the model (Kline, 2016). As the SRMR is calculated as the square root of the squared covariance residual, it is a measure of the mean absolute correlation residual. The SRMR ranges in values between 0.0 and 1.0, with 0.0 indicating perfect fit. Therefore, smaller values of SRMR correspond to better model fit (Brown, 2015).

RMSEA is a widely used parsimony correction index based on the χ^2 distribution to report the degree of model misspecification (Desa, 2018). Opposed to the more stringent χ^2 difference test that is exact fit, RMSEA is an error of approximation index. Since RMSEA is computed as a function of sample size and model degrees of freedom it is sensitive to the number of model parameters. An RMSEA value at 0.0 indicates a perfect fit and values closer to 0.0 indicate better fit (Brown, 2015).

The remaining two fit indices are the CFI and TLI. The CFI assesses the fit of the researcher's hypothesized model against the more restricted "null" model (Furr, 2018). The restricted model fixes the covariances to zero, and the indicator variances are not constrained (Brown, 2015). The value of the CFI ranges from 0.0 to 1.0, with values closer to 1.0 indicating a better model fit. The other popular index, the TLI, is a non-normed fit index that imposes a penalty for adding freely estimated parameters. This penalty compensates for the effect of the model complexity (Brown, 2015). The value of the CFI can fall outside of the 0.0 – 1.0 range, but values approaching 1.0 are closer to model fit.

Interpreting goodness-of-fit indices is more of an art, given the complexity of the various aspects of the statistical analyses and the importance of examining the interpretability and strength of the parameter estimates and the relationships the model does not sufficiently reproduce. While no true and fixed cutoff is established for fit indices, there are guidelines. Hu and Bentler (1999) conducted simulation studies using the Maximum Likelihood (ML) estimation, and suggested reasonably good fit is obtained where (1) SRMR is close to .08 or below; (2) RMSEA is close to .06 or below; and (3) CFI and TLI are close to .95 or greater. While these guidelines provide an evaluation of goodness-of-fit indices, researchers would

benefit from a combination of indices to evaluate the fit of CFA models due in part to issues of Type I and Type II error as well as variations of alternative analytic situations (Brown, 2015).

Modification Indices

In some instances, the fit indices indicate the model fits poorly, requiring the researcher to revise the hypothesis. In this instance, model modification and reanalysis is necessary to gain a deeper understanding of the test's dimensionality. To do this, an evaluation of the modification indices is crucial. A modification index indicates particular ways in which the measurement model can be improved to bring the model closer to the factor structure that may truly be underlying the test's items. It is the approximation of the difference in the overall χ^2 between a model with constrained or fixed parameters and a model where the parameters are freely estimated (Brown, 2015). Each modification index represents a parameter of the initial measurement model; thus, it can be calculated for each fixed and constrained parameter. A good-fitting model should produce modification indices that are less than the critical value of 3.84 (Brown, 2015). Therefore, if the modification index is greater than 3.84, freeing the fixed or constrained parameters will improve the model fit. Caution must be taken when making a modification within a CFA. Making modifications in a CFA must only occur one parameter at a time and must be validated and supported by prior research or theory (Furr, 2018).

Multiple-Group CFA

As mentioned prior, measurement invariance allows researchers to make comparisons across groups to determine equivalence of parameters (Furr, 2018). This invariance of a model across groups is simultaneously tested using a multiple-group confirmatory factor analysis (MGCFA) with nested model comparisons (Vandenberg & Lance, 2000). In conducting a MGCFA, researchers can examine the measurement models of more than one group and

compare those groups in terms of the models' parameters, including factor loadings, intercepts, and residual variances (Furr, 2018). In MGCFA, a stepwise comparison is employed where invariance is analyzed beginning with the least restricted solution and subsequently increasing in the restrictive constraints; equal factor loadings → equal intercepts → equal residual variances, etc. (Brown, 2015). The stepwise comparison is often recommended and follows the sequence: (1) test the CFA model separately in each group; (2) Conduct the simultaneous test of equal form; (3) Test the equality of factor loadings; (4) Test the equality of indicator intercepts; (5) Test the equality of indicator residual variances; (6) Test the equality of factor variances; (7) Test the equality of factor covariances; (8) Test the equality of latent means. When testing measurement invariance, steps 1-5 are employed, whereas steps 6 - 8 are tests of population heterogeneity (Brown, 2015).

Configural Invariance

Configural invariance, or equal form, is the first test for invariance and is the least restrictive of the four levels. Configural invariance establishes whether a factor structure is the same across groups, but constraints are placed on any of the parameters in each group. If the model is not consistent with the data, then measurement invariance would not hold true at any level. If configural invariance is achieved, the number of factors and the pattern of factor loadings is the same and it is concluded that the test items likely reflect the same latent variable across groups (Furr, 2018; Kline, 2016).

Weak Factorial Invariance

Weak factorial invariance, or metric invariance, is the next test for invariance and is more robust than configural invariance. Weak factorial invariance is established when the number of factors and the pattern of factor loadings is the same. Additionally, the exact values of the factor

loadings are the same. When weak factorial invariance is met, it is concluded that the test's items likely reflect the same latent variable across groups and the scale scores are on the same measurement metric (Furr, 2018). To test for weak factorial invariance, factor loadings must be constrained to be equal across groups. For example, test item 1 in group A will be constrained to equal the factor loading of test item 1 in group B.

Scalar Invariance

Scalar invariance, strong factorial invariance, or residual invariance is the third test for invariance and is more robust than weak factorial invariance (Furr, 2018). Scalar invariance is established when the number of factors is the same, the pattern of factor loadings is the same, the exact values of the factor loadings are the same, and the item intercepts are the same. Item intercept refers to the average score on an indicator given a true score of zero on the corresponding latent variable. When scalar invariance is met, it is concluded that if two people from different groups have the same level of the latent variable, then on average, they will respond similarly to the item. To test for scalar invariance, factor loadings and intercepts must be constrained to be equal across groups. If the model with the constraints fits worse than the weak factorial model, then the test does not meet scalar invariance (Furr, 2018).

Latent Mean Invariance

Latent mean invariance, or strict factorial invariance, is the most restrictive of the four invariance tests. When strict invariance is established, the number of factors is the same, the pattern of factor loadings is the same, the exact values of the factor loadings are the same, the item intercepts are the same, and the items' unique error variance are the same. To test for this, factor loadings, intercepts, and unique error variances are constrained to be equal across groups (Furr, 2018; Kline, 2016).

Multiple-Group CFA with Categorical Variables

While the aforementioned sequence is often employed for MGCFA, it is most appropriate for continuous variables rather than categorical variables. When variables are continuous, an MGCFA with ML procedure works very well, but it is not as effective when used with categorical variables (Curran et al., 1996). In fact, it was observed in their analyses that studies with categorical variables that used ML estimation had inflated chi-square values and lower parameter coverage for factor loadings. ML may also be a problematic estimation method when applied to binary data as the assumption of normality becomes violated with two response options. Furthermore, treatment of categorical variables as continuous can result in a bias in test statistics, standard errors, and subsequent inferences (Brown, 2015).

Millsap & Yun-Tein (2004) and Lubke & Muthén (2004) recommend the robust weighted least-squares (WLS) or mean- and variance-adjusted weighted least squares (WLSMV) estimation for categorical variables because it requires the measurement parameters to be equal across groups. Research has also indicated that WLSMV performs better than WLS for small sample sizes, which suggests WLSMV is the superior estimator to use for categorical variables (Sass, 2011; Schmitt, 2011). However, when using WLSMV, the change of chi-square cannot be used. Rather, researchers must utilize the DIFFTEST option in Mplus (Sass, 2011). For categorical items, the means and covariances are not enough to determine invariance, therefore another condition is applied, y^* (Desa, 2018). This other condition implies measurement invariance in both the factor model parameters and the threshold parameters (Desa, 2018). Thus, the correlations can be interpreted as an underlying continuous characteristic needed to produce a response for a categorical variable by means of the threshold parameter (Brown, 2015).

Thresholds link binary indicators to their underlying continuous latent variable by marking the point where respondents are likely to switch from a 0 to a 1 on an item. These thresholds, essentially z-scores associated with response probabilities, can be positive or negative. They can be converted to the likelihood of endorsing an item using a z-table for the standard normal distribution (Finney & DiStefano, 2013).

Necessary steps in completing a measurement invariance of a MGCFA with binary categorical variables are slightly different than simply moving through the steps outlined in measurement invariance of MGCFA with continuous variables. Similar to MGCFA with continuous indicators, the first step is to conduct a CFA separately in each group to ensure the measurement model is acceptable for each group. Once an appropriate model is identified, the next step of establishing an equal form model can proceed. However, since the item responses are binary, metric invariance cannot be completed before scalar invariance. Therefore, factor loadings and thresholds are constrained to equality across groups; the factor means are fixed to zero in one group and freely estimated in the other group; and the scale factors are fixed to one in one group and freely estimated in the other groups (Brown, 2015). The DIFFTEST option in Mplus is also utilized to compare it to the equal forms model with a non-significant model fit indicating full scalar invariance. When full scalar invariance is not achieved, partial measurement invariance can be pursued. For categorical variables, equality constraints for factor loadings and thresholds for a given item must be relaxed simultaneously. Additionally, the scale factor for the noninvariant item must also be fixed to one in all groups (Brown, 2015).

CHAPTER III

RATIONALE OF THE PRESENT STUDY

The purpose of the present study was to examine the measurement invariance of the MMPI-A-RF Internalizing and Externalizing Specific Problem Scales in male and female adolescents. Given the differences in adolescent development and cultural influences, it is important to evaluate measurement invariance. If full measurement invariance was found using the multiple-group CFA approach, the differences may reveal sex differences. If noninvariance or partial invariance was found, measurement bias in the MMPI-A-RF may be exhibited between the sexes.

In the present study, measurement invariance of the MMPI-A-RF Internalizing and Externalizing Specific Problem Scales were examined because these shorter scales are more likely to be unidimensional due to their narrower focus, compared to some of the other scales (i.e., RC Scales, Higher-Order Scales). This present study was meant to provide the first assessment of measurement invariance of gender in the MMPI-A/MMPI-A-RF. Future studies should further investigate the measurement invariance of multiple scales of the MMPI-A-RF with multiple different populations, including ethnicity, age, and gender, within setting-specific samples (e.g., medical, forensic, or inpatient).

No published studies have examined measurement invariance between males and females for any of the MMPI-A-RF scales, including the Specific Problem Scales. Therefore, this study makes an important contribution to the literature concerning the MMPI-A-RF. Due to the lack of published findings on the topic of measurement invariance with the MMPI-A-RF scales, the generation of hypotheses on the likely nature and extent of measurement invariance is challenging. However, in a systematic review by Dong and Dumas (2020), measurement

invariance of personality measures between genders was completed for 29 studies and all were supported with a configural invariance model. Of those studies, 25 had metric model invariance supported, and 13 of those studies had further shown full scalar invariance. Furthermore, in a multiple-group CFA conducted to examine the measurement invariance of the MMPI-2-RF Externalizing Scales across an American and Korean normative samples, partial scalar invariance with some gender noninvariant items was found (Wang et al., 2020). Given the results of the Dong and Dumas (2020) review and Wang et al. (2020) study, this study hypothesized that all MMPI-A-RF Internalizing and Externalizing Scales will reach partial measurement invariance.

CHAPTER IV

METHODOLOGY

Procedure

Part of the current study's sample was obtained from a Midwest community outpatient setting by remotely accessing their Electronic Health Record system. Following Eastern Virginia Medical School's Institutional Review Board (IRB) approval (IRB # 21-09-WC-0209), MMPI-A/MMPI-A-RF raw data from January 2014 (MMPI-A)/ May 2016 (MMPI-A-RF) through April 2023 were obtained. MMPI-A, MMPI-A-RF, and data from other measures were coded and entered into a new archival database for future research as well as for analyses for a peer's dissertation project. Two clinical psychology doctoral students employed a split-coding technique to double-enter the data. Datasets were compared, and if any discrepancies were identified, the original medical records were reviewed for accurate coding.

As I proposed to employ MGCFA, the sample size needed to be large. Furr (2018) recommended a CFA sample size range from a minimum of 50 (for simple measurement models) to 400 participants or more. To ensure adequate power and precision of the estimated factor loadings within an MGCFA, a larger sample size ($n > 400$) was needed (Meade & Bauer, 2007). It has also been noted that the sample size requirements when using WLSMV are less restrictive (i.e., 150 – 200 may be sufficient; Brown, 2015). Given that the models required for the analyses are simple models, the current study aimed to have a minimum of 200 protocols for each gender (400 total). Due to the insufficient sample size of the Midwest community outpatient setting dataset, additional data were obtained from Pearson Assessments, which consisted of completed MMPI-A and MMPI-A-RF protocols.

Participants

Midwest Archival Sample

The Midwest private practice sample was coded from archival records. The archival sample consisted of individuals ages 14 – 18 who completed psychological assessments since 2014. Testing was conducted for non-research purposes, and the testing batteries varied by the referral question. Individuals may have been administered testing batteries for ADHD, neurocognitive disorders, general psychological evaluations, diagnosis and treatment planning, or forensic evaluations. To obtain the raw data, investigators were provided remote access to the electronic medical record system. Data coded from the Midwest sample included raw test data (MMPI-A/MMPI-A-RF) and relevant demographic variables extracted from psychological reports. Additional psychological measures were also coded with the goal of using these measures in subsequent research studies. MMPI-A protocols were rescored as MMPI-A-RF protocols. A total of 572 cases were identified in the original database; 198 cases did not include an MMPI-A/MMPI-A-RF protocol and were consequently removed from the dataset. The final Midwest sample data included 374 protocols with 108 MMPI-A protocols (57 boys; 51 girls) and 266 MMPI-A-RF protocols (114 boys; 152 girls).

Pearson Sample

This archival sample was obtained from Pearson Assessments with administration dates between June 1, 2018 and October 31, 2023 (NCS Pearson, 2018 - 2023). The data were requested as MMPI-A-RF and MMPI-A protocols. Due to concerns that there would be an inadequate sample size of MMPI-A-RF protocols for MGCFA, MMPI-A protocols were also obtained. Since the MMPI-A-RF item pool is a subset of the MMPI-A item pool, MMPI-A protocols were rescored as MMPI-A-RF protocols. For the MMPI-A-RF, Pearson could only

provide data from the mail-in scoring service, as records are deleted after several months. MMPI-A data were available for a longer timeframe. MMPI-A data obtained from Pearson included age, gender, setting and MMPI-A item responses. MMPI-A-RF data obtained from Pearson included age, gender, years of education, ethnicity, and MMPI-A-RF item responses. The initial dataset consisted of 2,072 protocols with 1,705 MMPI-A protocols (874 boys; 831 girls) and 367 MMPI-A-RF protocols (202 boys; 165 girls).

Combined Data Samples

Upon the completion of all necessary data preparation (further outlined on page 45), the total sample for the study consisted of 1,622 valid protocols, 811 boys and 811 girls. As research has demonstrated that a large sample ($n > 400$) is necessary for adequate power in a CFA (Meade & Bauer, 2007), this sample size was deemed sufficient. The final total sample of boys ranged from 14 – 18 years old with a mean age of 15.79 years and a standard deviation of 1.22 years. The final total sample of girls ranged from 14-18 years old with a mean age of 15.71 years and a standard deviation of 1.16 years. The mean age of the entire sample was 15.75 years with a standard deviation of 1.19 years. See Table 1 for the distribution of gender, age, MMPI-A/MMPI-A-RF, ethnicity, setting, and years of education for the Midwest sample, Pearson sample, and combined sample.

Table 1. *Distribution of Gender, Age, MMPI-A/MMPI-A-RF, Ethnicity, Setting, and Years of Education for the Midwest Sample, Pearson Clinical Sample, and Combined Samples*

	Midwest	Pearson Clinical	Combined
N	234	1388	1622
Gender			
Boys	106 (45.3%)	705 (50.8%)	811 (50.0%)
Girls	128 (54.7%)	683 (49.2%)	811 (50.0%)
Age			
14	57 (24.4%)	247 (17.8%)	304 (18.7%)
15	61 (26.1%)	334 (24.1%)	395 (24.4%)
16	51 (21.8%)	363 (26.2%)	414 (25.5%)
17	61 (26.1%)	354 (25.5%)	415 (25.6%)
18	4 (1.7%)	90 (6.5%)	94 (5.8%)
Test			
MMPI-A-RF	151 (64.5%)	249 (17.9%)	400 (24.7%)
MMPI-A	83 (35.5%)	1139 (82.1%)	1222 (75.3%)
Ethnicity			
White	150 (64.1%)	79 (5.7%)	229 (14.1%)
Black	19 (8.1%)	9 (.6%)	28 (1.7%)
American Indian	2 (.9%)	-	2 (.1%)
Hispanic	18 (7.7%)	9 (.6%)	27 (1.7%)
Asian	5 (2.1%)	2 (.1%)	7 (.4%)
Other	16 (6.8%)	4 (.3%)	20 (1.2%)
Not Reported/Missing	24 (10.3)	1285 (92.6%)	1309 (80.7%)
Setting			
Outpatient Mental Health Center	234 (100%)	750 (54.0%)	984 (60.6%)
Inpatient Mental Health Center		22 (1.6%)	22 (1.4%)
Correctional		73 (5.3%)	73 (4.5%)
Drug-Alcohol Treatment		1 (.1%)	1 (.1%)
General Medical		1 (.1%)	1 (.1%)
School		53 (3.8%)	53 (3.3%)
Not Reported/Missing		488 (35.2%)	488 (30.0%)
Years of Education			
7		1 (.1%)	1 (.1%)
8		10 (.7%)	10 (.6%)
9		25 (1.8%)	25 (1.5%)
10		27 (1.9%)	27 (1.7%)
11		30 (2.2%)	30 (1.8%)
12		9 (.6%)	9 (.6%)
13		1 (.1%)	1 (.1%)
14		1 (.1%)	1 (.1%)
Not Reported/Missing	234 (100%)	1284 (92.5%)	1518 (93.6%)

Note: Pearson MMPI-A data do not include ethnicity or years of education. Pearson MMPI-A-RF data do not include setting.

