


Summer 2015

Examining the Measurement Invariance of the Minnesota Multiphasic Personality Inventory-2-Restructured Form Internalizing Specific Problem Scales in African- American and Caucasian Men

Megan Anne Brokenborough
Old Dominion University

Follow this and additional works at: https://digitalcommons.odu.edu/psychology_etds

 Part of the [Clinical Psychology Commons](#), [Personality and Social Contexts Commons](#), and the [Quantitative Psychology Commons](#)

Recommended Citation

Brokenborough, Megan A.. "Examining the Measurement Invariance of the Minnesota Multiphasic Personality Inventory-2-Restructured Form Internalizing Specific Problem Scales in African- American and Caucasian Men" (2015). Doctor of Psychology (PsyD), dissertation, Psychology, Old Dominion University, DOI: 10.25777/d32k-3a35
https://digitalcommons.odu.edu/psychology_etds/5

This Dissertation is brought to you for free and open access by the Psychology at ODU Digital Commons. It has been accepted for inclusion in Psychology Theses & Dissertations by an authorized administrator of ODU Digital Commons. For more information, please contact digitalcommons@odu.edu.

**EXAMINING THE MEASUREMENT INVARIANCE OF THE MINNESOTA
MULTIPHASIC PERSONALITY INVENTORY-2-RESTRUCTURED FORM
INTERNALIZING SPECIFIC PROBLEM SCALES IN AFRICAN-AMERICAN
AND CAUCASIAN MEN**

by

Megan Anne Brokenborough
B.A. June 2009, University of California Irvine
M.A. August 2013, Norfolk State University

A Dissertation Submitted to the Faculties of Eastern Virginia Medical School,
Norfolk State University, and 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 2015

Approved by:

Richard W. Handel (Director)
Eastern Virginia Medical School

Robert P. Archer (Member)
Eastern Virginia Medical School

Desideria S. Hacker (Member)
Norfolk State University

Serina A. Neumann (Member)
Eastern Virginia Medical School

James F. Paulson (Member)
Old Dominion University

ABSTRACT

EXAMINING THE MEASUREMENT INVARIANCE OF THE MINNESOTA MULTIPHASIC PERSONALITY INVENTORY-2-RESTRUCTURED FORM INTERNALIZING SPECIFIC PROBLEM SCALES IN AFRICAN-AMERICAN AND CAUCASIAN MEN

Megan Anne Brokenborough
Virginia Consortium Program in Clinical Psychology, 2015
Director: Dr. Richard W. Handel

Test bias has long been an area of investigation in the personality assessment literature, including the MMPI-2-RF. Research on previous versions of the MMPI and MMPI-2-RF has pointed to mixed results. The current study aims to examine test bias on the MMPI-2-RF's nine Internalizing Specific Problem Scales by examining measurement invariance using MIMIC modeling and investigating differential item functioning (DIF). After removal of invalid protocols, the first sample consisted of 2,980 protocols from various settings requested from Pearson (255 African American and 2,755 Caucasian protocols). The second sample consisted of 1,379 valid protocols from psychiatric inpatient settings (1,245 Caucasian and 133 African American protocols). MIMIC modeling was conducted using delta parametrization and the WLSMV estimator in Mplus (Muthén and Muthén, 1998-2012). Latent continuous response variables and threshold estimates were used to accommodate categorical indicators. Results of the MIMIC modeling pointed to latent mean differences in four of the nine and two of the nine scales in the Pearson and inpatient samples, respectively. In both samples, latent mean differences were found between African Americans and Caucasians on the Multiple Specific Fears scale. Evidence of DIF was seen in seven of the nine scales in both the Pearson and inpatient samples. However, only a total of four items were found to

functioning differently on the Inefficacy and Multiple Specific Fears scales across both samples. These results have implications for the MMPI-2-RF's invariance across African American and Caucasian test takers and overall psychological assessment standards involving fairness in testing.

This dissertation is dedicated to the faculty of curiosity that spurs the pursuit of knowledge purely for the sake of learning.

A philosopher knows that in reality he knows very little. That is why he constantly strives to achieve true insight. Socrates was one of these rare people. He knew that he knew nothing about life and about the world. And now comes the important part: it troubled him that he knew so little.

- *Sophie's World* by Jostein Gaarder, p. 67

ACKNOWLEDGEMENTS

There have been many people, some no longer with me, who have contributed to my academic success, eventual completion of this dissertation, and ultimately my degree. Rather than paint finite strokes, I wish to acknowledge all of the important people in my life who have provided me with support, advice, and encouragement along the way. Each of you has played an important role in helping me complete this project and my degree. My mother has a special place in this painting of gratitude because without her unconditional support, I would not have had access to the canvas. My late father consistently gave me various mediums and combinations of color with which to experiment. My step-father and sisters have been a source of inspiration, never letting me lose sight of the whole picture. I would also like to acknowledge my dissertation committee for their time and effort in helping me complete this project. My dissertation chair has consistently exceeded expectations in the time, energy, and dedication he has given our joint projects. He has helped me come to enjoy something I previously dreaded – statistics.

TABLE OF CONTENTS

	Page
LIST OF TABLES	vii
LIST OF FIGURES	ix
Chapter	
I. INTRODUCTION	1
II. LITERATURE REVIEW	11
A HISTORY OF THE MMPI	11
A HISTORY OF TEST BIAS RESEARCH	31
TEST BIAS RESEARCH ON THE MMPI/MMPI-2 WITH MINORITY POPULATIONS	33
TEST BIAS RESEARCH ON THE MMPI/MMPI-2 WITH AFRICAN AMERICAN POPULATIONS	42
ESTABLISHING MEASUREMENT INVARIANCE	56
III. RATIONALE OF THE PRESENT STUDY	71
IV. METHODOLOGY	75
PARTICIPANTS	75
INSTRUMENTS	80
STATISTICAL ANALYSES	84
V. RESULTS	89
DESCRIPTIVE STATISTICS	89
POPULATION HETEROGENEITY AND DIFFERENTIAL ITEM FUNCTIONING (MIMIC MODELS)	95
VI. SUMMARY AND DISCUSSION	138
BASELINE MODELS	138
MIMIC MODELS	139
DIFFERENTIAL ITEM FUNCTIONING	141
IMPLICATIONS OF FINDINGS	145
STUDY LIMITATIONS AND STRENGTHS	150
FUTURE DIRECTIONS	152
REFERENCES	154
VITA	179

LIST OF TABLES

Table	Page
1. Number of Valid (and Invalid) Protocols by Validity Scale for Each Sample	76
2. Demographic Information of the Inpatient Sample from the Minneapolis Veterans Affairs Medical Center (VAMC) and Hennepin County Medical Center (HCMC)	78
3. Correlated Indicator Error Terms for MMPI-2-RF items in the CFA Models for Both the Outpatient and Inpatient Sample by Scale	86
4. Descriptive Statistics and Reliability Coefficients for the Internalizing Specific Problem Scales for the Pearson Sample by Ethnicity	90
5. Descriptive Statistics and Reliability Coefficients for the Internalizing Specific Problem Scales for the Inpatient Sample by Ethnicity	91
6. Specific Problem Scale Correlations by Ethnicity for the Pearson Sample	93
7. Specific Problem Scale Correlations by Ethnicity for the Inpatient Sample	94
8. Suicidal/Death Ideation (SUI) Baseline CFA Factor Loadings, Thresholds, and Model Fit Indices for the Pearson and Inpatient Samples	96
9. Helplessness/Hopelessness (HLP) Baseline CFA Factor Loadings, Thresholds, and Model Fit Indices for the Pearson and Inpatient Samples	100
10. Self-Doubt (SFD) Baseline CFA Factor Loadings, Thresholds, and Model Fit Indices for the Pearson and Inpatient Samples	104
11. Inefficacy (NFC) Baseline CFA Factor Loadings, Thresholds, and Model Fit Indices for the Pearson and Inpatient Samples	108

Table	Page
12. Stress/Worry (STW) Baseline CFA Factor Loadings, Thresholds, and Model Fit Indices for the Pearson and Inpatient Samples	113
13. Anxiety (AXY) Baseline CFA Factor Loadings, Thresholds, and Model Fit Indices for the Pearson and Inpatient Samples	118
14. Anger Proneness (ANP) Baseline CFA Factor Loadings, Thresholds, and Model Fit Indices for the Pearson and Inpatient Samples	122
15. Behavior Restricting Fears (BRF) Baseline CFA Factor Loadings, Thresholds, and Model Fit Indices for the Pearson and Inpatient Samples	127
16. Multiple Specific Fears (MSF) Baseline CFA Factor Loadings, Thresholds, and Model Fit Indices for the Pearson and Inpatient Samples	132

LIST OF FIGURES

Figure	Page
1. The MIMIC model for the Suicidal/Death Ideation (SUI) Scale in the Pearson sample	97
2. The MIMIC model for the Suicidal/Death Ideation (SUI) Scale in the inpatient sample	99
3. The MIMIC model for the Helplessness/Hopelessness (HLP) Scale in the Pearson sample	101
4. The MIMIC model for the Helplessness/Hopelessness (HLP) Scale in the inpatient sample	103
5. The MIMIC model for the Self-Doubt (SFD) Scale in the Pearson sample	105
6. The MIMIC model for the Self-Doubt (SFD) Scale in the inpatient sample	107
7. The MIMIC model for the Inefficacy (NFC) Scale in the Pearson sample	110
8. The MIMIC model for the Inefficacy (NFC) Scale in the inpatient sample	112
9. The MIMIC model for the Stress/Worry (STW) Scale in the Pearson sample	115
10. The MIMIC model for the Stress/Worry (STW) Scale in the inpatient sample	117
11. The MIMIC model for the Anxiety (AXY) Scale in the Pearson sample	119
12. The MIMIC model for the Anxiety (AXY) Scale in the inpatient sample	121
13. The MIMIC model for the Anger Proneness (ANP) Scale in the Pearson sample	124