Instrument

MMPI-A-RF

The MMPI-A-RF (Archer et al., 2016) is a self-report measure of personality characteristics and psychological functioning of adolescents between the ages of 14 and 18. It comprises 241 true and false items. The measure utilizes linear T-scores for the Validity Scales. Uniform T-scores are used for the Substantive Scales to provide percentile comparability (Tellegen & Ben-Porath, 1992). The T-score cut-off for a clinically significant elevation is a $T \geq 60$. For Internalizing Specific Problem Scales, the test-retest correlations ranged from .24 (BRF) to .73 (SFD) with internal consistency alpha coefficients ranging from .37 (AXY for boys) to .61 (HLP for girls) in the normative sample and from .28 (SPF for inpatient boys) to .80 (HLP for girls in school settings) in the development subsamples (Archer et al., 2016). For the Externalizing Specific Problem Scales, the test-retest correlations ranged from .46 (SUB) to .71 (CNP), with internal consistency alpha coefficients in the normative sample ranging from .29 (NPI for girls) to .62 (CNP for boys) and from .41 (NPI for girls in outpatient and school settings) to .78 (NSA for inpatient boys and SUB for inpatient girls; Archer et al., 2016).

Some of the reliability estimates stated above are rather low. However, this is likely due to the small number of items in some of the scales (e.g., BRF is composed of three items). Adequate reliability in Cronbach's alpha coefficients has generally been regarded as acceptable when alpha is greater than .70. However, Schmitt (1996) noted that while lower reliability estimates attenuate the upper limit of validity, there may be cases where a measure has other desirable characteristics such as meaningful content coverage and reasonable unidimensionality. In these cases, low reliability may not be a major barrier to using the measure (Schmitt, 1996). Additionally, reliability is not a fixed value; it can be impacted by the variance of the sample.

While it can provide an estimate of the measurement error present, it does not determine the effect of measurement error on individual test scores (Harvill, 1991).

The standard error of measurement represents the average size of the error scores, and the larger the standard error of measurement, the less reliable the test scores (AERA et al., 2014; Furr, 2018). The standard error of measurement is generally a more informative index than a reliability or generalizability coefficient because it can provide a more direct measure of the relative or absolute score of the individual test scores (AERA et al., 2014). The standard error of measurements for the Internalizing Scales ranged from 5 to 8 T-score points, and the standard error of measurements for the Externalizing Scales ranged from 5 to 10 T-score points (Archer et al., 2016). As indicated above, many of these standard error of measurements are considered adequate and are within the same range as those that are calculated with other MMPI-A-RF measures with higher reliability estimates (Archer et al., 2016). Finally, an important consideration relates to the context in which the MMPI-A-RF is used. MMPI-A-RF interpretative statements are hypotheses that are supported or refuted with data from other sources (e.g., feedback session information, other test data). Therefore, even if a scale has a relatively large SEM, interpretive statements would always be considered as one piece of information within a broader array of data.

Statistical Analyses

Data Preparation

As noted earlier, to prepare the data for analyses, the MMPI-A protocols were rescored as MMPI-A-RF protocols. Data from both the Midwest sample and the Pearson clinical sample were combined (n = 2,446). Prior to further analyses, data were examined for missing values, invalid protocols, and additional data cleaning errors. First, 218 protocols were removed because

the age was missing or was not between fourteen and eighteen (reducing the total sample size to *n* = 2,228). 71 protocols were subsequently excluded because they had missing responses (*n* = 2,157). Next, following validity criteria set forth by the test developers (i.e., VRIN T-scores > 74, TRIN T-scores > 74, CRIN T-scores > 74, F-r T-scores > 89, L scores > 79, or K scores >74; Archer et al., 2016), 493 protocols were removed due to being invalid (*n* = 1,664). To have an equal number of cases between boys and girls, 42 boy MMPI-A protocols were randomly selected in SPSS and removed from the data set resulting in a final sample of 1,622 protocols (811 boys, 811 girls). Since items on several of the Internalizing and Externalizing Scales are not all keyed in the same direction (i.e., a "false" response is in the keyed direction in some cases), all data were recoded so keyed responses were coded as one and unkeyed response were coded as zero. Finally, MPlus files were created for each Internalizing and Externalizing Specific Problem Scale by creating text (.txt) files from the SPSS files.

Data Analyses

Every Specific Problem Scale is a standalone scale and is not dependent upon other MMPI-A-RF scales for interpretation (i.e., these scales are not subscales). Therefore, individual CFA models with categorical variables (i.e., the individual MMPI-A-RF items for each scale) were evaluated separately for boys and for girls to determine if a one-factor model for each of the Externalizing and Internalizing Specific Problems Scales was acceptable for each group. This process initially resulted in nine separate measurement models for Internalizing Specific Problem Scales (18 total analyses) and six separate measurement models (12 total analyses) for Externalizing Specific Problem Scales. These analyses used the WLSMV estimator in Mplus. In each case, the indicators consisted of the item responses for a given scale, and the latent variable was the construct measured by the MMPI-A-RF scale. For example, Antisocial Attitudes has six

items, so there were six binary indicators for the latent variable "Antisocial Attitudes." Overall, 15 CFA analyses were conducted separately by gender. If the model fit for a given scale was not acceptable, items were examined to determine if error terms could be correlated to improve the model.

This model modification step only occurred if it was supported by both modification indices and a theoretical rationale (in this case, similar item content). If a correlated error term was only necessary for one gender, the term was also correlated in the other gender to maintain model consistency. If the model fit of certain scales could not be improved or modifications were not empirically and theoretically justified, further measurement invariance of those Specific Problem Scales did not proceed. If both groups demonstrated an acceptable model fit, measurement invariance analyses were conducted to determine if the test's items reflected the same latent variables across groups. Both groups were analyzed simultaneously, and model fit was examined to establish configural invariance. If configural invariance was achieved, the models proceeded to be tested for measurement invariance.

Following the recommendations set forth by Brown (2015), measurement invariance was examined by constraining the thresholds and factor loadings to equality and fixing the scale factors to 1.00 in one group and freely estimating the second group. Additionally, the factor mean was fixed to zero in one group and freely estimated in the second group. By constraining the thresholds and factor loadings to equality, the change in model fit between the configural invariance model and the measurement invariance model were examined. If there was a non-significant model fit difference, measurement invariance was established and any difference between the estimated latent means was interpreted.

Partial measurement invariance was examined if full measurement invariance was not established. In reference to Brown (2015), this was done by constraining the factor loadings and thresholds one item at a time while freeing the remaining items. Scale factors for the noninvariant items were fixed at 1.00 in both groups. Noninvariant items were examined by comparing the partial measurement invariance models with the configural invariance model. If the comparison yielded a non-significant model fit difference, then partial scalar invariance was obtained; items that are noninvariant would produce statistically significant model fit differences. Latent means were calculated for scales that reached measurement and partial measurement invariance.

Goodness-of-fit Indices

To evaluate the model fit for each step in the CFA and measurement invariance, RMSEA, CFI, and TLI were examined. Model fit was evaluated using the guidance set forth by Hu and Bentler (1999), where RMSEA values < .08 for acceptable fit, values below .05 showed good model fit, and values equal to or greater than 0.1 were rejected. In examining CFI and TLI, values > .95 were considered to have good fit, and values between .92 - .94 were considered to have an adequate fit. When comparing two models with continuous indicators, a chi-squared difference test is often used. As these analyses were comparing categorical variables, the χ^2 difference test was not utilized; rather the DIFFTEST for WLSMV in Mplus was calculated (Brown, 2015; Sass, 2011).

CHAPTER V

RESULTS

Descriptive Statistics

Descriptive statistics of the combined sample were calculated using SPSS. The endorsement frequencies in the keyed direction as well as chi-square tests and phi-coefficients for the Internalizing Scales items across boys' and girls' samples are presented in Table 2. Given the large number of statistical tests in this manuscript, alpha was at .01 for all analyses. As seen in Table 2, the girls' sample had higher endorsement frequencies for all Internalizing Scales compared to the boys' sample with many items having statistically significant differences ($p \leq .001$). However, many of these effect sizes were rather small (< .20). Table 2 also presents the corrected item-total correlations, internal consistency reliability coefficients, and standard error of measurement for each Internalizing Scale. Alpha coefficients of the Internalizing Scales ranged from .34 (SPF) to .78 (HLP) in the boys' sample and .31 (BRF) to .78 (HLP) in the girls' sample. The majority of the Cronbach's alpha coefficients for each scale were comparable across genders with the greatest difference in the BRF Scale (boys $\alpha = .39$, girls $\alpha = .31$). Many of the scales obtained alpha coefficients greater than .55 in both groups with the exception of the BRF (boys $\alpha = .39$, girls $\alpha = .31$) and SPF (both gender groups $\alpha = .34$) Scales.

Table 2. *Endorsement Frequencies, Chi-Square, Phi, Internal Consistency Coefficients, Standard Error of Measurement, and Corrected Item-Total Correlations for Internalizing Scale Items.*

Scale	Item	Endorsement Frequencies		χ^2	ϕ		Item-Total *r*	
		Boys [a]	Girls [a]				Boys [a]	Girls [a]
ANP								
	54	53.1%	**64.0%**	19.67***	.110		.52	.52
	93	56.0%	**64.2%**	11.54***	.084		.46	.39
	161	57.6%	**69.3%**	23.99***	.122		.44	.44
	165	34.6%	**36.6%**	.69	.021		.44	.43
	229	38.3%	**48.1%**	15.68***	.098		.39	.43
						α	.69	.69
						SEM	6	7
AXY								
	55	17.9%	**32.9%**	48.43***	.173		.43	.45
	71	37.0%	**54.6%**	50.79***	.177		.49	.51
	153	56.1%	**79.2%**	98.48***	.246		.46	.37
	209	27.5%	**45.5%**	56.70***	.187		.37	.36
						α	.65	.64
						SEM	8	8
BRF								
	50	9.7%	**11.0%**	.664	.020		.18	.09
	65	22.7%	**30.0%**	11.06***	.083		.25	.20
	123	27.7%	**52.2%**	100.75***	.249		.26	.25
						α	.39	.31
						SEM	8	9
HLP								
	56	33.8%	**44.3%**	18.72***	.107		.35	.45
	60	42.0%	**66.8%**	100.42***	.249		.43	.45
	62	34.2%	**40.7%**	7.39**	.068		.50	.47
	119	26.9%	**37.0%**	19.07***	.108		.62	.62
	121	50.9%	**60.0%**	13.67***	.092		.46	.45
	162	27.6%	**39.2%**	24.48***	.123		.45	.52
	169	23.1%	**40.0%**	53.62***	.182		.35	.33
	194	18.5%	**32.2%**	40.15***	.157		.50	.50
	228	40.2%	**50.7%**	17.97***	.105		.39	.43
	239	49.4%	**59.3%**	15.90***	.099		.39	.31
						α	.78	.78
						SEM	7	7
NFC								
	51	43.0%	**56.2%**	28.24***	.132		.47	.40
	112	60.0%	**78.9%**	68.06***	.205		.51	.47
	159	40.1%	**63.7%**	91.04***	.237		.42	.43
	224	51.3%	**71.1%**	67.31***	.204		.54	.54
						α	.70	.67
						SEM	7	7

Table 2. Continued.

Scale	Item	Endorsement Frequencies		χ^2	ϕ		Item-Total *r*	
		Boys [a]	Girls [a]				Boys [a]	Girls [a]
OCS								
	15	25.2%	**35.3%**	19.66***	.110		.19	.34
	87	40.6%	**46.0%**	4.86*	.055		.36	.40
	139	36.6%	**48.0%**	21.38***	.115		.35	.36
	221	53.5%	**74.6%**	78.30***	.220		.42	.44
						α	.55	.60
						SEM	8	8
SFD								
	73	53.3%	**74.8%**	82.00***	.225		.57	.55
	79	64.2%	**81.6%**	62.09***	.196		.61	.56
	145	39.5%	**53.4%**	31.65***	.140		.45	.45
	202	64.9%	**86.2%**	99.82***	.248		.48	.44
	234	54.7%	**72.9%**	57.69***	.189		.62	.60
						α	.77	.75
						SEM	6	6
SPF								
	44	22.4%	**26.5%**	3.63	.047		.19	.21
	109	44.9%	**49.7%**	3.76*	.048		.16	.10
	147	53.3%	**54.1%**	.122	.009		.18	.19
	213	25.5%	**31.1%**	6.15*	.062		.20	.22
						α	.34	.34
						SEM	7	8
STW								
	4	49.1%	**70.2%**	74.88***	.215		.37	.38
	48	72.7%	**86.3%**	45.83***	.168		.58	.58
	67	77.1%	**82.5%**	7.40**	.068		.41	.31
	77	42.2%	**58.1%**	40.40***	.158		.29	.35
	129	79.8%	**91.1%**	41.97***	.161		.41	.42
	198	60.3%	**78.7%**	64.55***	.199		.48	.44
	203	70.0%	**85.5%**	55.67***	.185		.40	.39
						α	.71	.69
						SEM	6	6

Note. ANP = Anger Proneness. AXY = Anxiety. BRF = Behavior Restricting Fears. HLP = Helplessness/Hopelessness. NFC = Inefficacy. OCS = Obsessions/Compulsions. SFD = Self-Doubt. SPF = Specific Fears. STW = Stress/Worry. Cronbach's alpha and item endorsements obtained from raw data in the keyed direction. Higher endorsement between the two genders are bolded for ease of identification purposes and do not necessarily represent significant differences. SEM = Standard Error of Measurement.

[a] $n = 811$ for each group.

* ≤ .05. ** ≤ .01. *** ≤ .001.

As seen in Table 2, the standard error of measurements for the Internalizing Scales were highly comparable to data presented in the MMPI-A-RF manual (Archer et al., 2016) and ranged from 6 to 8 T-score points for the boys' sample and 6 to 9 T-score points for the girls' sample with the BRF Scale having the largest at 8 T-score points for the boys' sample and 9 T-score points for the girls' sample. In examining the corrected item-total correlations for the Internalizing Scales, item 50 in the BRF Scale had the lowest item-total correlation (r = .09) for the girls' sample and item 109 in the SPF Scale had the lowest item-total correlation (r = .16) for the boys' sample. A majority of the items had similar item-total correlations with the exceptions of item 15 (OCS), item 67 (STW), and item 56 (HLP). Specific Problem Scale intercorrelations by gender for the Internalizing Scales can be found in Table 3.

Table 3. *Internalizing Specific Problem Scale Correlations by Gender*

	HLP	SFD	NFC	OCS	STW	AXY	ANP	BRF	SPF
HLP		.615**	.586**	.431**	.424**	.471**	.309**	.276**	-.037
SFD	.595**		.570**	.380**	.508**	.411**	.212**	.213**	<.001
NFC	.652**	.548**		.477**	.512**	.460**	.336**	.350**	.058
OCS	.415**	.323**	.388**		.438**	.482**	.365**	.349**	.053
STW	.445**	.506**	.502**	.417**		.580**	.298**	.309**	.111**
AXY	.526**	.454**	.508**	.487**	.553**		.339**	.450**	.149**
ANP	.268**	.211**	.313**	.341**	.215**	.254**		.246**	.029
BRF	.316**	.261**	.337**	.293**	.325**	.434**	.154**		.181**
SPF	.008	.008	.094**	.053	.122**	.120**	.017	.184**	

Note. Boys (N = 811) correlations in the upper shaded diagonal. Girls (N = 811) correlations in the lower diagonal. Raw scale totals were used in the analysis. HLP = Helplessness/Hopelessness. SFD = Self-Doubt. NFC = Inefficacy. STW = Stress/Worry. AXY = Anxiety. ANP = Anger Proneness. BRF = Behavior Restricting Fears. SPF = Specific Fears.
** Correlation statistically significant at 0.01 level.

Given the differences in item endorsement frequencies, not surprisingly, statistically significant differences ($p < .01$) were subsequently observed for mean scale scores with the girls' sample having higher T-scores for all Internalizing Scales (Table 4). Furthermore, Cohen's *d* values were computed for each t-test of the Internalizing Scales. Following Cohen's guidelines ($d = .20, .50$, and $.80$ for small, medium, and large effects; Cohen, 1992), values indicated a medium effect size for the differences between AXY ($d = .57$), NFC ($d = .54$), SFD ($d = .56$), and STW ($d = .59$), and a small effect size for the differences between ANP ($d = .26$), BRF ($d = .40$), HLP ($d = .44$), and OCS ($d = .38$). The effect size for the SPF ($d = .14$) was less than small (i.e., < .20).