Figure	Page
14. The MIMIC model for the Anger Proneness (ANP) Scale in the inpatient sample	126
15. The MIMIC model for the Behavior Restricting Fears (BRF) Scale in the Pearson sample	129
16. The MIMIC model for the Behavior Restricting Fears (BRF) Scale in the inpatient sample	131
17. The MIMIC model for the Multiple Specific Fears (MSF) Scale in the Pearson sample	134
18. The MIMIC model for the Multiple Specific Fears (MSF) Scale in the inpatient sample	136

CHAPTER I

INTRODUCTION

The Minnesota Multiphasic Personality Inventory-2-Restructured Form (MMPI-2-RF; Tellegen & Ben-Porath, 2008/2011) is the third version of the MMPI test for use with adults. The MMPI-2-RF is built around the Restructured Clinical (RC) scales (Tellegen et al., 2003). The RC scales were originally released for use with the MMPI-2 (Butcher et al., 2001). The scale development techniques used to create the RC scales were subsequently used to develop other scales on the MMPI-2-RF (Ben-Porath, 2012). While keeping the external correlates of the scales in consideration, the resulting scales were examined and tailored for maximum reliability, convergent and discriminant validity, and meaningfulness. The MMPI-2-RF is a more concise measure than the MMPI-2; reducing the item pool from 567 to 338 items and contains nine Validity Scales, three Higher-Order Scales, nine RC Scales, two Interest Scales, 23 Specific Problem (SP) Scales, and revised Personality Psychopathology Five (PSY-5) Scales (Ben-Porath & Tellegen, 2008/2011).

The SP scales were developed to highlight characteristics included in or related to, yet not exclusively or saliently addressed by one of the RC scales (Ben-Porath, 2012). Based on conceptual considerations and empirical analyses, four sets of SP Scales were developed, the Somatic/Cognitive, Internalizing, Externalizing, and Interpersonal scales. The Somatic/Cognitive SP scales assess symptoms related to physical and cognitive symptoms (Ben-Porath, 2012). The Internalizing SP scales assess dimensions related to suicidality, helplessness, self-doubt, anxiety, and fears (Ben-Porath & Tellegen, 2008/2011). The Externalizing SP Scales assess adolescent

conduct problems, substance abuse, aggression, and activation. The Interpersonal SP scales place a range of interpersonal functioning at the forefront.

While the MMPI-2-RF normative sample is ethnically diverse, such diversity does not guarantee that the scales function the same way with all ethnic groups. To investigate possible ethnic differences in scale functioning, studies of possible test bias are still needed. Early studies on test bias with the MMPI and MMPI-2 examined mean T-score differences, simply any differences on mean T-scores between groups. More contemporary research with the MMPI-2 and MMPI-2-RF has examined test bias in two different forms – predictive and measurement bias. As will be discussed later, very little research has been conducted in the area of measurement invariance as a means of assessing for measurement bias.

Predictive bias can be seen when a test leads to systematic inaccuracies in the prediction of an external variable based on group membership (Millsap, 1997). This type of bias is usually assessed in terms of intercept or slope bias using moderated multiple regression. Intercept bias involves examining whether a predictor systematically under- or overpredicts the criterion variable for the different groups (Anastasi & Urbina, 1997; Nunnally & Burnstein, 1994). Slope bias suggests varying prediction accuracy and can be seen when there is difference in the magnitude of the correlation between the predictor and criterion for the different groups (Arbisi, Ben-Porath, & McNulty, 2002). The other type of bias, measurement bias, involves systematic inaccuracies in the data a test provides about a characteristic or latent variable based on group membership and can be assessed using measurement invariance tests (Millsap, 1997).

Early research investigating test bias on the MMPI examined mean T-score differences between groups. However, this method is problematic as mean score differences do not necessarily automatically equate with test bias. Such differences instead may simply reflect underlying group differences in symptoms or setting (Archer, Griffin, & Aiduk, 1995). Early test bias research with the original MMPI and MMPI-2 comparing Hispanic Americans, Native Americans, and Asian Americans with Caucasians focused on mean T-score differences. Thus, this previous research has been methodologically limited and has not conclusively demonstrated whether or not test bias existed between these groups and Caucasians.

In comparing Hispanic Americans and Caucasians, one possible explanation for mean T-score differences on the MMPI and MMPI-2 is actual differences in the base rates of psychopathology between groups in a given sample. Therefore, simply comparing mean scale scores is an inadequate method to examine the possibility of test bias. Nevertheless, a number of studies have been conducted comparing Hispanic Americans and Caucasians on the MMPI and MMPI-2 and results indicate that mean T-score differences exist but no consistent patterns have been found (Hall, Bansal, & Lopez, 1999; Velasquez and Callahan, 1990a). Some studies examining these differences in the MMPI-2 have questioned whether score differences may be related to acculturation (Canul & Cross, 1994; Lessenger, 1997).

Research comparing mean T-scores in Native American and Caucasian populations has also found no clear pattern of differences. Some studies on both the MMPI and MMPI-2 have found higher scores among Native Americans (Klein, Rozytko, Flint, & Roberts, 1973; Lacey, 2004; Prewett, 2012) whereas others found

no meaningful differences (Page & Bozlee, 1982; Venn, 1988). Examination of confounding variables on the MMPI-2 (e.g., education, acculturation) has accounted for some of these differences (Pace et al., 2006) and examination of external correlates has indicated that these differences may be related to underlying symptomology (Greene, Robin, Albaugh, Caldwell, & Goldman, 2003).

Examination of mean T-score differences in Asian Americans and Caucasians has also been undertaken with the MMPI and MMPI-2. As with other group comparisons, some findings point to statistically significant T-score differences (Lee, Cheung, Man, & Hsu, 1992; Kwan, 1999; Sue & Sue, 1974). Some studies have attributed these differences to acculturation or other factors (Greene, 1987; Sue, Keefe, Enomoto, Durvasula, & Chao, 1996; Tsushima & Onorato, 1982; Tsushima & Stoddard, 1990). Again, it is difficult to interpret whether mean T-score differences indicate test bias or differences in underlying characteristics.

Early research on the original MMPI also compared mean T-scores of African American and Caucasian test takers. Research comparing low income African American and Caucasian men and women found inconsistent results (Harrison & Kass, 1967, McGill, 1980). In examining groups of students on the MMPI, some research demonstrated that African Americans scored higher on certain scales while Caucasians scored higher on others (Ball, 1960; Moore & Handal, 1980). However, other research found differences by ethnicity and gender on mean scores (McDonald and Gynther, 1962). Controlling for demographic variables in various populations (i.e., students, inpatients, and forensic patients) has minimized the score differences in some studies (Bertelson, Marks, & May, 1982; Butcher, Ball, & Ray, 1964) but not

others (Butcher, Braswell, & Raney, 1983; Holcomb & Adams, 1982; King, Carroll, & Fuller, 1977; McDonald & Gynther, 1962).

Inpatient and forensic populations have also provided inconsistent results with regard to whether T-score differences exist on the MMPI between African American and Caucasian groups (Costello, Fine, & Blau, 1973; Davis, 1975; Davis & Jones, 1974; McCreary & Padilla, 1977; Smith & Graham, 1981). Studies examining such differences in African Americans and Caucasians in substance abuse treatment have generally found lower scale elevations for African American test takers (Penk et al., 1982; Penk, Woodward, Robinowitz, & Hess, 1978). Many of the apparent inconsistencies in these studies may be due largely to sampling error. Meta-analytic techniques are effective methods to minimize the influence of sampling error inherent in individual studies.

A meta-analysis comparing mean T-scores of African American and Caucasian men and African American and Caucasian women on the MMPI and MMPI-2 found that African Americans scored higher on some scales but lower on others (Hall, Bansal, & Lopez, 1999). However, the aggregate effect sizes for both men and women were small. Greene (1987) argued that mean T-score differences of less than five points are probably too small to be clinically meaningful. Thus, while there has been evidence of statistically significantly different T-scores between African Americans and Caucasians on the MMPI-2 in various settings, some research points to the clinical meaningfulness of these differences (Castro, Gordon, Brown, Anestis, & Joiner, 2008; Munley, Morris, Murraray, & Baines, 2001) whereas others found such

differences lacked clinical significance (Frueh, Gold, de Arellano, & Brady, 1997; Timbrook & Graham, 1994) based on the five T-score point criterion.

In attempts to explore test bias in a more sophisticated manner, research began examining external correlates and predictive bias in MMPI-2 data. This research has assessed protocols from various settings and some findings have pointed to slight underprediction of psychopathology for African Americans (Arbisi et al., 2002; Timbrook & Graham, 1994) for certain scales. On the other hand, Monnot, Quirk, Hoerger, and Brewer (2009) found that the MMPI-2 overpredicted psychopathology in African Americans for some scales but not for others. Studies by Arbisi et al. (2002) and Monnot et al. (2009) both employed linear regression with binary dependent variables. However, the appropriate analytic technique with dichotomous dependent variables is binary logistic regression rather than Ordinary Least Squares regression. It is unknown if the results of these studies would have been altered by the use of the more appropriate binary logistic regression procedure. Finally, other research has demonstrated a lack of predictive bias when comparing African Americans and Caucasians scores on the MMPI-2 and MMPI-2-RF (Castro et al., 2008; McBride, 2013; McNulty, Graham, Ben-Porath, & Stein, 1997).