Table 4. *Independent Sample T-tests for Internalizing Scales*

Scale	Gender [a]	*M*	*SD*	*t*	*df*	*p*-value	Cohen's d
ANP							
	Boys	50.90	11.40				
	Girls	53.92	11.88	-5.233	1617.343	< .001	-.260
AXY							
	Boys	54.61	13.27				
	Girls	62.38	14.06	-11.435	1614.615	< .001	-.568
BRF							
	Boys	51.33	10.62				
	Girls	55.72	11.29	-8.065	1620	< .001	-.401
HLP							
	Boys	54.22	14.34				
	Girls	60.89	15.87	-8.887	1603.734	< .001	-.441
NFC							
	Boys	53.93	12.42				
	Girls	60.63	12.24	-10.943	1620	< .001	-.543
OCS							
	Boys	51.64	11.68				
	Girls	56.39	13.29	-7.651	1593.728	< .001	-.380
SFD							
	Boys	55.03	13.17				
	Girls	62.16	12.16	-11.317	1609.721	< .001	-.562
SPF							
	Boys	47.83	8.65				
	Girls	49.13	9.24	-2.924	1620	.004	-.145
STW							
	Boys	55.17	11.90				
	Girls	62.08	11.39	-11.943	1620	< .001	-.593

Note. ANP = Anger Proneness. AXY = Anxiety. BRF = Behavior Restricting Fears. HLP = Helplessness/Hopelessness. NFC = Inefficacy. OCS = Obsessions/Compulsions. SFD = Self-Doubt. SPF = Specific Fears. STW = Stress/Worry. Means and standard deviations were obtained from the Unrounded, Untruncated T-scores.

[a] $n = 811$ for each group.

The endorsement frequencies in the keyed direction as well as chi-square tests and phi-coefficients for the Externalizing Scales items across male and female groups are reported in Table 5. In terms of descriptive statistics, boys generally had higher endorsement frequencies for Externalizing Scale items with statistically significant differences (≤ .01) in 14 out of the 24

items. Although, many of these effect sizes were rather small (< .20), the CNP Scale had three items (33, 88, and 127) with phi coefficients at -.26, -.29, and -.26 respectively. The girls' sample ranged from having one to four items within a scale where they had higher item endorsement frequencies (with the exception of the CNP scale where the boys' sample had statistically significant [$p < .001$] higher endorsement frequencies for the majority of items). Table 5 also presents corrected item-total correlations and internal consistency reliability coefficients for each Externalizing Scale. Alpha coefficients of the Externalizing Scales ranged from .66 (SUB) to .75 (NSA) in the boys' sample and .65 (ASA) to .73 (SUB) in the girls' sample. The majority of the Cronbach's alpha coefficients for each scale were similar across genders with the greatest difference in the SUB Scale (boys $\alpha = .66$, girls $\alpha = .73$).

Table 5. *Endorsement Frequencies, Chi-Square, Phi, Internal Consistency Coefficients, Standard Error of Measurement, and Corrected Item-Total Correlations for Externalizing Scale Items.*

Scale	Item	Endorsement Frequencies		χ^2	ϕ		Item-Total *r*	
		Boys [a]	Girls [a]				Boys [a]	Girls [a]
AGG								
	16	**61.5%**	56.6%	4.08*	-.050		.41	.46
	36	7.4%	**10.0%**	3.42	.046		.29	.32
	41	63.9%	**67.4%**	2.3	.038		.39	.41
	130	**42.3%**	39.5%	1.35	-.029		.52	.58
	149	23.3%	**23.7%**	.031	.004		.51	.47
	186	**23.1%**	12.3%	32.04***	-.141		.38	.29
	233	**43.8%**	34.9%	13.39***	-.091		.43	.43
	240	**22.1%**	14.2%	17.02***	-.102		.34	.37
						α	.72	.72
						SEM	6	6
ASA								
	35	51.2%	**53.8%**	1.09	.026		.44	.45
	80	**43.8%**	33.2%	19.26***	-.109		.43	.36
	99	55.5%	**56.7%**	.25	.012		.34	.23
	171	61.4%	**69.2%**	10.80***	.082		.29	.33
	193	**59.7%**	57.1%	1.12	-.026		.44	.49
	219	**58.4%**	54.7%	2.26	-.037		.43	.44
						α	.67	.65
						SEM	7	7
CNP								
	14	**52.9%**	40.8%	23.78***	-.121		.46	.42
	33	**45.6%**	21.5%	106.25***	-.256		.49	.44
	88	**44.5%**	17.4%	139.63***	-.293		.57	.53
	110	**19.5%**	12.7%	13.81***	-.092		.37	.33
	127	**51.0%**	25.8%	109.52***	-.260		.58	.56
	148	**12.3%**	9.2%	4.00*	-.050		.30	.24
	238	**30.3%**	28.4%	.761	-.022		.33	.45
						α	.73	.71
						SEM	7	6
NPI								
	19	22.4%	**22.8%**	.032	.004		.47	.49
	64	**18.2%**	14.7%	3.77*	-.048		.54	.50
	111	**22.6%**	15.9%	11.57***	-.084		.42	.42
	146	**30.8%**	26.6%	3.48	-.046		.52	.46
	160	**24.3%**	20.0%	4.38*	-.052		.38	.39
						α	.71	.70
						SEM	6	6

Table 5. Continued.

Scale	Item	Endorsement Frequencies		χ^2	ϕ		Item-Total *r*	
		Boys [a]	Girls [a]				Boys [a]	Girls [a]
NSA								
	29	**37.1%**	32.7%	3.52	-.047		.59	.50
	75	48.1%	**50.7%**	1.089	.026		.54	.52
	104	**19.4%**	17.1%	1.34	-.029		.39	.32
	136	49.9%	**50.3%**	.022	.004		.55	.52
	195	69.9%	**76.3%**	8.48**	.072		.46	.42
	241	58.9%	**70.2%**	22.31***	.117		.37	.35
						α	.75	.70
						SEM	7	7
SUB								
	43	**14.5%**	11.0%	4.66*	-.054		.57	.60
	72	**4.7%**	4.3%	.129	-.009		.35	.45
	166	13.3%	**13.9%**	.131	.009		.50	.55
	235	**26.4%**	20.5%	7.92**	-.070		.44	.55
						α	.66	.73
						SEM	6	5

Note. AGG = Aggression. ASA = Antisocial Attitudes. CNP = Conduct Problems. NPI = Negative Peer Influence. NSA = Negative School Attitudes. SUB = Substance Abuse. Cronbach's alpha and item endorsements obtained from raw data in the keyed direction. Higher endorsement between the two genders are bolded for ease of identification purposes and do not necessarily represent significant differences. SEM = Standard Error of Measurement.

[a] n = 811 for each group.

* ≤ .05. ** ≤ .01. ***≤. 001.

As seen in Table 5, the standard error of measurements for the Externalizing Scales were generally comparable to the data in the MMPI-A-RF manual ranging from 6 to 7 T-score points for the boys' sample and 5 to 7 T-score points for the girls' sample. However, the standard error of measurements for the SUB and NPI Scales were lower. In the MMPI-A-RF normative sample, the SUB Scale was 7 T-score points for both genders, and in this sample, it was 6 T-score points for the boys' sample and 5 T-score points for the girls' sample. For the NPI Scale, the normative sample SEM was 10 T-score points for both genders, but the SEM was 6 for both genders in the

present study. In examining the corrected item-total correlations for the Externalizing Scales, the lowest item-total correlation (r = .29) for the boys' sample (item 171) and the lowest item-total correlation (r = .23) for the girls' sample (item 99) were in the ASA Scale. A majority of the items had similar item-total correlations with the exceptions of item 186 (AGG), item 99 (ASA), item 238 (CNP), and items 72 and 235 (SUB). Table 6 includes the Specific Problem Scale intercorrelations by gender for the Externalizing Scales.

Table 6. *Externalizing Specific Problem Scale Correlations by Gender*

	NSA	ASA	CNP	SUB	AGG	NPI
NSA		.473**	.206**	.176**	.429**	.188**
ASA	.434**		.218**	.257**	.501**	.306**
CNP	.198**	.302**		.409**	.418**	.513**
SUB	.200**	.275**	.421**		.254**	.399**
AGG	.360**	.509**	.423**	.267**		.336**
NPI	.223**	.264**	.426**	.404**	.332**	

Note. Boys (N = 811) correlations in the upper shaded diagonal. Girls (N = 811) correlations in the lower diagonal. Raw scale totals were used in the analysis. NSA = Negative School Attitudes. ASA = Antisocial Attitudes. CNP = Conduct Problems. SUB = Substance Abuse. AGG = Aggression. NPI = Negative Peer Influence.
** Correlation statistically significant at 0.01 level.

Regarding differences in scale means, statistically significant differences (p < .01) were observed with the boys' sample having higher T-scores in the AGG, CNP, and NPI Externalizing Scales (Table 7). Cohen's d values were computed for each t-test of the Externalizing Scales. Following Cohen's guidelines (d = .20, .50, and .80 for small, medium, and large effects; Cohen, 1992), values indicated a medium effect size for CNP (d = .54) and a less than small (< .20) effect size for AGG (d = .14), ASA (d = .04), NPI (d = .13), NSA (d = .06), and SUB (d = .10).

Table 7. *Independent Sample T-tests for Externalizing Scales*

Scale	Gender [a]	*M*	*SD*	*t*	*df*	*p*-value	Cohen's d
AGG							
	Boys	49.31	11.59				
	Girls	47.68	10.90	2.909	1620	.004	.144
ASA							
	Boys	51.15	11.67				
	Girls	50.65	11.27	.892	1620	.372	.044
CNP							
	Boys	55.42	12.95				
	Girls	48.94	11.12	10.818	1583.695	< .001	.537
NPI							
	Boys	49.38	11.16				
	Girls	47.93	10.38	2.714	1620	.007	.135
NSA							
	Boys	56.64	13.85				
	Girls	57.40	12.84	-1.150	1610.943	.250	-.057
SUB							
	Boys	48.36	9.71				
	Girls	47.41	9.61	1.977	1620	.048	.098

Note. AGG = Aggression. ASA = Antisocial Attitudes. CNP = Conduct Problems. NPI = Negative Peer Influence. NSA = Negative School Attitudes. SUB = Substance Abuse. Means and standard deviations were obtained from the Unrounded, Untruncated T-scores.
[a] $n = 811$ for each group.

Confirmatory Factor Analysis

The CFA model fit of a one-factor model for each of the Internalizing Scale across boys and girls is presented in Table 8. In examining the goodness-of-fit indices of CFI, TLI, and RMSEA, most of the scales indicated adequate to good model fit across groups with the exception of the OCS (TLI = .887; RMSEA = .096) and SFD (RMSEA = .115) Scales for boys and the OCS (TLI = .804; RMSEA = .152) and SFD (RMSEA = .110) Scales for girls.

Modification indices were examined to determine whether freeing parameters would result in an improved model fit for the OCS Scale and the SFD Scale. For the OCS Scale, items 139 and 221 were the only items that indicated a justifiable model improvement in both boys'

($\Delta\chi^2 = 14.574$, $df = 1$, $p < .001$) and girls' ($\Delta\chi^2 = 31.310$, $df = 1$, $p < .001$) groups after calculating the modification indices. Further review of item 139 and item 221 revealed both items had similar content related to ruminating thoughts. As such, the errors for items 139 and 221 were correlated to improve the model fit, thereby making it acceptable for further analyses, as indicated in Table 8 (OCS 139w221). In reviewing the items in the SFD Scale, items 73 and 202 had similar content related to self-confidence and indicated a justifiable model improvement in both boys' ($\Delta\chi^2 = 32.249$, $df = 1$, $p < .001$) and girls' ($\Delta\chi^2 = 30.715$, $df = 1$, $p < .001$) groups after calculating the modification indices. As such, the errors for items 73 and 202 were correlated to improve the model fit, thereby making it acceptable for further analyses as indicated in Table 8 (SFD 73w202).

Table 8. *Confirmatory Factor Analysis of Internalizing Scales Across Genders*

	Model [a]	χ^2	*df*	*p*	CFI	TLI	RMSEA (95% CI)
ANP							
	Boys	6.466	5	.263	.998	.997	.019 (< .001 - .055)
	Girls	12.757	5	.025	.992	.983	.044 (.014 - .074)
AXY							
	Boys	3.279	2	.194	.998	.995	.028 (< .001 - .081)
	Girls	1.41	2	.565	1.000	1.000	< .001 (< .001 - .059)
BRF							
	Boys	0	0	< .001	1.00	1.00	< .001 (< .001 - < .001)
	Girls	0	0	< .001	1.00	1.00	< .001 (< .001 - < .001)
HLP							
	Boys	151.088	35	< .001	.956	.944	.064 (.054 - .075)
	Girls	92.288	35	< .001	.980	.974	.045 (.034 - .056)
NFC							
	Boys	.418	2	.811	1.000	1.000	< .001 (< .001 - .043)
	Girls	10.877	2	.004	.989	.967	.074 (.035 - .120)
OCS							
	Boys	16.875	2	< .001	.962	.887	.096 (.057 - .140)
	Girls	39.371	2	< .001	.935	.804	.152 (.113 - .195)
139w221	Boys	.099	1	.752	1.00	1.00	< .001 (< .001 - .064)
	Girls	.559	1	.454	1.00	1.00	< .001 (< .001 - .084)
SFD							
	Boys	58.889	5	< .001	.979	.958	.115 (.090 - .143)
	Girls	54.477	5	< .001	.975	.949	.110 (.085 - .138)
73w202	Boys	14.483	4	.005	.996	.990	.057 (.027 - .090)
	Girls	7.825	4	.098	.998	.995	.034 (< .001 - .070)
SPF							
	Boys	3.290	2	.193	.981	.943	.028 (< .001 - .081)
	Girls	.002	2	.998	1.000	1.000	< .001 (< .001 - < .001)
STW							
	Boys	17.053	14	.253	.998	.997	.016 (< .001 - .040)
	Girls	15.238	14	.362	.999	.998	.010 (< .001 - .036)

Note. CFI = Comparative fit index. TLI = Tucker-Lewis index. RMSEA = Root Mean Square Error of Approximation. CI = Confidence Interval. ANP = Anger Proneness. AXY = Anxiety. BRF = Behavior Restricting Fears. HLP = Helplessness/Hopelessness. NFC = Inefficacy. OCS = Obsessions/Compulsions. SFD = Self-Doubt. SPF = Specific Fears. STW = Stress/Worry. 139w221 = correlated errors of item 139 and item 221. 73w202 = correlated errors of item 73 and item 202.

[a] n = 811 for each group.

The standardized factor loadings and thresholds for the Internalizing Scales are presented in Table 9. Standardized factor loadings greater than or equal to .30 or .40 are often considered salient in applied research (Brown, 2015). Thresholds mark the point where respondents are likely to switch from a 0 to a 1 on an item. These thresholds are essentially z-scores associated with response probabilities, and they can be converted to the likelihood of endorsing an item using a z-table for the standard normal distribution (Finney & DiStefano, 2013).

For the ANP Scale, factor loadings for the boys' sample varied from .597 to .807 with thresholds ranging from – .191 to .395. For the girls' sample, factor loadings ranged from .611 to .816 with thresholds ranging from - .504 to .342. For the AXY Scale, factor loadings for the boys' sample varied from .600 to .811 with thresholds ranging from – .154 to .920. For the girls' sample, factor loadings ranged from .560 to .836 with thresholds ranging from - .812 to .442. The BRF Scale indicated not only low factor loadings (.201) but also a Heywood Case in the girls' sample. Item 123 revealed a standardized factor loading of 1.212. For a one-factor model, a standardized loading greater than 1.00 (more than 100% of the variable's variance is explained by the factor) is referred to as a Heywood Case, and it implies the residual variance is negative and an improper factor solution was obtained (Wang et al., 2023). Thresholds for the girls' sample ranged from - .054 to 1.228. For the boys' sample, factor loadings ranged from .483 to .636 with thresholds ranging from .590 to 1.296.

For the HLP Scale, factor loadings for the boys' sample varied from .523 to .895 with thresholds ranging from – 0.023 to .897. For the girls' sample, factor loadings ranged from .449 to .877 with thresholds ranging from - .435 to .463. For the NFC Scale, factor loadings for the boys' sample varied from .648 to .826 with thresholds ranging from – .255 to .251. For the girls' sample, factor loadings ranged from .632 to .872 with thresholds ranging from – .803 to - .157.

For the OCS Scale with correlated error terms for items 139 and 221, factor loadings for the boys' sample were lower and varied from .389 to .893 with thresholds ranging from – .088 to .670. For the girls' sample, factor loadings ranged from .421 to .814 with thresholds ranging from - .662 to .378. For the SFD Scale with correlated error terms for items 73 and 202, factor loadings for the boys' sample varied from .463 to .928 with thresholds ranging from -.381 to .267. For the girls' sample, factor loadings ranged from .575 to .917 with thresholds ranging from - 1.089 to -.085. For the SPF Scale, factor loadings for both groups were lower with the boys' sample ranging from .353 to .539 with thresholds ranging from – .082 to .757. For the girls' sample, factor loadings ranged from .219 to .563 with thresholds ranging from - .104 to .628. For the STW Scale, factor loadings for the boys' sample varied from .452 to .897 with thresholds ranging from – .834 to .194. For the girls' sample, factor loadings ranged from .526 to .954 with thresholds ranging from – 1.348 to .204.