Thus, as with other minority groups, MMPI and MMPI-2 research findings related to the presence of test bias in African Americans are not entirely consistent, and the extent of the clinical significance of small to moderate effect sizes in the over- or under-prediction of external variables is unknown. While the examination of predictive bias provides more information than mean T-score differences, such information may still prove limited. Using an external correlate as a criterion operates

under the assumption that the external criterion is not biased, which may or may not be the case.

While the examination of predictive bias provides more information than mean T-score differences, it does not address the question of measurement bias. Test bias research has been moving toward the investigation of bias internally, or measurement bias testing. Measurement bias is typically assessed using measurement invariance testing (Millsap, 2011). Measurement invariance, as applied in psychometrics, is a concept that item responses relate to a latent variable in the same way across groups. Measurement invariance can be assessed using Multiple-Group Confirmatory Factor Analysis (MG-CFA) or Multiple Indicator Multiple Causes (MIMIC) modeling (Brown, 2006; Kim, Yoon, & Lee, 2012). The present study employed MIMIC modeling as a means of examining measurement invariance. The rationale for using MIMIC modeling will be discussed in the literature review.

Confirmatory factor analysis (CFA), a theory-driven structural equation modeling procedure, is at the heart of MIMIC modeling. CFA, using a fitting function, produce estimates of model parameters, including factor loadings, error variances, and factor variances. These estimates can be pre-specified to be fixed to a certain value, constrained to a range of values, or freely estimated. CFA delivers parameter estimates that are geared at maximizing the probability that the sample and predicted variance/covariance matrix are not statistically significantly different. Goodness-of-fit indices are then examined to evaluate the fit of the specified model based on whether the solution best represents the observed variances and covariances

from the input data. Modification indices can be used to evaluate the impact of freeing certain parameters.

Assessing for measurement invariance using MIMIC modeling begins with assessing the baseline CFA model on the full sample, merging groups (Brown, 2006). If the model demonstrates adequate fit, MIMIC modeling involves adding dummy-coded covariates, representing group membership, to the baseline CFA model to examine their effect on the latent variable (Schumacker & Lomax, 2012). Specifically, the latent variable is regressed upon the covariates to examine latent mean differences across levels of the covariate (e.g., ethnicity; Kim, Yoon, & Lee, 2012; Schumacker & Lomax, 2012). A single input matrix is used that contains variances and covariances of the latent factor and observed covariates (Brown, 2006). A significant direct effect of an observed covariate on a latent factor points to group differences on latent means, also known as population heterogeneity.

To take measurement invariance testing a step further with MIMIC modeling, indicators can be regressed on the covariates to assess for differential item functioning. Differential item functioning (DIF) points to different measurement properties of an item based on group membership, holding any group mean differences constant (Woods, Oltmanns, & Turkheimer, 2011). Thus, a significant direct effect of the observed covariate on an indicator signifies group differences on the indicator's intercept and the presence of measurement noninvariance (Brown, 2006). An item demonstrating DIF is noninvariant because part of whether it is endorsed is based on group membership, not levels of underlying traits (Woods et al., 2011). MIMIC models, including assessment of DIF, can be tested with or without a hypothesis

regarding invariance (Brown, 2006). In an exploratory approach to MIMIC modeling, all direct effects between the covariate and indicators are set to zero and modification indices are examined for significant direct effects.

MIMIC modeling with categorical indicators, as would be the case with the MMPI-2-RF's dichotomous responses, varies slightly (Brown, 2006; Muthén & Asparouhov, 2002; Muthén & Muthén, 2009b). Latent continuous response variables and thresholds, tetrachoric correlations, and different fitting functions must be used. Ultimately, the core of using this analytic technique rests in the assumption that each binary (true-false) MMPI-2-RF item is actually measuring a continuous underlying variable.

The goals of the current study were to evaluate the possibility of population heterogeneity and differential item functioning in the MMPI-2-RF Internalizing Specific Problem Scales in African American and Caucasian samples using MIMIC modeling. Research comparing the MMPI-2 in African American and Caucasian populations has provided inconsistent results while research comparing the two groups on MMPI-2-RF Specific Problems scales is nonexistent. The SP Scales were chosen because, given their narrow bandwidth focus (Tellegen & Ben-Porath, 2008/2011), they are more likely to be unidimensional than other MMPI-2-RF Scales. The Internalizing SP Scales were chosen because they represent one of the defined subsets of MMPI-2-RF scales.