Table 9. *Standardized Factor Loadings and Thresholds of Items for Internalizing Scales*

		Boys [a]		Girls [a]	
Internalizing Scales	Items	Factor Loadings	Thresholds	Factor Loadings	Thresholds
ANP					
	54	.807	-.079	.816	-.358
	93	.706	-.150	.611	-.365
	161	.678	-.191	.718	-.504
	165	.695	.395	.699	.342
	229	.597	.296	.655	.048
AXY					
	55	.760	.920	.783	.442
	71	.788	.332	.836	-.116
	153	.811	-.154	.699	-.812
	209	.600	.598	.560	.113
BRF					
	50	.483	1.296	.201	1.228
	65	.610	.749	.307	.525
	123	.636	.590	1.212	-.054
HLP					
	56	.523	.418	.636	.144
	60	.614	.201	.663	-.435
	62	.722	.408	.660	.236
	119	.895	.616	.877	.332
	121	.672	-.023	.657	-.255
	162	.668	.594	.738	.274
	169	.540	.737	.488	.255
	194	.782	.897	.732	.463
	228	.579	.248	.611	-.017
	239	.579	.014	.449	-.236
NFC					
	51	.720	.176	.632	-.157
	112	.806	-.255	.786	-.803
	159	.648	.251	.690	-.352
	224	.826	-.032	.872	-.558
OCS					
	15	.328	.670	.578	.378
	87	.584	.239	.652	.101
	139	.671	.342	.664	.051
	221	.858	-.088	.874	-.662

Table 9. Continued

		Boys [a]		Girls [a]	
Internalizing Scales	Items	Factor Loadings	Thresholds	Factor Loadings	Thresholds
OCS					
139w221	15	.389	.670	.655	.378
	87	.893	.239	.814	.101
	139	.393	.342	.421	.051
	221	.542	-.088	.626	-.662
	139w221	.507		.569	
SFD					
	73	.779	-.082	.811	-.670
	79	.903	-.365	.876	-.901
	145	.650	.267	.699	-.085
	202	.692	-.381	.753	-1.089
	234	.911	-.119	.893	-.609
73w202					
	73	.711	-.082	.723	-.670
	79	.915	-.365	.895	-.901
	145	.666	.267	.720	-.085
	202	.594	-.381	.612	-1.089
	234	.928	-.119	.917	-.609
	73w202	.463		.575	
SPF					
	44	.498	.757	.563	.628
	109	.353	.129	.219	.008
	147	.420	-.082	.451	-.104
	213	.539	.658	.557	.494
STW					
	4	.575	.023	.606	-.529
	48	.897	-.605	.954	-1.095
	67	.666	-.741	.526	-.934
	77	.452	.194	.577	-.204
	129	.657	-.834	.765	-1.348
	198	.722	-.261	.706	-.795
	203	.625	-.525	.654	-1.056

Note. ANP = Anger Proneness. AXY = Anxiety. BRF = Behavior Restricting Fears. HLP = Helplessness/Hopelessness. NFC = Inefficacy. OCS = Obsessions/Compulsions. SFD = Self-Doubt. SPF = Specific Fears. STW = Stress/Worry. 139w221 = correlated errors of item 139 and item 221. 73w202 = correlated errors of item 73 and item 202.

[a] n = 811.

The CFA model fit of a one-factor model for each of the Externalizing Scales for boys and girls is presented in Table 10. In examining the goodness-of-fit indices of CFI, TLI, and RMSEA, most of the scales indicated adequate to good model fit across groups with the exception of the NPI (TLI = .821; RMSEA = .180) and SUB (TLI = .876; RMSEA = .141) Scales for boys and the NPI (TLI = .808; RMSEA = .178) and SUB (RMSEA = .109) Scales for girls.

Modification indices were examined to determine whether freeing parameters would result in an improved model fit for the NPI Scale and the SUB Scale. For the NPI Scale, items 19 and 146 were the only items that indicated a justifiable model improvement in both boys' ($\Delta\chi^2$ = 71.092, df = 1, $p < .001$) and girls' ($\Delta\chi^2$ = 64.463, df = 1, $p < .001$) groups after calculating the modification indices. Further review of item 19 and item 146 revealed both items had similar content related to parental approval of friends. As such, the errors for items 19 and 146 were correlated to improve the model fit, thereby making it acceptable for further analyses as indicated in Table 10 (NPI 19w146).

Table 10. *Confirmatory Factor Analysis of Externalizing Scales Across Genders*

	Model [a]	χ^2	*df*	*p*	CFI	TLI	RMSEA (95% CI)
AGG							
	Boys	50.635	20	< .001	.979	.971	.043 (.029 - .059)
	Girls	43.617	20	.001	.986	.981	.038 (.023 - .054)
ASA							
	Boys	6.208	9	.718	1.000	1.000	< .001 (< .001 - .030)
	Girls	4.932	9	.840	1.000	1.000	< .001 (< .001 - .023)
CNP							
	Boys	62.172	14	< .001	.978	.966	.065 (.049 - .082)
	Girls	96.737	14	< .001	.947	.920	.085 (.070 - .102)
NPI							
	Boys	136.218	5	< .001	.911	.821	.180 (.155 - .207)
	Girls	133.242	5	< .001	.904	.808	.178 (.152 - .205)
19w146	Boys	22.745	4	< .001	.987	.968	.076 (.048 - .108)
	Girls	35.361	4	< .001	.977	.941	.098 (.070 - .129)
NSA							
	Boys	18.070	9	.034	.995	.992	.035 (.009 - .059)
	Girls	29.469	9	< .001	.984	.974	.053 (.032 - .075)
SUB							
	Boys	34.303	2	< .001	.959	.876	.141 (.102 - .184)
	Girls	21.231	2	< .001	.986	.957	.109 (.070 - .153)
72w166	Boys	10.499	1	.001	.988	.927	.108 (.056 - .172)
	Girls	2.703	1	.100	.999	.992	.046 (< .001 - .115)

Note. CFI = Comparative fit index. TLI = Tucker-Lewis index. RMSEA = Root Mean Square Error of Approximation. CI = Confidence Interval. AGG = Aggression. ASA = Antisocial Attitudes. CNP = Conduct Problems. NPI = Negative Peer Influence. NSA = Negative School Attitudes. SUB = Substance Abuse. 19w146 = correlated errors of item 19 and item 146. 72w166 = correlated errors of item 72 and item 166.

[a] $n = 811$ for each group.

For the SUB Scale, a review of items 72 and 166 revealed both items had similar content related to the consumption of excessive alcohol. As such, the errors for items 72 and 166 were correlated to improve the model fit. Correlating the error terms indicated a justifiable model improvement in both boys' ($\Delta\chi^2 = 21.898$, $df = 1$, $p < .001$) and girls' ($\Delta\chi^2 = 16.131$, $df = 1$, $p < .001$) groups after calculating the modification indices. However, RMSEA continued to exhibit poor model fit, and further examination of modification indices to improve model fit was not

justified. Therefore, further measurement invariance testing could not be completed for the SUB Scale. The standardized factor loadings and thresholds for the Externalizing Scales are presented in Table 11. For the AGG Scale, factor loadings for the boys' sample varied from .546 to .812 with thresholds ranging from – .355 to 1.447. For the girls' sample, factor loadings ranged from .526 to .882 with thresholds ranging from - .452 to 1.282. For the ASA Scale, factor loadings for the boys' sample varied from .464 to .701 with thresholds ranging from – .290 to .157. For the girls' sample, factor loadings ranged from .359 to .777 with thresholds ranging from - .501 to .435. For the CNP Scale, factor loadings for the boys' sample varied from .482 to .892 with thresholds ranging from – .073 to 1.159. For the girls' sample, factor loadings ranged from .492 to .872 with thresholds ranging from .232 to 1.326.

For the NPI Scale with correlated error terms for items 19 and 146, factor loadings for the boys' sample varied from .516 to .972 with thresholds ranging from .501 to .906. For the girls' sample, factor loadings ranged from .512 to .963 with thresholds ranging from .624 to 1.051. For the NSA Scale, factor loadings for the boys' sample varied from .535 to .884 with thresholds ranging from – 0.522 to .865. For the girls' sample, factor loadings ranged from .549 to .789 with thresholds ranging from - .717 to .949. For the SUB Scale with correlated error terms for items 72 and 266, factor loadings for the boys' sample varied from .590 to .985 with thresholds ranging from .631 to 1.676. For the girls' sample, factor loadings ranged from .582 to .947 with thresholds ranging from .825 to 1.715.

Table 11. *Standardized Factor Loadings and Thresholds of Items for Externalizing Scales*

		Boys [a]		Girls [a]	
Externalizing Scales	Items	Factor Loadings	Thresholds	Factor Loadings	Thresholds
AGG					
	16	.644	-.293	.692	-.166
	36	.617	1.447	.607	1.282
	41	.638	-.355	.725	-.452
	130	.776	.194	.882	.267
	149	.812	.729	.711	.717
	186	.603	.737	.526	1.159
	233	.641	.157	.626	.388
	240	.546	.770	.643	1.072
ASA					
	35	.691	-.029	.723	-.094
	80	.684	.157	.609	.435
	99	.533	-.138	.359	-.169
	171	.464	-.290	.537	-.501
	193	.701	-.245	.777	-.179
	219	.665	-.213	.695	-.119
CNP					
	14	.647	-.073	.649	.232
	33	.701	.110	.692	.791
	88	.878	.138	.872	.939
	110	.597	.860	.585	1.141
	127	.892	-.026	.861	.650
	148	.553	1.159	.492	1.326
	238	.482	.515	.677	.572
NPI					
	19	.833	.757	.836	.745
	64	.832	.906	.859	1.051
	111	.626	.753	.644	.998
	146	.864	.501	.809	.624
	160	.662	.697	.728	.843
19w146					
	19	.516	.757	.539	.745
	64	.972	.906	.963	1.051
	111	.685	.753	.707	.998
	146	.585	.501	.512	.624
	160	.713	.697	.775	.843
	19w146	.752		.731	

Table 11. Continued

		Boys [a]		Girls [a]	
Externalizing Scales	Items	Factor Loadings	Thresholds	Factor Loadings	Thresholds
NSA					
	29	.884	.329	.789	.449
	75	.773	.048	.771	-.017
	104	.661	.865	.598	.949
	136	.804	.002	.779	-.008
	195	.732	-.522	.702	-.717
	241	.535	-.226	.549	-.529
SUB					
	43	.905	1.056	.914	1.228
	72	.779	1.676	.867	1.715
	166	.823	1.112	.850	1.083
	235	.761	.631	.863	.825
72w166					
	43	.985	1.056	.947	1.228
	72	.590	1.676	.753	1.715
	166	.715	1.112	.782	1.083
	235	.764	.631	.878	.825
	72w166	.620		.582	

Note. AGG = Aggression. ASA = Antisocial Attitudes. CNP = Conduct Problems. NPI = Negative Peer Influence. NSA = Negative School Attitudes. SUB = Substance Abuse. 19w146 = correlated errors of item 19 and item 146. 72w166 = correlated errors of item 72 and item 166.
[a] n = 811.

Measurement Invariance Tests

Internalizing Scales

Measurement invariance tests were conducted for each Internalizing Scale across sexes with the exception of the BRF Scale due to the Heywood case in the girls' sample. Bollen (1987) recommended multiple solutions to address Heywood cases, including dropping the problematic indicator, obtaining a larger sample, increasing the number of indicators per factor, fixing the improper estimate to a plausible value, or using an inequality restriction. Since the MMPI-A-RF is an established testing instrument and the BRF Scale only has three items, dropping an item

and increasing the number of items for the factor are not possible solutions until the test is revised. For the purposes of the study, fixing the estimate or restricting the inequality would limit the implications and further measurement invariance testing. Therefore, the BRF Scale was dropped from further analyses.

In examining measurement invariance, the configural model fit was acceptable for all remaining tested Internalizing Scales (see Tables 12 – 19). As configural invariance was established, scalar invariance was tested by setting the equality constraints on both factor loadings and thresholds across groups. The model fit difference between the scalar model and the configural model met full scalar invariance for the ANP Scale, $\Delta\chi^2_{diff}$ (3) = 5.781, p = .1228 (see Table 12); the AXY Scale, $\Delta\chi^2_{diff}$ (2) = 7.371, p = .0251 (see Table 13); SFD Scale $\Delta\chi^2_{diff}$ (3) = 12.432, p = .006 (see Table 17); SPF Scale $\Delta\chi^2_{diff}$ (2) = .493, p = .7817 (see Table 18); and the STW Scale $\Delta\chi^2_{diff}$ (5) = 7.358, p = .1954 (see Table 19). Partial scalar invariance was calculated for the HLP, NFC, and OCS Scales (see Tables 14 – 16). Freely estimating item factor loadings and thresholds resulted in statistically significant DIFFTEST p values signifying noninvariant items across sexes. For the HLP Scale, nine items reached noninvariance (56, 62, 119, 121, 162, 169, 194, 228, 239; see Table 14). For the NFC Scale, two items (112, 224; see Table 15) were found to be noninvariant. For the OCS Scale, two items (15, 139; see Table 16) were found to be noninvariant.

Table 12. *Fit Indices for Invariance Models across Genders for Anger Proneness (ANP)*

Model/Item	χ^2 (*df*)	*p*	CFI	TLI	RMSEA (95% CI)	Model Comparison	Δ χ^2 (*df*)	*p*	Δ CFI	Δ TLI	Δ RMSEA
Boys [a]	6.47 (5)	.263	.998	.997	.019 (< .001 - .055)						
Girls [a]	12.76 (5)	.025	.992	.983	.044 (.014 - .074)						
Model 1: Configural [b]	19.49 (10)	.034	.995	.990	.034 (.009 - .057)						
Model 2: Full MI [b]	25.19 (13)	.021	.994	.990	.034 (.013 - .054)	2 vs 1	5.781 (3)	.122	.001	< .001	< .001

Note. CFI = Comparative fit index. TLI = Tucker-Lewis index. RMSEA = Root Mean Square Error of Approximation. CI = Confidence Interval. MI = Measurement Invariance. Statistically significant p value ≤ .001.
[a] *n* = 811. [b] *n* = 1,622.

Table 13. *Fit Indices for Invariance Models across Genders for Anxiety (AXY)*

Model/Item	χ^2 (*df*)	*p*	CFI	TLI	RMSEA (95% CI)	Model Comparison	Δ χ^2 (*df*)	*p*	Δ CFI	Δ TLI	Δ RMSEA
Boys [a]	3.28 (2)	.194	.998	.995	.028 (< .001 - .081)						
Girls [a]	1.14 (2)	.565	1.00	1.00	< .001 (< .001 - .059)						
Model 1: Configural [b]	4.49 (4)	.344	1.00	.999	.012 (< .001 - .056)						
Model 2: Full MI [b]	12.77 (6)	.046	.995	.990	.037 (.004 - .066)	2 vs 1	7.37 (2)	.025	.005	.009	.025

Note. CFI = Comparative fit index. TLI = Tucker-Lewis index. RMSEA = Root Mean Square Error of Approximation. CI = Confidence Interval. MI = Measurement Invariance. Statistically significant p value ≤ .001.
[a] *n* = 811. [b] *n* = 1,622.

Table 14. *Fit Indices for Invariance Models across Genders for Helplessness/Hopelessness (HLP)*

Model/Item	χ^2 (*df*)	*p*	CFI	TLI	RMSEA (95% CI)	Model Comparison	$\Delta\chi^2$ (*df*)	*p*	Δ CFI	Δ TLI	Δ RMSEA
Boys [a]	151.09 (35)	<.001	.956	.944	.064 (.054 - .075)						
Girls [a]	92.29 (35)	<.001	.980	.974	.045 (.034 - .056)						
Model 1: Configural [b]	243.92 (70)	<.001	.968	.959	.055 (.048 - .063)						
Model 2: Full MI [b]	298.05 (78)	<.001	.960	.953	.059 (.052 - .066)	2 vs 1	50.63 (8)	<.001	.008	.006	.004
Model 3: Partial MI [b]											
56	298.87 (77)	<.001	.959	.952	.060 (.053 - .067)	56 vs 1	50.50 (7)	<.001	.001	.001	.001
60	262.77 (77)	<.001	.966	.960	.055 (.047 - .062)	60 vs 1	21.39 (7)	.003	.006	.007	.004
62	285.86 (77)	<.001	.962	.955	.058 (.051 - .065)	62 vs 1	39.79 (7)	<.001	.002	.002	.001
119	289.92 (77)	<.001	.961	.954	.058 (.051 - .066)	119 vs 1	42.71 (7)	<.001	.001	.001	.001
121	288.11 (77)	<.001	.961	.955	.058 (.051 - .065)	121 vs 1	41.48 (7)	<.001	.001	.002	.001
162	299.42 (77)	<.001	.959	.952	.060 (.053 - .067)	162 vs 1	50.93 (7)	<.001	.001	.001	.001
169	288.40 (77)	<.001	.961	.955	.058 (.051 - .065)	169 vs 1	42.75 (7)	<.001	.001	.002	.001
194	297.61 (77)	<.001	.960	.953	.059 (.052 - .067)	194 vs 1	49.18 (7)	<.001	<.001	<.001	<.001
228	298.50 (77)	<.001	.959	.953	.060 (.053 - .067)	228 vs 1	49.87 (7)	<.001	.001	<.001	.001
239	299.25 (77)	<.001	.959	.952	.060 (.053 - .067)	239 vs 1	50.66 (7)	<.001	.001	.001	.001

Note. CFI = Comparative fit index. TLI = Tucker-Lewis index. RMSEA = Root Mean Square Error of Approximation. CI = Confidence Interval. MI = Measurement Invariance. Statistically significant p value ≤ .001. Noninvariant items are bolded.
[a] n = 811. [b] n = 1,622.

Table 15. *Fit Indices for Invariance Models across Genders for Inefficacy (NFC)*

Model/Item	χ^2 (*df*)	*p*	CFI	TLI	RMSEA (95% CI)	Model Comparison	$\Delta\chi^2$ (*df*)	*p*	Δ CFI	Δ TLI	Δ RMSEA
Boys [a]	.42 (2)	.811	1.00	1.00	< .001 (< .001 - .043)						
Girls [a]	10.88 (2)	.004	.989	.967	.074 (.035 - .120)						
Model 1: Configural [b]	11.58 (4)	.020	.996	.987	.048 (.017 - .082)						
Model 2: Full MI [b]	26.26 (6)	<.001	.989	.977	.065 (.041 - .091)	2 vs 1	14.09 (2)	<.001	.007	.01	.017
Model 3: Partial MI [b]											
51	16.96 (5)	.004	.993	.984	.054 (.027 - .084)	51 vs 1	5.48 (1)	.019	.003	.003	.006
112	25.08 (5)	<.001	.989	.973	.070 (.045 - .099)	112 vs 1	11.88 (1)	<.001	.007	.014	.022
159	11.57 (5)	.041	.996	.991	.040 (.008 - .071)	159 vs 1	< .001 (1)	.997	<.001	.004	.008
224	24.36 (5)	<.001	.989	.974	.069 (.043 - .098)	224 vs 1	11.28 (1)	<.001	.007	.013	.021

Note. CFI = Comparative fit index. TLI = Tucker-Lewis index. RMSEA = Root Mean Square Error of Approximation. CI = Confidence Interval. MI = Measurement Invariance. Statistically significant p value ≤ .001. Noninvariant items are bolded.
[a] n = 811. [b] n = 1,622.

Table 16. *Fit Indices for Invariance Models across Genders for Obsessions/Compulsions (OCS)*

Model/Item	χ^2 (*df*)	*p*	CFI	TLI	RMSEA (95% CI)	Model Comparison	Δ χ^2 (*df*)	*p*	Δ CFI	Δ TLI	Δ RMSEA
Boys 139 W 221[a]	.10 (1)	.752	1.00	1.00	<.001 (<.001 - .064)						
Girls 139 W 221[a]	.56 (1)	.454	1.00	1.00	<.001 (<.001 - .084)						
Model 1: Configural[b]	.62 (2)	.732	1.00	1.00	<.001 (<.001 - .049)						
Model 2: Full MI[b]	37.93 (4)	<.001	.965	.895	.102 (.074 - .113)	2 vs 1	30.429 (2)	<.001	.035	.105	.102
Model 3: Partial MI[b]											
15	14.57 (3)	.002	.988	.952	.069 (.036 - .106)	15 vs 1	10.716 (1)	.001	.012	.048	.069
87	2.23 (3)	.525	1.00	1.00	<.001 (<.001 - .053)	87 vs 1	1.193 (1)	.274	<.001	<.001	< .001
139	40.39 (3)	<.001	.961	.845	.124 (.092 - .159)	139 vs 1	29.168 (1)	<.001	.039	.155	.124
221	3.55 (3)	.314	.999	.998	.015 (<.001 - .053)	221 vs 1	2.280 (1)	.131	.001	.002	.015

Note. CFI = Comparative fit index. TLI = Tucker-Lewis index. RMSEA = Root Mean Square Error of Approximation. CI = Confidence Interval. MI = Measurement Invariance. Statistically significant p value ≤ .001. Noninvariant items are bolded.
[a] n = 811. [b] n = 1,622.

Table 17. *Fit Indices for Invariance Models across Genders for Self-Doubt (SFD)*

Model/Item	χ^2 (*df*)	*p*	CFI	TLI	RMSEA (95% CI)	Model Comparison	$\Delta\chi^2$ (*df*)	*p*	Δ CFI	Δ TLI	Δ RMSEA
Boys [a]	14.48 (4)	.005	.996	.990	.057 (.027 - .090)						
Girls [a]	7.83 (4)	.098	.998	.995	.034 (< .001 - .070)						
Model 1: Configural [b]	22.36 (8)	.004	.997	.992	.047 (.025 - .071)						
Model 2: Full MI [b]	35.44 (11)	<.001	.995	.990	.052 (.034 - .072)	2 vs 1	12.432 (3)	.006	.002	.002	.005

Note. CFI = Comparative fit index. TLI = Tucker-Lewis index. RMSEA = Root Mean Square Error of Approximation. CI = Confidence Interval. MI = Measurement Invariance. Statistically significant p value ≤ .001.
[a] n = 811. [b] n = 1,622.

Table 18. *Fit Indices for Invariance Models across Genders for Specific-Fears (SPF)*

Model/Item	χ^2 (*df*)	*p*	CFI	TLI	RMSEA (95% CI)	Model Comparison	$\Delta\chi^2$ (*df*)	*p*	Δ CFI	Δ TLI	Δ RMSEA
Boys [a]	3.29 (2)	.193	.981	.943	.028 (<.001 - .081)						
Girls [a]	.002 (2)	.998	1.00	1.00	<.001 (<.001 - <.001)						
Model 1: Configural [b]	3.19 (4)	.527	1.00	1.00	<.001 (<.001 - .048)						
Model 2: Full MI [b]	3.60 (6)	.731	1.00	1.00	<.001 (<.001 - .033)	2 vs 1	.493 (2)	.781	<.001	<.001	<.001

Note. CFI = Comparative fit index. TLI = Tucker-Lewis index. RMSEA = Root Mean Square Error of Approximation. CI = Confidence Interval. MI = Measurement Invariance. Statistically significant p value ≤ .001.
[a] n = 811. [b] n = 1,622.

Table 19. *Fit Indices for Invariance Models across Genders for Stress/Worry (STW)*

Model/Item	χ^2 (*df*)	*p*	CFI	TLI	RMSEA (95% CI)	Model Comparison	$\Delta\chi^2$ (*df*)	*p*	Δ CFI	Δ TLI	Δ RMSEA
Boys [a]	17.05 (14)	.253	.998	.997	.016 (< .001 - .040)						
Girls [a]	15.24 (14)	.362	.999	.998	.010 (< .001 - .036)						
Model 1: Configural [b]	32.32 (28)	.261	.998	.997	.014 (< .001 - .032)						
Model 2: Full MI [b]	40.20 (33)	.181	.997	.996	.016 (< .001 - .032)	2 vs 1	7.358 (5)	.195	.001	.001	.002

Note. CFI = Comparative fit index. TLI = Tucker-Lewis index. RMSEA = Root Mean Square Error of Approximation. CI = Confidence Interval. MI = Measurement Invariance. Statistically significant p value ≤ .001.
[a] n = 811. [b] n = 1,622.

Latent means were calculated for seven of the nine Internalizing Scales (ANP, AXY, NFC, OCS, SFD, SPF, and STW) across sexes (setting latent means for the boys' group to zero). Latent means could not be compared for the BRF Scale because it did not reach full or partial measurement invariance. For the HLP Scale, mean comparisons on the latent factors could not be made as the majority of items were noninvariant. For the NFC and OCS Scales, mean comparisons were made after freeing the equality constraints on the factor loadings and thresholds of the noninvariant items, while additionally maintaining the constraints on the other items in the model. No statistically significant differences were found in the SPF Scale ($Z = .926$, $p = .355$). A statistically significant difference was found in the remaining six latent means where the girls' sample showed a statistically significant higher mean for ANP ($Z = 4.878$, $p = < .001$); AXY ($Z = 8.296$, $p = < .001$); NFC ($Z = 8.400$, $p = < .001$); OCS ($Z = 3.064$, $p = .002$); SFD ($Z = 7.728$, $p = < .001$); and STW ($Z = 10.424$, $p = < .001$).

Externalizing Scales

Measurement invariance tests were conducted for each Externalizing Scale by gender, except for the SUB Scale since the boys' group did not indicate good model fit and using modification indices to improve model fit were not justified. The configural model fit was acceptable for all remaining tested scales (see Tables 20 – 24).

As configural invariance was established, scalar invariance was tested by setting the equality constraints on both factor loadings and thresholds across groups. The model fit difference between the scalar model and the configural model met full scalar invariance for the ASA Scale, $\Delta\chi^2_{diff}$ (4) = 11.705, $p = .0197$ (see Table 21); NPI Scale, $\Delta\chi^2_{diff}$ (3) = 4.303, $p = .2305$ (see Table 23); and the NSA Scale, $\Delta\chi^2_{diff}$ (4) = 7.937, $p = .0939$ (see Table 24). Partial scalar invariance was calculated for the AGG (see Table 20) and CNP Scales (see Table 22).

Freely estimating item factor loadings and thresholds resulted in statistically significant DIFFTEST *p* values signifying noninvariant items across sexes. For the AGG Scale, seven items reached noninvariance (16, 36, 41, 130, 149, 233, 240; see Table 20). For the CNP Scale, all seven items (14, 33, 88, 110, 127, 148, 238; see Table 22) were found to be noninvariant.

Table 20. *Fit Indices for Invariance Models across Genders for Aggression (AGG)*

Model/Item	χ^2 (*df*)	*p*	CFI	TLI	RMSEA (95% CI)	Model Comparison	Δ χ^2 (*df*)	*p*	Δ CFI	Δ TLI	Δ RMSEA
Boys [a]	50.64 (20)	<.001	.979	.971	.043 (.029 - .059)						
Girls [a]	43.62 (20)	.001	.986	.981	.038 (.023 - .054)						
Model 1: Configural [b]	94.10 (40)	<.001	.983	.976	.041 (.030 - .052)						
Model 2: Full MI [b]	122.00 (46)	<.001	.976	.971	.045 (.036 - .055)	2 vs 1	24.990 (6)	<.001	.007	.005	.004
Model 3: Partial MI [b]											
16	118.96 (45)	<.001	.977	.971	.045 (.035 - .055)	16 vs 1	21.761 (5)	<.001	.006	.005	.004
36	115.29 (45)	<.001	.978	.973	.044 (.034 - .054)	36 vs 1	19.579 (5)	.001	.005	.003	.003
41	119.63 (45)	<.001	.977	.971	.045 (.035 - .055)	41 vs 1	22.291 (5)	<.001	.006	.005	.004
130	118.67 (45)	<.001	.977	.971	.045 (.035 - .055)	130 vs 1	21.432 (5)	<.001	.006	.005	.004
149	120.26 (45)	<.001	.976	.971	.045 (.036 - .055)	149 vs 1	22.692 (5)	<.001	.007	.005	.004
186	107.38 (45)	<.001	.980	.976	.041 (.031 - .051)	186 vs 1	13.414 (5)	.019	.003	<.001	<.001
233	116.55 (45)	<.001	.978	.972	.044 (.034 - .054)	233 vs 1	20.035 (5)	.001	.005	.004	.003
240	123.98 (45)	<.001	.975	.969	.047 (.037 - .056)	240 vs 1	25.942 (5)	<.001	.008	.007	.006

Note. CFI = Comparative fit index. TLI = Tucker-Lewis index. RMSEA = Root Mean Square Error of Approximation. CI = Confidence Interval. MI = Measurement Invariance. Statistically significant p value ≤ .001. Noninvariant items are bolded.
[a] n = 811. [b] n = 1,622.

Table 21. *Fit Indices for Invariance Models across Genders for Antisocial Attitudes (ASA)*

Model/Item	χ^2 (*df*)	*p*	CFI	TLI	RMSEA (95% CI)	Model Comparison	$\Delta\chi^2$ (*df*)	*p*	Δ CFI	Δ TLI	Δ RMSEA
Boys [a]	6.21 (9)	.718	1.00	1.00	<.001 (<.001 - .030)						
Girls [a]	4.93 (9)	.840	1.00	1.00	<.001 (<.001 - .023)						
Model 1: Configural [b]	11.12 (18)	.889	1.00	1.00	<.001 (<.001 - .015)						
Model 2: Full MI [b]	24.66 (22)	.313	.998	.998	.012 (<.001 - .033)	2 vs 1	11.705 (4)	.019	.002	.002	.012

Note. CFI = Comparative fit index. TLI = Tucker-Lewis index. RMSEA = Root Mean Square Error of Approximation. CI = Confidence Interval. MI = Measurement Invariance. Statistically significant p value ≤ .001.
[a] n = 811. [b] n = 1,622.

Table 22. *Fit Indices for Invariance Models across Genders for Conduct Problems (CNP)*

Model/Item	χ^2 (*df*)	*p*	CFI	TLI	RMSEA (95% CI)	Model Comparison	$\Delta\chi^2$ (*df*)	*p*	Δ CFI	Δ TLI	Δ RMSEA
Boys [a]	62.17 (14)	<.001	.978	.966	.065 (.049 - .082)						
Girls [a]	96.74 (14)	<.001	.947	.920	.085 (.070 - .102)						
Model 1: Configural [b]	159.78 (28)	<.001	.964	.947	.076 (.065 - .088)						
Model 2: Full MI [b]	227.71 (33)	<.001	.947	.933	.085 (.075 - .096)	2 vs 1	59.594 (5)	<.001	.017	.014	.009
Model 3: Partial MI [b]											
14	210.37 (32)	<.001	.952	.937	.083 (.072 - .094)	14 vs 1	44.625 (4)	<.001	.012	.01	.007
33	216.23 (32)	<.001	.950	.935	.084 (.074 - .095)	33 vs 1	48.502 (4)	<.001	.014	.012	.008
88	214.34 (32)	<.001	.951	.935	.084 (.073 - .095)	88 vs 1	46.969 (4)	<.001	.023	.012	.008
110	235.98 (32)	<.001	.945	.928	.089 (.078 - .099)	110 vs 1	65.988 (4)	<.001	.019	.019	.013
127	226.51 (32)	<.001	.947	.931	.087 (.076 - .097)	127 vs 1	55.373 (4)	<.001	.017	.016	.011
148	239.14 (32)	<.001	.944	.927	.089 (.079 - .100)	148 vs 1	72.208 (4)	<.001	.02	.02	.013
238	172.27 (32)	<.001	.962	.950	.074 (.063 - .084)	238 vs 1	17.020 (4)	.001	.002	.003	.002

Note. CFI = Comparative fit index. TLI = Tucker-Lewis index. RMSEA = Root Mean Square Error of Approximation. CI = Confidence Interval. MI = Measurement Invariance. Statistically significant p value ≤ .001. Noninvariant items are bolded.
[a] n = 811. [b] n = 1,622.

Table 23. *Fit Indices for Invariance Models across Genders for Negative Peer Influence (NPI)*

Model/Item	χ^2 (*df*)	*p*	CFI	TLI	RMSEA (95% CI)	Model Comparison	$\Delta\chi^2$ (*df*)	*p*	Δ CFI	Δ TLI	Δ RMSEA
Boys [a]	22.75 (4)	<.001	.987	.968	.076 (.048 - .108)						
Girls [a]	35.36 (4)	<.001	.977	.941	.098 (.070 - .129)						
Model 1: Configural [b]	58.77 (8)	<.001	.982	.955	.088 (.068 - .110)						
Model 2: Full MI [b]	57.42 (11)	<.001	.983	.970	.072 (.054 - .091)	2 vs 1	4.303 (3)	.230	.001	.015	.016

Note. CFI = Comparative fit index. TLI = Tucker-Lewis index. RMSEA = Root Mean Square Error of Approximation. CI = Confidence Interval. MI = Measurement Invariance. Statistically significant p value ≤ .001.
[a] n = 811. [b] n = 1,622.

Table 24. *Fit Indices for Invariance Models across Genders for Negative School Attitudes (NSA)*

Model/Item	χ^2 (*df*)	*p*	CFI	TLI	RMSEA (95% CI)	Model Comparison	$\Delta\chi^2$ (*df*)	*p*	Δ CFI	Δ TLI	Δ RMSEA
Boys [a]	18.07 (9)	.034	.995	.992	.035 (.009 - .059)						
Girls [a]	29.47 (9)	<.001	.984	.974	.053 (.032 - .075)						
Model 1: Configural [b]	47.65 (18)	<.001	.991	.985	.045 (.030 - .061)						
Model 2: Full MI [b]	54.78 (22)	<.001	.990	.986	.043 (.029 - .057)	2 vs 1	7.937 (4)	.093	.001	.001	.002

Note. CFI = Comparative fit index. TLI = Tucker-Lewis index. RMSEA = Root Mean Square Error of Approximation. CI = Confidence Interval. MI = Measurement Invariance. Statistically significant p value ≤ .001.
[a] n = 811. [b] n = 1,622.

Latent means were calculated for three of the six scales across sexes (setting latent means for the boys' group to zero). For the AGG and CNP Scales, mean comparisons on the latent factors could not be made as the majority of items were noninvariant. Further mean comparisons of the latent factors could not be made on the SUB Scale because configural invariance was not obtained. No statistically significant differences ($p < .01$) in means between the two groups were found for the ASA ($Z = -2.312$, $p = .021$), NSA ($Z = .925$, $p = .355$), and NPI ($Z = -1.319$, $p = .187$) Scales.