Interestingly, while measurement invariance research has been building in the psychological assessment literature (Carle, Millsap, & Cole, 2008; Culhane, Morera, Watson, & Millsap, 2009, 2011; Woods et al., 2011), only one study thus far has

examined the measurement invariance of the MMPI-2 and did so using the English MMPI-2 and Korean MMPI-2 RC Scales (Ketterer, 2011). No studies have explored the measurement invariance of the MMPI-2-RF Scales in American samples. This study is meant to build upon previous test bias research within the MMPI literature, but also advance this research by providing the first assessment of measurement invariance in the MMPI-2-RF in African American and Caucasian populations. The present study was approved by the Eastern Virginia Medical School's Institutional Review Board, approval number 14-08-NH-0177.

CHAPTER II

LITERATURE REVIEW

A History of the MMPI

The family of MMPI assessments (MMPI/MMPI-2/MMPI-2-RF) have been and are currently used to measure personality and psychopathology. The MMPI-2 is one of the most frequently used psychological tests around the world and usually reported to be the most used measure of personality and psychopathology (Camara, Nathan, & Puente, 2000; Graham, 2006).

MMPI. The original MMPI (Hathaway & McKinley, 1943) was developed using a criterion keying approach, meaning that the clinical scales were created by choosing items that were endorsed by patients known to have a particular psychopathology and not endorsed by others (Hathaway & McKinley, 1940, 1942; McKinley & Hathaway, 1940, 1942, 1944). The developers conducted statistical analyses to identify eight sets of items that distinguished test takers who belonged to eight different diagnostic groups from “non-patients” or those without any such psychological problems (Ben-Porath & Tellegen, 2008/2011). The eight diagnostic groups, and resulting scales, were Hypochondriasis (Hs), Depression (D), Hysteria (Hy), Psychopathic Deviance (Pd), Paranoia (Pa), Psychasthenia (anxiety; Pt), Schizophrenia (Sc), and Hypomania (Ma). A scale measuring Masculinity/Femininity (Mf) was introduced later, in an attempt to assess for homosexual tendencies at a time when homosexuality was considered a psychological disorder. A scale measuring Social Introversion (Si) was also added later, resulting in the ten Clinical Scales.

Although the development of the clinical scales was novel and appeared promising, attempts to replicate their validity as indicators of diagnostic categories varied (Ben-Porath & Tellegen, 2008/2011). Some scales appeared to be moderately successful in predicting diagnostic group membership, while other scales lacked such validity. As a result, the original plan of using the MMPI as a diagnostic instrument was abandoned. However, researchers and clinicians began to notice that individual Clinical scales and constellations of scores on the Clinical Scales were, in fact, empirically related to personality characteristics and psychopathology. MMPI research then moved to identifying these correlates in a wide range of settings and populations for their use in applied assessment. In addition to identifying empirical correlates of scores on individual MMPI scales, some researchers developed elaborate, configural “cookbook” systems for MMPI scales (e.g., Gilberstadt & Duker, 1965).

Twenty years after its birth, the MMPI had taken on a new life. Rather than using it as a diagnostic tool, clinicians were using the MMPI to assess for personality characteristics, symptoms of psychopathology, and behavioral tendencies (Ben-Porath & Tellegen, 2008/2011). Code types were prominent in interpretation and empirical correlates of code types dominated research and interpretation. However, around this time, researchers also began looking at the item content of the MMPI rather than only external correlates (Wiggins, 1966). Consequently, more direct and easily communicated content-based scales began to be developed.

Restandardization project and the MMPI-2. After many decades of clinical use, it became necessary to revise the MMPI due to a number of salient issues (e.g., outdated norms, outdated or unclear wording of items, the omission of important areas

of psychopathology such as suicide attempts, drug use, and treatment related behaviors; Butcher et al., 1989; Graham, 2006). The University of Minnesota Press commissioned a restandardization project in 1982, with a goal of revising the existing MMPI (Ben-Porath & Tellegen, 2008/2011). The most pressing need in the restandardization project was new norms. The initial MMPI norms were based on a mostly Caucasian, working-class, rural sample with an average of eight years of education from around the University of Minnesota. Since the MMPI had gained popularity and was being used across the United States and abroad, these norms were no longer appropriate. To this end, the revised norms for the MMPI-2 were collected from different areas of the United States with an attempt to represent the census data from the time (Schinka & LaLone, 1997). In the end, 2,600 people (1,462 women and 1,138 men) constituted the MMPI-2's more nationally-representative normative group (Ben-Porath & Tellegen, 2008/2011).