CHAPTER VI

DISCUSSION

The present study aimed to examine the measurement invariance of the MMPI-A-RF Internalizing and Externalizing Specific Problem Scales in male and female adolescents. It was hypothesized that all MMPI-A-RF Internalizing and Externalizing Scales will reach at least partial scalar invariance. This hypothesis was largely supported for the Internalizing Scales, but less so for the Externalizing scales. Five of the nine Internalizing Scales (Anger Proneness, Anxiety, Self-Doubt, Specific Fears, and Stress/Worry) obtained full measurement invariance. The Behavior Restricting Fears Scale was dropped from further analyses due to the Heywood case in the girls' sample. For the remaining three Internalizing Scales, two (Inefficacy and Obsessions/Compulsions) met partial measurement invariance. The last scale (Helplessness/Hopelessness) reached configural invariance but not partial measurement invariance. Three of the six Externalizing Scales (Antisocial Attitudes, Negative Peer Influence, and Negative School Attitudes) obtained full measurement invariance. For the remaining three Externalizing Scales, two (Aggression and Conduct Problems) met configural invariance but not partial measurement invariance. The final Externalizing Scale (Substance Abuse) did not meet configural invariance.

Internalizing Scales

As previously indicated, configural invariance was observed for eight of the nine Internalizing Scales tested, indicating that intercorrelations of the items are explained well between genders, and the same items are associated with a singular latent factor in each group. For scales that reached full measurement invariance (i.e., ANP, AXY, SFD, SPF, and STW), implications are made that the MMPI-A-RF is measuring the same constructs in the same way

between genders, thus allowing for meaningful comparisons between the two groups and further supporting the use of nongendered norms. For scales that reached partial measurement invariance (i.e., NFC and OCS), the invariant items maintained consistent interpretation and significance across groups, meaning the invariant items have equivalent expected scores across male and female samples for any given level of the underlying latent factor. However, there were some items that were noninvariant, which indicates that the expected scores for these items were not equivalent across gender samples.

The BRF Scale failed to meet configural invariance which means the pattern of loadings of items on the latent factor differs between boys and girls. There are generally two options researchers can take when configural noninvariance is encountered: (1) omitting items to redefine the construct or (2) assume the construct is noninvariant and discontinue invariance testing (Putnick & Bornstein, 2016). Since the MMPI-A-RF is an established testing instrument and the BRF Scale only has three items, dropping an item and increasing the number of items for the factor are not possible solutions until the test is revised.

Though the BRF Scale was dropped for further measurement invariance analyses, the items were examined for potential explanations of the noninvariance. First, it is notable that the internal consistency coefficients were low in the boys' sample (α = .39) and girls' sample (α = .31). These values corresponded to the standard error of measurement values of 8 and 9 T-score points, respectively, which were the highest SEM values of all Internalizing Scales. In examining the factor loadings, there were large differences between boys and girls. The boys' sample had factor loadings that ranged from .483 to .636 and the girls had one factor loading considered below the accepted range of .30 or .40 (Brown, 2015) at a value of .201. Furthermore, item 123 was greater than 1.0 for the girls' sample thus indicating a Heywood case. As previously

mentioned, Bollen (1987) recommended multiple solutions to address Heywood cases, including dropping the problematic indicator, obtaining a larger sample, increasing the number of indicators per factor, fixing the improper estimate to a plausible value, or using an inequality restriction. For the purposes of the study, fixing the estimate or restricting the inequality would limit the implications and further measurement invariance testing.

The Helplessness/Hopelessness (HLP) Scale reached configural invariance indicating the basic organization of the construct is supported in the two groups, that is the items reflected the same latent variable between groups. With evidence of configural invariance, full measurement invariance was examined but not obtained, indicating there are differences in factor loadings, thresholds, or both. Partial invariance testing was conducted by relaxing equality constraints on the factor loading and threshold of one item at a time while simultaneously keeping equality constraints on the remaining items within the factor. However, further examination of the scale indicated partial measurement invariance was not obtained because the majority (Vandenberg & Lance, 2000) of the items (nine out of ten items; 90%) were noninvariant apart from item 60. Throughout this document, only item numbers are presented because the test publisher does not permit the reproduction of MMPI-A-RF item content in theses and dissertations.

Although there is no systematic method to determine if the noninvariance is due to differences in factor loadings, thresholds, or both (Wang et. al., 2020), differences in factor loadings, thresholds, and item endorsements were examined in an attempt to explain the noninvariance. Of these, raw item endorsement frequencies are the least informative because any differences in item endorsement frequencies (and corresponding scale T-scores) could simply reflect actual differences between the groups rather than test bias. Therefore, these descriptions are only provided within the context of already established MGCFA noninvariance. The boys'

and girls' sample had relatively high and consistent factor loadings ranging from .523 to .895 in the boys' sample and .449 to .877 in the girls' sample. This would suggest each item is contributing to the latent construct to a similar degree between groups. Though items 169 and 239 in the girls' sample had lower factor loadings in comparison to the remaining items in each sample, they were still considered acceptable (Brown, 2015).

In examining the item thresholds, there were differences between the boys' and girls' samples for a majority of the items. These results indicate that while girls may endorse these items more often, they are not related to increased latent levels of helplessness or hopelessness in the same way that they are for boys.

Although simple differences in raw item endorsement frequencies alone do not indicate that a measure is biased, the differences between the boys' and girls' samples could potentially be explained by a myriad of possibilities or due to a combination of social, psychological, and biological factors. In general, adolescent girls report higher levels of internalizing symptoms, including feelings of hopelessness (Kann et al., 2018;) compared to adolescent boys who engaged in more externalizing behaviors (Memmott-Elison et al., 2020). Studies have also identified that the prevalence rate of depressive symptoms and depressive mood is elevated and more common in female adolescents (Shorey et al., 2021) compared to males who displayed greater difficulties in concentration and psychomotor agitation/hindrance (Crockett et al., 2020). Crockett et al. (2020) further revealed that higher levels of dysfunctional thoughts and perceived social support were associated with subthreshold depression in females and males.

Adolescent girls may face more considerable societal pressures regarding their appearance, performance, and behaviors (Silva et al., 2020), which may lead to feelings of inadequacy and low self-esteem if they do not measure up to the standards set forth by the

onslaught of social media. Not fitting into the social stereotype may lead to social or societal exclusion which would contribute to feelings of loneliness. The difficulties encountered through social exclusions could also lead to hopelessness with their social connections. The traditional gender roles may contribute to adolescent girls' perception of their opportunities and capabilities, leading them to believe they are less competent or valuable than boys. Internalizing these messages could enforce feelings of hopelessness about their future and impose helplessness in changing it.

Research indicates heredity may play a role in explaining why adolescent girls may be more susceptible to depression (Zhao et al., 2020). The literature also suggests that the biological changes and fluctuations of hormones in adolescents can affect their psychological well-being depending on when these hormonal changes begin. For example, early sexual maturation in girls may result in higher risks of depression (Ge et al., 2001). These body changes can also lead to body image concerns (Silva et al., 2020). Coupled with the influx of social media portrayals, influencers with unrealistic body standards, and beauty hacks, adolescent girls may develop feelings of helplessness and hopelessness in changing their appearance leading to poorer body image. Adolescent girls may also be more likely to experience trauma through sexual harassment or assault, and without proper support or healthy coping skills, these experiences may significantly contribute to feelings of helplessness or hopelessness. All of this information notwithstanding, additional research will be required to evaluate why the content of specific MMPI-A-RF items relates differently to the latent variable for boys versus girls.

The Inefficacy (NFC) Scale reached configural invariance indicating the basic organization of the construct is supported in the two groups, that is, the items reflect the same latent variable between groups. With evidence of configural invariance, full measurement

invariance was examined but not obtained, indicating there are differences in factor loadings, thresholds, or both. Partial invariance testing was conducted by relaxing equality constraints on the factor loading and threshold of one item at a time while simultaneously keeping equality constraints on the remaining items within the factor. Further examination of the scale indicated partial measurement invariance was obtained because at least half of the items achieved invariance. Items 112 and 224 were the two items found noninvariant. In comparing observed T-score means, the girls' sample reported more symptoms of inefficacy compared to the boys' sample with a medium effect size (Cohen's d = -.543) which is consistent with the literature discussed below.

Factor loadings, thresholds, and item endorsement frequencies were assessed in an attempt to shed light on the noninvariance of the two items. In examining the factor loadings, the boys' and girls' sample had relatively high factor loadings ranging from .648 to .826 in the boys' sample and .632 to .872 in the girls' sample. This would suggest each item is contributing to the latent construct to a similar degree between groups. However, in examining the thresholds and converting them to probabilities using a z-table, items 112, 159, and 224 had larger probability differences between the boys' and girls' samples. This suggests the girls' sample may experience or endorse these items more, but those items may not be related to increased levels of inefficacy in the same way that they are for boys. Although differences in raw item endorsement frequencies alone do not indicate that a measure is biased, the girls' sample had statistically significantly ($p \leq .001$) higher item endorsement frequencies compared to the boys' sample for all four items with phi coefficients ranging from .132 to .237.

Potential reasons why the observed inefficacy scores could be interpreted differently in the girls' sample could be due to the societal expectations of gender stereotypes (Hoffmann et

al., 2004). Adolescent girls may endorse giving up more quickly because they receive societal messages through social media and family members that prioritize stereotypical traits such as nurturing, empathy, and compliance more than the traits that would encourage them to take risks or pursue goals inconsistent with gender stereotypes (Ward & Grower, 2020). This in combination with a lack of visible female representation within certain fields, could alter their perceptions and confidence in their abilities, thus inhibiting them from taking risks or pursuing their goals. Additionally, adolescent girls could be facing hardships from their own peers. During adolescent development, a shift occurs from familial relationships to the development of relative independence and involvement with peers. Specifically, for adolescent girls, their peer relationships place a greater value on their friendships and social support (American Psychiatric Association, 2013). This heightened peer influence could contribute to their inefficacy if they perceive their peers as more successful, knowledgeable, or capable than themselves. Their social comparison and fear of judgement could inhibit them from taking risks or encourage them to give up more easily, thereby increasing their feelings of inadequacy and self-doubt. Nevertheless, the key point for future research will be to evaluate why certain items relate differently to the latent variable for boys versus girls.

The Obsessions/Compulsions (OCS) Scale reached configural invariance indicating the basic organization of the construct is supported in the two groups, that is the items reflected the same latent variable between groups. With evidence of configural invariance, full measurement invariance was examined but not obtained, indicating there are differences in factor loadings, thresholds, or both. Partial invariance testing was conducted by relaxing equality constraints on the factor loading and threshold of one item at a time while simultaneously keeping equality constraints on the remaining items within the factor. Further examination of the scale indicated

partial measurement invariance was obtained because at least half of the items achieved invariance. Items 15 and 139 were the two items found noninvariant.

Factor loadings, thresholds, and item endorsement frequencies were examined in an attempt to explain the noninvariance observed in the two items. In examining the factor loadings, the boys' sample was much lower for item 15 (.389) and item 139 (.393) compared to the girls' sample (.655 and .421 respectively). For the boys' sample, item 15 had a low factor loading, a high threshold, and little correlation with other items, indicating it may be exhibiting item bias and may not be a strong indicator for the OCS construct for the boys' sample. In examining the item thresholds, item 221 has an observable difference between the boys' and girls' sample. After converting the item thresholds to probabilities using the z-table, item 221 indicated the largest difference in probability between the boys' and girls' samples where there is a 46.5 % (25.4 % for girls' sample) probability that item 221 = 0 and a 53.5 % (74.6% for girls' sample) probability that item 221 = 1 in the boys' sample at the same level of latent obsession/compulsions. This suggests that girls may endorse this item more, but the item may not be related to increased levels of the OCS construct. Furthermore, in the OCS Scale, the girls' sample had statistically significantly higher item endorsement frequencies for all items compared to the boys' sample. Below is a brief description of some of the literature in this area. However, as noted earlier, item endorsement frequency differences or observed T-score differences are not in of themselves indicative of bias as any differences could reflect actual differences between the groups on the underlying construct.

The girls' sample may have higher item endorsement frequencies on these items because of the increased anxiety and uncertainty that many adolescent girls face from physical, emotional, and social changes (Benton et al., 2021). Furthermore, adolescents are often faced

with significant stressors such as academic pressures, social challenges, and changes in family dynamics. The increased anxiety could lead to the rumination of cultural and social factors brought about by peers, family, and stereotypical gender roles (Benton et al., 2021). Girls may also experience societal pressures to excel or be perfect academically, socially, and in their appearance. The perfectionistic tendencies could additionally lead to obsessive thoughts about making mistakes, being lesser than their peers or family expectations, and criticism leading to compulsive behaviors such as reassurance-seeking or repetitive thoughts aimed at achieving unrealistic standards. Superstitious thinking could serve as a coping mechanism to manage stress and anxiety.

The aforementioned five scales (Anger Proneness [ANP], Anxiety [AXY], Self-Doubt [SFD], Specific Fears [SPF], and Stress/Worry [STW]) reached full measurement invariance indicating the mean differences in the latent construct captures the mean differences in the shared variance of the items and that there is equivalence of the item loadings on the latent factors. In comparing the latent means of boys and girls for the five scales, the girls' sample had statistically significantly higher means than the boys' sample in the ANP, AXY, SFD, and STW Scales. However, no statistically significant difference between the two genders was observed in the SPF Scale. In general, these findings are consistent with the literature where girls often experience higher rates of internalizing behaviors in comparison to boys (American Psychiatric Association, 2013; Romano et. al., 2001).

Externalizing Scales

Configural invariance was observed for five of the six Externalizing Scales indicating that intercorrelations of the items are explained well between genders, and the same items are associated with a singular latent factor in each group. For the three scales (i.e., ASA, NPI, and

NSA) that reached full measurement invariance, implications are made that the MMPI-A-RF is measuring the same constructs in the same way between genders, thus allowing for meaningful comparisons between the two groups. For the remaining two scales (i.e., AGG and CNP), while configural invariance was obtained, partial measurement invariance could not be achieved because the majority of items in each scale were noninvariant indicating the expected scores of these items were not equivalent across gender samples.

The Aggression (AGG) Scale reached configural invariance indicating the basic organization of the construct is supported in the two groups, that is the items reflected the same latent variable between groups. With evidence of configural invariance, full measurement invariance was examined but not obtained, indicating there are differences in factor loadings, thresholds, or both. Partial invariance testing was conducted by relaxing equality constraints on the factor loading and threshold of one item at a time while simultaneously keeping equality constraints on the remaining items within the factor. However, further examination of the scale indicated partial measurement invariance was not obtained because the majority (Vandenberg & Lance, 2000) of the items (seven out of eight items; 87.5%) were noninvariant. Item 186 was the only item found invariant.

To account for the noninvariance, factor loadings, thresholds, and item endorsement frequencies were evaluated. The factor loadings in the boys' and girls' sample were relatively high ranging from .546 to .812 in the boys' sample and .526 to .882 in the girls' sample. These factor loadings suggest each item is contributing to the latent construct to a similar degree between groups. However, a few items yielded very large thresholds. Item 36, for example, was 1.447 in the boys' sample and 1.282 in the girls' sample. Furthermore, in examining the item thresholds, the largest differences, after converting the thresholds to probabilities with the z-

table, were between the boys' and girls' samples for items 186, 233, and 240. These results suggest that boys may experience or endorse these items more, but those items may not be related to increased levels of latent aggression relative to girls. For example, converting item 186 thresholds to probabilities using a z-table, results would mean that there is a 76.9 % (87.7 % for girls' sample) probability that item 186 = 0 and a 23.1 % (12.3% for girls' sample) probability that item 186 = 1 in the boys' sample at the same level of latent aggression. Not surprisingly, when examining raw item endorsement frequencies for items 186 and 240, the boys' sample had statistically significantly higher endorsement frequencies.

Potential reasons why these items may be endorsed more frequently may include the substantial increase in testosterone levels in boys, which has been associated with increased aggression and dominance-related behaviors (Archer, 2006). The increase in physical assaults could also be due to the differences in brain structure, particularly in the amygdala and prefrontal cortex, which are involved in emotion regulation and impulse control (Coccaro et al., 2011).

Cultural norms, socialization, peer influences, and stereotypes may also be factors that contribute to the differences in how the boys' sample endorsed higher rates of aggression compared to the girls' sample (Perry & Pauletti, 2011). Within some societies, the idealistic behaviors and attitudes associated with masculinity are generally attributes that are in relation to power and control (Connell & Messerschmidt, 2005). Research has suggested that those who subscribe to more traditional masculine roles and beliefs, view asking for help or talking about their feelings as a more feminine activity (Vogel et al., 2011). This toxic masculinity is further amplified through entertainment and social media, thereby normalizing and reinforcing aggressive behaviors as a means of resolving conflicts or obtaining goals. These aggressive behaviors may then also be further shaped by their peer influences. Boys may be taught not to

express their vulnerabilities or sensitivities but to suppress or mask their emotions, leading to frustration and anger that would, in turn, manifest as physical aggression (Zahn-Waxler et al., 2008). Through frequent occurrences, boys may begin to resort to aggression as a coping mechanism. However, here too, simple differences in mean scores or endorsement frequencies do not in of themselves indicate that a measure is biased. Further research with different populations will be necessary to evaluate the utility of these items across genders (especially given that the factor loadings are acceptable in both groups in the present study).

The Conduct Problems (CNP) Scale reached configural invariance indicating the basic organization of the construct is supported in the two groups, that is the items reflected the same latent variable between groups. With evidence of configural invariance, full measurement invariance was examined but not obtained, indicating there are differences in factor loadings, thresholds, or both. Partial invariance testing was conducted by relaxing equality constraints on the factor loading and threshold of one item at a time while simultaneously keeping equality constraints on the remaining items within the factor. However, further examination of the scale indicated partial measurement invariance was not obtained because all of the items were noninvariant.

An analysis of factor loadings, thresholds, and item endorsement frequencies were conducted to elucidate the noninvariance. In examining the factor loadings, the boys' and girls' sample had relatively high and consistent factor loadings ranging from .482 to .892 in the boys' sample and .492 to .872 in the girls' sample. This would suggest each item is contributing to the latent construct to a similar degree between groups. Though item 238 in the boys' sample and item 148 in the girls' sample had lower factor loadings in comparison to the remaining items in each sample, they were still considered acceptable (Brown, 2015).

A few items yielded very large thresholds. Item 148, for example, was 1.159 in the boys' sample and 1.326 in the girls' sample. Furthermore, in examining the item thresholds, there were greater differences in the probabilities between the boys' and girls' samples for items 33, 88, and 127. These findings suggest that endorsement of these items may not be related to increased levels of the latent variable conduct problems for boys in the same way that they are for girls. For example, converting item 88 thresholds to probabilities using a z-table, results would mean that there is a 55.5 % (82.6 % for girls' sample) probability that item 88 = 0 and a 44.5 % (17.4% for girls' sample) probability that item 88 = 1 in the boys' sample. Not surprisingly, when examining item endorsements for these three items (33, 88, 127), the boys' sample had statistically significantly higher endorsement frequencies with near medium effect sizes.

Similar to the AGG Scale, these differences may be partially explained by the higher levels of testosterone that predispose boys to engage in behaviors that violate rules or laws due to increased aggression, risk-taking, and impulsive behaviors (Archer, 2019). In the school setting, impulsivity may be related to attention deficit hyperactivity disorder, a diagnosis more common in boys than girls (American Psychiatric Association, 2013) and may be associated with conduct problems. Boys may also have lesser developed emotional regulation skills making them more prone to externalizing behaviors (Lahey et al., 2000). Gender norms and stereotypes, as well as peer pressure, may also play a role in shaping adolescent boys' behaviors by encouraging boys to seek status and dominance by challenging authority and displaying more toxic masculinity (Card & Little, 2006). This could lead to disruptive and delinquent behaviors that increase rates of conduct problems in the schools and with the law.

Boys may also have greater difficulties adapting to the structured and sedentary nature of traditional school environments, which would exacerbate conduct problems and potentially

escalate into legal issues. In schools, teachers may have different expectations and perceptions of boys' behavior, potentially leading to a bias in identifying and reporting conduct problems among boys. Within the familial structure, boys who grow up in single-parent households, experience inconsistent or harsh parenting, or from lower socioeconomic status may face additional stressors and environmental risk factors that can increase the likelihood of engaging in delinquent behaviors (Hoeve et al., 2012). In relation to the law, boys have a higher crime commitment rate than girls (McCabe et al., 2002). Furthermore, the degree of exposure to violence and crime, and lack of visible positive role models who demonstrate prosocial behaviors, can contribute to the development of negative conduct problems that could ultimately lead to legal issues (Kalvin & Bierman, 2017). Psychometrically, the CNP Scale also did not reach measurement invariance as shown in Park's (2018) study that examined measurement invariance across adolescent cultures (American, Korean). But again, the core issue for future research will be to evaluate why the specific content of noninvariant items functions differently for boys versus girls holding the latent variable constant.

The Substance Abuse Scale (SUB) is the only Externalizing Scale that did not reach full measurement, partial measurement, or configural invariance. No invariance testing could be computed for this scale because it never reached an acceptable model fit for the boys' sample indicating interpretation of items between boys and girls was different and the factor structure is not consistent across groups. With these psychometric differences, comparisons of the means could be problematic because the underlying constructs are not measured equivalently. In examining the items, the factor loadings range from .753 (Item 72) to .947 (Item 43) for the girls' sample and .590 (Item 72) to .985 (Item 43) for the boys' sample. In examining the items independently, item 72 was different from the other items in that it relates to how others may

perceive the adolescent as opposed to the adolescent's personal report. It is possible that boys interpret this item differently resulting in a poor factor structure or a different underlying construct compared to the girls' sample. Additionally, very few adolescents endorsed this item (less than 5% for each gender).

The aforementioned three scales (Antisocial Attitudes [ASA], Negative Peer Influence [NPI], and Negative School Attitudes [NSA]) reached full measurement invariance indicating the mean differences in the latent construct capture the mean differences in the shared variance of the items and that there is equivalence of the item loadings on the latent factors. For the ASA, NPI, and NSA Scales, full measurement invariance was found, and they had similar latent means in the boys' and girls' samples. That is, no statistically significant difference ($p < .01$) was found between the two genders.

Strengths and Implications

Broadly, no published studies have examined measurement invariance between males and females for any of the MMPI-A-RF scales, including the Specific Problem Scales. Therefore, this study makes an important contribution to the literature as being the first to examine the measurement invariance between boys and girls for the MMPI-A-RF Externalizing and Internalizing Scales. This study identified through the use of a contemporary psychometric method that many of the existing MMPI-A-RF Externalizing and Internalizing Specific Problems Scales obtained measurement invariance in varying degrees.

This study has identified several scales where test developers, researchers, and practitioners should be cognizant of the influence of noninvariant items (i.e., the HLP, NFC, OCS, AGG, and CNP Scales). Some potential solutions outlined for non-invariance include (a) deleting the noninvariant items, (b) using all the items and assuming any differences are small

and do not influence the results, (c) interpret the scores independently and preclude group comparisons, or (d) simply avoid using the scale (Sass, 2011). Option "d" could be applied, but it would limit the clinical and practical data that may be beneficial for treatment. Option "b" could potentially be feasible if the scale reached partial invariance. Option "c" could be useful in maintaining test comparability of prior versions but would require the test developers to create test scores that would be computed based upon the noninvariance of group membership – an arduous task that can only be justified by having further psychometric support from future studies. Importantly, the test publisher is in early discussions about revisiting the test, and the present results could be considered along with other updates.

Limitations and Future Directions

Though this study makes important contributions to the literature, several limitations must be noted. Recruitment methods may limit generalization from the results. Due to the nature of the statistical analyses, the Midwest sample size was too small. Therefore, the sample had to include data from not only the Midwest sample but also from the Pearson sample. Due to the merging of datasets, further descriptive analyses could not be completed for the entire sample as the individual datasets did not include all extraneous variables (e.g., diagnoses, reason for testing), limiting the ability of the researcher to have proportionate data samples. Consequently, this study did not include alternative covariates or controls for other confounding or contributing variables in the analyses.

In addition to including the two datasets, there were not sufficient MMPI-A-RFs to complete the analysis. Therefore, MMPI-As from each dataset were converted to MMPI-A-RFs to reach a sufficient data size. The MMPI-A can easily be rescored as an MMPI-A-RF since it uses the same items; nonetheless, there could potentially be complications due to the differences

in the testing measures (e.g., test item length). However, research with adults has demonstrated that there are comparable scores between the MMPI-2 and MMPI-2-RF on the 51 scales (Tellegen & Ben-Porath, 2008/2011). The archival data were from an extended time period, January 2014 through April 2023 from the Midwest sample and June 2018 to October 2023 from the Pearson sample. Though prior research indicated the original sample was comparable to today's adolescents (Archer et al., 2016), it is still possible that the difference could have had a minor impact on this data analysis, and it would be negligent not to identify it as a limitation of the study.

In terms of additional limitations, The Midwest sample may not accurately represent other regions of the United States. Although the Pearson sample may be more geographically diverse, no information is available regarding where the measures were administered. Therefore, the overall sample is heterogenous in terms of setting and geography, and the reasons for conducting psychological assessments in the Pearson sample are also unknown. Future studies should examine measurement invariance in more well-defined samples.

The psychometric properties could also be a limitation of the study as it pertains to the exploratory nature of the partial invariance analyses (Brown, 2015). Some scales had a larger number of parameters, and it is possible that some of the parameters differed by chance. Additionally, the larger sample size can inflate the chi-square which is sensitive to sample size. To address these concerns, Vandenberg and Lance's (2000) approach was employed, in which partial invariance should not be utilized when a large number of items were found to be noninvariant. Indeed, when partial invariance was not obtained, comparisons between the genders were not calculated. However, potential reasons to explain the noninvariance between the genders were made in an attempt to justify why there may have been differences observed.

Special limitations existed for the BRF Scale. When the girls' and boys' samples were combined, the scale did not indicate model misspecification. However, when CFAs were conducted separately by gender, the girls' sample did exhibit model misspecification, specifically through item 123 – a Heywood case. Though the model fit perfectly (just-identified), the composition of the data was not in line with the model, and there is homogeneity in the factor structure. It is possible that the misspecification is explained by the small number of indicators. However, careful consideration must be taken when interpreting this scale, particularly for adolescent girls.

Particular attention may be brought to the SUB Scale, which did not reach configural invariance, thus indicating that the basic structure of the scale is not equivalent across groups. This would necessitate further examination of the scale's items and cautious interpretations of results across genders. Implications for the SUB Scale would include the test developers needing to reexamine and possibly modify the scale to ensure it measures the same construct across different groups in newer editions of the instrument. While removing or adding items is not an option for the current MMPI-A-RF, perhaps re-examination of the SUB Scale items or redefining the factors can shed more light on potential solutions for future editions to the adolescent instrument. However, further research would need to be conducted to not only gain clarification, but also to determine if this scale continues to demonstrate psychometric difficulties.

Future research to address and improve the BRF and SUB Scales could include reviewing and possibly modifying the current items, increasing the number of items in each scale, and identifying a more appropriate and accurate factor structure. Though the recommended minimum number of items for a single factor model is three items, increasing the number would allow for more variation and identification of the factor. In addition, having more items would allow test

developers the opportunity to identify the factor structure of the scale. Further research could examine how agoraphobia relates to adolescents to better assess the utility of the BRF Scale. Future research studies should also focus on the SUB Scale to identify how this scale could be revised and improved. One possibility is that there may be separate constructs (e.g., alcohol, other substances) underlying substance abuse. Until the next revision of the MMPI-A-RF is released, clinicians and practitioners may still utilize the BRF and SUB Scales, though caution is recommended. The MMPI-A-RF Specific Problem Scales can be examined independently of the Higher Order Scales. Therefore, if elevations on the BRF and/or SUB Scales are observed, it could lead to clarification questions during the feedback session with the client. The instrument and Scales alone do not provide a definitive answer to assessment questions. Rather, the MMPI—A-RF should always be used in conjunction with other sources of data such as clinical judgment, a feedback session, a clinical interview, other psychometric measures, and additional collateral data as appropriate.

Future studies should further investigate the measurement invariance of multiple scales of the MMPI-A-RF with multiple differing populations, including ethnicity and age. It would also better serve the testing instrument if measurement invariance could be examined in differing populations within setting-specific samples (e.g., medical, forensic, or inpatient).

This study is also limited to comparisons between boys and girls and did not explore individuals who had identified as transgender. The decision of examining boys and girls was mainly due to the instrument only providing the two options (male and female) and not having indications of the adolescent's identity with the Pearson sample. Given the natural fluidity of gender as a spectrum, as opposed to strictly binary, future studies may wish to examine measurement invariance based on more categories of gender identity beyond the traditional

binary identities. In addition, future revisions to the instrument should focus on incorporating contemporary definitions of gender.

REFERENCES

Alperin, J. J., Archer, R. P., & Coates, G. D. (1996). Development and effects of an MMPI-A K-correction procedure. *Journal of Personality Assessment, 67*(1), 155 – 156. https://doi.org/10.1207/s15327752jpa6701_12

Altinay, M. (2020). Transgender: The T in LGBTQ²IAPA. In P. Levounis & E. Yarbrough, (Eds), *Pocket guide to LGBTQ mental health. Understanding the spectrum of gender and sexuality* (pp. 61-87). American Psychiatric Association Publishing.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* American Educational Research Association.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). American Psychiatric Association.

Anastasi, A., & Urbina, S. (1997). Psychological testing (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Archer, J. (2006). Testosterone and human aggression: An evaluation of the challenge hypothesis. *Neuroscience & Biobehavioral Reviews, 30*(3), 319–345. https://doi.org/10.1016/j.neubiorev.2004.12.007

Archer, J. (2019). The reality and evolutionary significance of human psychological sex differences. *Biological Reviews, 94*(4), 1381–1415. Portico. https://doi.org/10.1111/brv.12507

Archer, R. P. (1984). Use of the MMPI with adolescents: A review of salient issues. *Clinical Psychology Review, 4,* 241 – 251. https://doi.org/10.1016/0272-7358(84)90002-3

Archer, R. P. (1987). *Using the MMPI with adolescents.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Archer, R. P. (2005). *MMPI-A: Assessing adolescent psychopathology* (3rd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Archer, R. P. (2016). Introducing the Minnesota Multiphasic Personality Inventory-Adolescent-Restructured Form (MMPI-A-RF). *European Scientific Journal,* [e-ISSN 1857-7431]

Archer, R. P. (2017). *Assessing adolescent psychopathology: MMPI-A/MMPI-A-RF* (4th ed.). New York: Routledge Press.

Archer, R. P., & Handel, R. W. (2019, June 13). *Using the MMPI-A-RF with adolescents* [Conference session]. MMPI Workshops & 54th Annual Symposium, Minneapolis, MN.

Archer, R. P., Handel, R. W., Ben-Porath, Y. S., & Tellegen, A. (2016). *Administration, Scoring, Interpretation, and Technical Manual.* Minneapolis, MN: University of Minnesota Press

Archer, R.P., Krishnamurthy, R., Kaufman, A.S., & Kaufman, N.L. (Eds.). (2002). Essentials of MMPI-A Assessment. New York, NY: John Wiley & Sons, Inc.

Archer, R. P., & Newsom, C. R. (2000). Psychological test usage with adolescent clients: Survey update. *Assessment, 7*(3), 227 – 235. https://doi:10.1177/107319110000700303

Ben-Porath, Y. S., & Archer, R. P. (2014). The MMPI instruments. In R. P. Archer, & S. R. Smith, (Eds), *Personality assessment (2nd ed.)* (pp. 89-146). New York, NY: Routledge.

Ben-Porath, Y. S. (2012). *Interpreting the MMPI-2-RF*. Minneapolis, MN: University of Minnesota Press.

Ben-Porath, Y. S., & Forbey, J. D. (2003) *Non-gendered norms for the MMPI-2*. Minneapolis, MN: University of Minnesota Press.

Ben-Porath, Y. S., & Tellegen, A. (2008/2011). *MMPI-2-RF (Minnesota Multiphasic Personality Inventory-2 Restructured Form) manual for administration, scoring, and interpretation*. Minneapolis, MN: University of Minnesota Press.

Ben-Porath, Y. S., & Tellegen, A. (2020). *Minnesota Multiphasic Personality Inventory – 3 (MMPI-3): Manual for administration, scoring, and interpretation.* University of Minnesota Press.

Benton, T. D., Boyd, R. C., & Njoroge, W. F. M. (2021). Addressing the global crisis of child and adolescent mental health. *JAMA Pediatrics, 175*(11), 1108. https://doi.org/10.1001/jamapediatrics.2021.2479

Berenbaum, S. A., Martin, C. L., & Ruble, D. N. (2008). Gender development. In W. Damon & R. M. Lerner (Eds.). *Child and adolescent development: An advanced course.* John Wiley & Sons.

Bollen, K. A. (1987). Outliers and improper solutions: A confirmatory factor analysis example. *Sociological Methods and Research*, 15(4), 375 – 384. https://doi.org/10.1177/0049124187015004002

Brown, T. A. (2015). *Confirmatory Factor Analysis for Applied Research* (2nd ed.). The Guilford Press.

Bryant, W. T., Livingston, N. A., McNulty, J. L., Choate, K. T., & Brummel, B. J. (2021). Examining Minnesota Multiphasic Personality Inventory-2-Restructured Form (MMPI-2-RF) scale scores in a transgender and gender diverse sample. *Psychological Assessment, 33*(12), 1239–1246. https://doi:10.1037/pas0001087

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

Butcher, J. N., Williams, C. L., Graham, J. R., Archer, R. P., Tellegen, A., Ben-Porath, Y. S., & Kaemmer, B. (1992). Minnesota Multiphasic Personality Inventory-Adolescent Version (MMPI-A): Manual for administration, scoring and interpretation. Minneapolis, MN: University of Minnesota Press.

Capwell, D. F. (1945). Personality patterns of adolescent girls: II. Delinquents and nondelinquents. *Journal of Applied Psychology, 29*(4)*,* 256 – 265. https://doi.org/10.1037/h0054701

Card, N. A., & Little, T. D. (2006). Proactive and reactive aggression in childhood and adolescence: A meta-analysis of differential relations with psychosocial adjustment. *International Journal of Behavioral Development, 30*(5), 466–480. https://doi.org/10.1177/0165025406071904

Christie, D., & Viner, R. (2005). ABC of adolescence: Adolescent development. *British Medical Journal, 330,* 301 – 304.