Another main goal of the restandardization project was revision of the test items (Ben-Porath & Tellegen, 2008/2011). Items that were not scored on any of the main scales, deemed offensive due to concern with religious beliefs or sexist verbiage, or those that made reference to bowel or bladder functioning were excluded from the MMPI-2. Some of the items also contained outdated language or reference to cultural norms and thus were revised. Despite these major changes, all wanted continuity between the MMPI and MMPI-2 and thus the items on the Clinical Scales were only altered slightly and only a few were eliminated from the test. Of the 383 items scored on the Validity and Clinical Scales on the original MMPI, 372 were maintained in the MMPI-2. In total, 64 MMPI-2 items were revised from the original MMPI. Research

indicated that such revisions did not impact the psychometric properties of the scales (Ben-Porath & Butcher, 1989). Further, the code types created by the MMPI and MMPI-2 norms appeared to be generally compatible when considering the effect of measurement error (Ben-Porath & Tellegen, 1995; Graham, Timbrook, Ben-Porath, & Butcher, 1991).

In all, the restandardization project undertaken by Butcher, Dahlstrom, Graham, Tellegen, and Kraemmer (1989) provided the MMPI-2 with a wealth of improvements, including more representative norms and a new means of calculating standard scores (Ben-Porath & Tellegen, 2008/2011). Two new response inconsistency scales were also developed to identify random or fixed responding, The Variable Response Inconsistency Scale (VRIN) and the True Response Inconsistency Scale (TRIN). The F_B Scale was introduced to evaluate infrequent responding to items in the later portion of the test. Finally, MMPI-2 Content Scales (Butcher, Graham, Williams, & Ben-Porath, 1990) were developed to replace the Content Scales in the original MMPI. The new Content Scales, in line with the original, allowed for more streamlined assessment of the symptomology measured by the Clinical Scales, but also evaluated symptoms or problems not covered by the Clinical Scales.

After the release of the MMPI-2, research on the test continued and a revised edition of the test manual was published (Butcher et al., 2001). The revised test manual introduced a host of new scales (Ben-Porath & Tellegen, 2008/2011). Arbisi and Ben-Porath (1995) introduced the Infrequency-Psychopathology scale (F_p) as a supplement to the F scale in identifying infrequent responding. The F_p scale, however, identifies infrequent endorsing of items by the normative sample and psychiatric

inpatients (Graham, 2006). As a result, an elevated score on the F_p scale is more likely to indicate an attempt to over-report psychopathology. The Superlative Self-Presentation (S) Scale (Butcher & Han, 1995), another validity scale, was also introduced and assesses a tendency to present as highly virtuous, free from psychological difficulties, and morally and socially flawless (Graham, 2006).

Content Component Scales (Ben-Porath & Sherwood, 1993), which assess specific sub-areas of the Content Scales, were also introduced in the revised MMPI-2 manual (Ben-Porath & Tellegen, 2008/2011). The Personality Psychopathology Five (PSY-5; Harkness, McNulty, & Ben-Porath, 1995; Harkness, McNulty, Ben-Porath, & Graham, 2002) Scales, which measure both normal and abnormal personality traits, were also included in the revised manual (Ben-Porath & Tellegen, 2008/2011; Graham, 2006). The original MMPI Hostility (Ho) Scale was revised and introduced in the MMPI-2 manual (Ben-Porath & Tellegen, 2008/2011). Following release of the revised MMPI-2 manual, the Symptom Validity Scale (FBS; Lees-Haley, English, & Glenn, 1991) was added to the standard set of Validity Scales (Ben-Porath & Tellegen, 2008/2011). Ben-Porath and Forbey (2003) also created non-gendered norms for the MMPI-2.

The Restructured Clinical (RC) Scales. Despite the advances made in the MMPI-2, the core of the MMPI-2, the Clinical Scales, remained essentially unchanged (Ben-Porath & Tellegen, 2008/2011). While this was advantageous for continuity between the MMPI and MMPI-2, psychometric problems with the Clinical Scales were troubling. The range of the item content on a single Clinical Scale and resulting item overlap and high intercorrelations between scales creates structural heterogeneity

among the Clinical Scales. Such heterogeneity ultimately leaves the convergent and discriminant validity of scores on the scales lacking. The RC Scales (Tellegen et al., 2003) were developed to improve the psychometric properties of scores on the Clinical Scales by reducing their heterogeneity and increasing their distinctiveness (Ben-Porath & Tellegen, 2008/2011). Further, each RC scale assesses one of the areas identified as a core part of one or more of the Clinical Scales, resulting in easier and more refined access to particular clinical symptoms. At the time of publication in 2003, the developers of the RC Scales recommended that they were used in conjunction with the Clinical Scales in interpretation.