Coccaro, E. F., Sripada, C. S., Yanowitch, R. N., & Phan, K. L. (2011). Corticolimbic function in impulsive aggressive behavior. *Biological Psychiatry, 69*(12), 1153–1159. https://doi.org/10.1016/j.biopsych.2011.02.032

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155 - 159. https://doi.org/10.1037/0033-2909.112.1.155

Connell, R. W., & Messerschmidt, J. W. (2005). Hegemonic masculinity. *Gender & Society, 19*(6), 829–859. https://doi.org/10.1177/0891243205278639

Crockett, M. A., Martínez, V., & Jiménez-Molina, Á. (2020). Subthreshold depression in adolescence: Gender differences in prevalence, clinical features, and associated factors. *Journal of Affective Disorders, 272*, 269–276. https://doi.org/10.1016/j.jad.2020.03.111

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods, 1*(1),16–29. https://doi.org/10.1037/1082-989x.1.1.16

Curtis, A. C. (2015). Defining adolescence. *Journal of Adolescent and Family Health, 7*(2), 1 – 40.

Dahlstrom, W. G., Welsh, G. S., & Dahlstrom, L. E. (1972). *An MMPI handbook: Vol. 1. Clinical interpretation* (Rev. ed.). Minneapolis, MN: University of Minnesota Press.

Desa, D. (2018). Understanding non-linear modeling of measurement invariance in heterogeneous populations. *Advances in Data Analysis and Classification, 12*(4). 841-865. https://doi.org/10.1007/s11634-016-0240-3

Dong, Y., & Dumas, D. (2020). Are personality measures valid for different populations? A systematic review of measurement invariance across cultures, gender, and age. *Personality and Individual Differences, 160,* p 109956. https://doi.org/10.1016/j.paid.2020.109956

Finney, S. J., & DiStefano, C. (2013). Non-normal and categorical data in structural equation modeling. In G. R. Handcock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed; pp. 269- 301). Charlotte, NC: Information Age Publishing, Inc.

Furr, R. M. (2018). *Psychometrics an introduction* (3rd ed.). Sage.

Ge, X., Conger, R. D., & Elder, G. H. (2001). Pubertal transition, stressful life events, and the emergence of gender differences in adolescent depressive symptoms. *Developmental Psychology, 37*, 404 – 417. https://doi.org/10.1037/0012-1649.37.3.404

Graber, J. A., Lewinsohn, P. M., Seeley, J. R., & Brooks-Gunn, J. (1997). Is psychopathology associated with the timing of pubertal development? *Journal of the Academy of Child and Adolescent Psychiatry, 36*, 1768 – 1776. https://doi.org/10.1097/00004583-199712000-00026

Graham, J. R. (2012). *MMPI-2 assessing personality and psychopathology (5th ed).* New York, NY: Oxford University Press.

Groth-Marnat, G. (2009). Minnesota multiphasic personality inventory. In G. Groth-Marnat, (Ed), *Handbook of psychological assessment (5th ed.)* (pp. 207-294). Hoboken, New Jersey: John Wiley & Sons, Inc.

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. https://doi.org/10.1080/00223891.2012.754360

Handel, R. W. (2016). An introduction to the Minnesota Multiphasic Personality Inventory – Adolescent – Restructured Form (MMPI-A-RF). *Journal of Clinical Psychology Medical Settings, 23,* 361 – 373. https://doi10.1007/s10880-016-9475-6

Harkness, A. R., McNulty, J. L., & Ben-Porath, Y. S. (1995). The personality psychopathology five (PSY-5): Constructs and MMPI-2 scales. *Psychological Assessment, 7*(1), 104–114. https://doi:10.1037/1040-3590.7.1.104

Harvill, L. M. (1991). An NCME instructional module on standard error of measurement. *Educational Measurement: Issues and Practice, 10*(2), 33–41. https://doi:10.1111/j.1745-3992.1991.tb00195.x

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

Hathaway, S. R., & Monachesi, E. D. (1963). *Adolescent personality and behavior: MMPI patterns of normal, delinquent, dropout, and other outcomes.* University of Minnesota Press.

Hilts, D., & Moore, J. M., Jr. (2003). Normal range MMPI-A profiles among psychiatric inpatients. *Assessment, 10* (3), 266-272. https://doi.org/10.1177/1073191103255494

Hoeve, M., Stams, G. J. J. M., van der Put, C. E., Dubas, J. S., van der Laan, P. H., & Gerris, J. R. M. (2012). A Meta-analysis of attachment to parents and delinquency. *Journal of Abnormal Child Psychology, 40*(5), 771–785. https://doi.org/10.1007/s10802-011-9608-1

Hoffmann, M. L., Powlishta, K. K., & White, K. J. (2004). An examination of gender differences in adolescent adjustment: The effect of competence on gender role differences in symptoms of psychopathology. *Sex Roles*, *50*(11/12), 795–810. https://doi.org/10.1023/b:sers.0000029098.38706.b1

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1-55. https://doi.org/10.1080/10705519909540118

Janus, M. D., Tolbert, H., Calestro, K., & Toepfer, S. (1996). Clinical accuracy ratings of MMPI approaches for adolescents: Adding ten years and the MMPI-A. *Journal of Personality Assessment, 67*(2), 364-383. https://doi.org/10.1207/s15327752jpa6702_11

Kalvin, C. B., & Bierman, K. L. (2017). Child and adolescent risk factors that differentially predict violent versus nonviolent crime. *Aggressive Behavior, 43*(6), 568–577. Portico. https://doi.org/10.1002/ab.21715

Kann, L., McManus, T., Harris, W. A., Shanklin, S. L., Flint, K. H., Queen, B., … Ethier, K. A. (2018). Youth risk behavior surveillance – United States, 2017. *Morbidity and Mortality Weekly Report. Surveillance Summaries, 67*(8), 1 – 114. https://doi.org/10.15585/mmwr.ss6708a1

Kazdin, A. E. (2000). Adolescent development, mental disorders, and decision making of delinquent youths. In P. Grisso and R. G. Schwartzz (Eds.) *Youth on trial: A developmental perspective on juvenile justice* (pp 33 – 65). University of Chicago Press.

Kline, R.B. (2016). *Principles and practice of structural equation modeling* (4th ed.). Guildford Press.

Klinge, V., & Strauss, M. E. (1976). Effects of scoring norms on adolescent psychiatric patients' MMPI profiles. *Journal of Personality Assessment, 40*(1), 13-17. https://doi.org/10.1207/s15327752jpa4001_3

Klinge, V., Lachar, D., Grisell, J., & Berman, W. (1978). Effects of scoring norms on adolescent psychiatric drug users' and nonusers' MMPI profiles. *Adolescents*, *13*, 1-11.

Krishnamurthy, R. (2016). Gender considerations in self-report personality assessment interpretation. In V. M. Brabender & J. L. Mihura (Ed.), *Handbook of gender and sexuality in psychological assessment.* pp. 128 – 148. Routledge.

Lahey, B. B., Schwab-stone, M., Goodman, S. H., Waldman, I.D., Canino, G., Rathouz, P. J., … Jensen, P. S. (2000). Age and gender differences in oppositional behavior and conduct problems: A cross-sectional household study of middle childhood and adolescence. *Journal of Abnormal Psychology, 109*(3), 488-503. https://doi.org/10.1037/0021-843x.109.3.488

Long, K. S., Graham, J. R., & Timbrook, R. E. (1994). Socioeconomic status and MMPI-2 interpretation. *Measurement and Evaluation in Counseling and Development, 27* (3), 158 - 177.

Lubke, G., & Muthén, B. (2004). Applying multigroup confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. *Structural Equation Modeling: A Multidisciplinary Journal, 11*(4), 514–534. https://doi.org/10.1207/s15328007sem1104_2

Maras, A., Laucht, M., Gerdes, D., Wilhelm, C., Lewicka, S., Haack, D., … Schmidt, M. H. (2003). Association of testosterone and dihydrotestosterone with externalizing behavior in adolescent boys and girls. *Psychoneuroendocrinology, 28*(7), 932 – 940. https://doi.org/10.1016/s0306-4530(02)00119-1

Marks, P. A., & Briggs, P. F. (1972). Adolescent norm tables for the MMPI. In W. G. Dahlstrom, G. S. Welsh, & L. E. Dahlstrom (Eds.), An MMPI handbook: Vol. 1. Clinical interpretation (rev. ed., pp. 388–399). Minneapolis: University of Minnesota Press.

Marks, P. A., Seeman, W., & Haller, D. L. (1974). *The actuarial use of the MMPI with adolescents and adults.* Baltimore, MD: Williams & Wilkins.

Mattern, K. D., & Patterson, B. D. (2013). Test of slope and intercept bias in college admissions: A response to Aguinis, Culpepper, and Pierce (2010). *Journal of Applied Psychology, 98*(1), 134-147. https://doi:10.1037/a0030610

McCabe, K. M., Lansing, A. E., Garland, A., & Hough, R. (2002). Gender differences in psychopathology, functional impairment, and familial risk factors among adjudicated delinquents. *Journal of the American Academy of Child & Adolescent Psychiatry, 41*(7), 860–867. https://doi.org/10.1097/00004583-200207000-00020

McLaughlin, K. A., & King, K. (2015). Developmental trajectories of anxiety and depression in early adolescence. *Journal of abnormal child Psychology, 43*(2), 311 – 323. https://doi.org/10.1007/s10802-014-9898-1

McNulty, J. L., Harkness, A. R., Ben-Porath, Y. S., & Williams, C. L. (1997). Assessing the personality psychopathology five (PSY-5) in adolescents: New MMPI-A scales. *Psychological Assessment*, *9*(3), 250-259. https://doi.org/10.1037/1040-3590.9.3.250

Meade, A. W., & Bauer, D. J. (2007). Power and precision in confirmatory factor analytic tests of measurement invariance. *Structural Equation Modeling, 14*(4), 611-635. doi:10.1080/10705510701575461

Memmott-Elison, M. K., Holmgren, H. G., Padilla-Walker, L. M., & Hawkins, A. J. (2020). Associations between prosocial behavior, externalizing behaviors, and internalizing symptoms during adolescence: A meta-analysis. *Journal of Adolescence, 80*(1), 98–114. Portico. https://doi.org/10.1016/j.adolescence.2020.01.012

Millsap, R. E. (2011). *Statistical approaches to measurement invariance.* Routledge.

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research, 39*(3):479–515. https://doi.org/10.1207/s15327906mbr3903_4

NCS Pearson. (2018 - 2023). MMPI-A and MMPI-A-RF Protocols. Unpublished raw data.

Park, K. Y. (2018). *Examining the measurement invariance of the MMPI-A-RF externalizing scales across Korean and American adolescent normative samples* (doi:10.25777/ak6h-hf74) [Doctoral dissertation, Old Dominion University]. ODU Digital Commons.

Pearson Assessments US. [uploader] Handel, R. W. [author]. (2021, June 2). *MMPI-A-RF: Basic overview* [Video]. YouTube. https://www.youtube.com/watch?v=5JW2GFsVA1Q

Pearson Assessments US. [uploader] Ben-Porath, Y. [author]. (2021). *Introduction to the MMPI-3* [Webinar]. https://www.brainshark.com/1/player/pearsonassessments?pi=zIMzuKDavz6paZz0&r3f1=&fb=0

Perry, D. G., & Pauletti, R. E. (2011). Gender and adolescent development. *Journal of Research on Adolescence, 21*(1), 61 – 74. https://doi.org/10.1111/j.1532-7795.2010.00715.x

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review, 41*, 71–90. https://doi.org/10.1016/j.dr.2016.06.004

Raykov, T. (2002). Examining group differences in reliability of multiple-component instruments. *British Journal of Mathematical and Statistical Psychology, 55,* 145 – 158. https://doi.org/10.1348/000711002159743

Reynolds, C. R., Altmann, R. A., & Allen, D. N. (2021). *Mastering modern psychological testing: Theory and methods.* Springer International Publishing.

Romano, B., Tremblay, R. E., Vitaro, F., Zoccolillo, M., & Pagani, L. (2001). Prevalence of psychiatric disorders and the role of perceived impairment: Findings from an adolescent community sample. *Journal of Child Psychology and Psychiatry, 42,* 451 – 461. https://doi.org/10.1111/1469-7610.00739

Sass, D. A. (2011). Testing measurement invariance and comparing latent factor means within a confirmatory factor analysis framework. *Journal of Psychoeducational Assessment, 29* (4), 347 – 363. https://doi.org/10.1177/0734282911406661

Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*(4), 350–353. doi:10.1037/1040-3590.8.4.350.

Schmitt, T. A. (2011). Current methodological considerations in exploratory and confirmatory factor analysis. *Journal of Psychoeducational Assessment, 29*, 304-321. https://doi.org/10.1177/0734282911406653

Shorey, S., Ng, E. D., & Wong, C. H. J. (2021). Global prevalence of depression and elevated depressive symptoms among adolescents: A systematic review and meta-analysis. *British Journal of Clinical Psychology, 61*(2), 287–305. Portico. https://doi.org/10.1111/bjc.12333

Silva, S. A., Silva, S. U., Ronca, D. B., Gonçalves, V. S. S., Dutra, E. S., & Carvalho, K. M. B. (2020). Common mental disorders prevalence in adolescents: A systematic review and meta-analyses. *PLOS ONE, 15*(4), e0232007. https://doi.org/10.1371/journal.pone.0232007

Tellegen, A., & Ben-Porath, Y. S. (1992). The new uniform *T* scores for the MMPI-2: Rationale, derivation, and appraisal. *Psychological Assessment, 4*(2), 145 – 155. doi: 10.1037/10403590.4.2.145

Tellegen, A., & Ben-Porath, Y. S. (2008/2011). *MMPI-2-RF (Minnesota Multiphasic Personality Inventory-2 Restructured Form) technical manual.* Minneapolis, MN: University of Minnesota Press.

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: University of Minnesota Press.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4-70. https://doi.org/10.1177/109442810031002

Vogel, D. L., Heimerdinger-Edwards, S. R., Hammer, J. H., & Hubbard, A. (2011). "Boys don't cry": Examination of the links between endorsement of masculine norms, self-stigma, and help-seeking attitudes for men from diverse backgrounds. *Journal of Counseling Psychology, 58*(3), 368–382. https://doi.org/10.1037/a0023688

Wang, J., Han, K., Ketterer, H. L., Weed, N. C., Ben-Porath, Y. S., Kim, J.-H., & Moon, K. (2020). Evaluating the measurement invariance of MMPI-2-RF Restructured Clinical Scale 4 (Antisocial Behavior) between American and Korean clinical samples: Exploring cultural and translation issues affecting item responding. *Journal of Personality Assessment, 103*(4), 465–475. https://doi.org/10.1080/00223891.2020.1769111

Wang, S., Paul De Boeck, & Yotebieng, M. (2023). Heywood Cases in Unidimensional Factor Models and Item Response Models for Binary Data. *Applied Psychological Measurement, 47*(2), 141–154. https://doi.org/10.1177/01466216231151701

Ward, L. M., & Grower, P. (2020). Media and the development of gender role stereotypes. *Annual Review of Developmental Psychology, 2*(1), 177–199. https://doi.org/10.1146/annurev-devpsych-051120-010630

Wicherts, J. M., & Dolan, C. V. (2010). Measurement invariance in confirmatory factor analysis: An illustration using IQ test performance of minorities. *Educational Measurement: Issues and Practice, 29*(3), 39-47. https://doi.org/10.1111/j.1745-3992.2010.00182.x

Zahn-Waxler, C., Shirtcliff, E., & Marceau, K. (2008). Disorders of childhood and adolescence: Gender and psychopathology. *Annual Review of Clinical Psychology, 2008*(4), 275 – 303. https://doi.org/10.1146/annurev.clinpsy.3.022806.091358

Zhao, L., Han, G., Zhao, Y., Jin, Y., Ge, T., Yang, W., Cui, R., Xu, S., & Li, B. (2020). Gender differences in depression: Evidence from genetics. *Frontiers in Genetics, 11*. https://doi.org/10.3389/fgene.2020.562316

VITA

Thomas Jay Augustin
Virginia Consortium Program in Clinical Psychology
Norfolk, VA 23529

Education

Doctorate of Philosophy in Clinical Psychology — August 2024
Virginia Consortium Program in Clinical Psychology

Master of Science in Clinical Psychology — July 2018
Fort Hays State University

Bachelor of Science in Psychology & — December 2015
Bachelor of Music in Musical Theatre
University of Nebraska at Kearney

Research Interests

Thomas Augustin's research interests are primarily in the psychometric properties of psychological measures and within the intersectionality of law and psychology, but additional research interests can be found in educational psychology and research involving the fine and performing arts.

Teaching Interests

Thomas Augustin's teaching interests with graduate level students are a reflection of his research interests with psychological testing and the intersectionality of psychology and the law. For undergraduates, Thomas enjoys teaching research/experimental methods, abnormal psychology, personality, and other classes in relation to clinical and forensic psychology.

Doctoral Clinical Experiences

Doctoral Clinical Psychology Internship — August 2023 – July 2024
Center for Behavioral Medicine, Kansas City, MO

Advanced Practicum Clinical Psychology Trainee — August 2022 – May 2023
Alicia's Place, Virginia Beach, VA

Advanced Practicum Clinical Psychology Trainee — August 2021 – July 2022
Eastern State Hospital, Williamsburg, VA

Graduate Therapy Practicum Student — August 2020 – July 2021
Virginia Beach City Public Schools, Virginia Beach, VA

Graduate Testing/Assessment Practicum Student — January 2020 – May 2020
Neuropsychological Associates at Tidewater, Virginia Beach, VA