Demoralization is a central construct within the RC Scales. Demoralization is theorized to be a general factor that will inflate correlations between characteristics or psychopathology that should be independent in clinical assessment measures such as the MMPI (Tellegen, 1985). Demoralization is stated to be one side of an overarching mood dimension of Pleasant (happy, enthusiastic, content) versus Unpleasant (afraid, upset, sad) Arousal or Activation (excited, astonished, tense vs. relaxed, sleepy; Watson & Tellegen, 1985). Demoralization, on the Unpleasant end of the dimension, is the combination of high negative and low positive activation and thus identified as a risk factor for psychological problems (Tellegen, 1985; Watson & Tellegen, 1985).

Based on this theory, Demoralization, which is common in clinical settings, was seen as a common general factor accounting for shared variance amongst the clinical scales and thus contributing to the heterogeneity of the scales (Ben-Porath & Tellegen, 2008/2011; Tellegen et al., 2003). Further, the presence of Demoralization in such populations will likely lead to MMPI profiles with multiple scale elevations

that may or may not be related to the core characteristic the scale is attempting to measure. On the other hand, low levels of Demoralization may suppress Clinical Scale scores. Therefore, the minimization of Demoralization in the Clinical Scales was at the core of the RC Scale development project. As a result, the final nine RC Scales can prove helpful in determining what salient problems exist for the test taker apart from overarching Demoralization. Demoralization, as measured on the MMPI-2 and MMPI-2-RF, assesses general unhappiness and dissatisfaction.

The final nine RC Scales include a Demoralization (RCd) specific scale (Ben-Porath & Tellegen, 2008/2011). Somatic Complaints (RC1) assesses for diffuse health complaints and Low Positive Emotions (RC2) measures lack of positive emotional responsiveness. Cynicism (RC3) evaluates non-self-referential beliefs about distrust and generally not liking others. Antisocial Behavior (RC4) is measured by items related to rule breaking and irresponsible behavior. Ideas of Persecution (RC6) assesses for self-referential beliefs that others are threatening and Dysfunctional Negative Emotions (RC7) measures maladaptive anxiety, anger, and irritability. Aberrant Experiences (RC8) is measured by items related to unusual perceptions or thoughts. Finally, Hypomanic Activation (RC9) evaluates over-activation, aggression, impulsivity, and grandiosity.

Development of the RC Scales. The development of the RC scales is thoroughly outlined in a test monograph (Tellegen et al., 2003) and occurred in four steps (Ben-Porath & Tellegen, 2008/2011).

Step one. Based on the theory of Demoralization, Tellegen et al. (2003) tested the hypothesis that the MMPI-2 Clinical Scales contain a number of items assessing

this construct using four samples. The samples consisted of 832 men and 380 women involved in a residential substance abuse program and 232 men and 191 women at one of three psychiatric facilities in two states. First, the researchers used principal component analysis with a Varimax rotation to identify Demoralization items on Clinical Scales 2 and 7. Across all four samples, 14 items had a loading of at least $|.50|$ on the principal factor.

Second, distinctive positive emotionality and negative emotionality factors were examined in all four samples, requiring a four-factor rotation (Tellegen et al., 2003). Once appropriate items were located, brief measures of positive and negative emotionality were created. Tellegen et al. (2003) found 17 items that correlated with both of these measures (in opposite directions) of at least $|.25|$. Further factor analysis of those items in all four samples resulted in 12 items with loadings of at least $|.50|$ on the principal factor. In comparing the two sets of items (the 14 and 12 item set), 11 items overlapped. Ten of these items compose the final Demoralization scale.

The authors concluded that their hypotheses were accurate based on the content of the items and the factor analyses (Tellegen et al., 2003). Next, the remainder of the MMPI-2 item pool was examined for Demoralization items. Items not on Clinical Scales 2 and 7 were correlated with the measures of positive and negative emotionality. Based on these correlations, 23 items were identified for further exploration. After further analysis, 18 of those 23 items were retained in the final Demoralization Scale.

Step two. Three hypotheses guided the second step in development of the RC Scales, including the assumption that Demoralization is not a core part of any of the

Clinical Scales; removing Demoralization items will create more distinct and incrementally valid Clinical Scales; and item factor analysis of each Clinical Scale, combined with the Demoralization items, will yield a distinct Demoralization factor (Tellegen et al., 2003). Consequently, the second step in developing the RC scales involved conducting a separate item exploratory factor analysis (principal component analysis with Varimax rotation) of each of the Clinical Scales combined with the 23 identified Demoralization items (Ben-Porath & Tellegen, 2008/2011; Tellegen et al., 2003).

For the majority of the Clinical Scales, a two factor solution resulted in a Demoralization and discrete non-Demoralization component loading on the separate factors (Tellegen et al., 2003). In such cases, the second factor was identified as the core component of the scale. On three Clinical Scales, a three factor solution emerged. In such cases, the first factor contained Demoralization items. The second factor consisted of a number of items related to other Clinical Scales and the third factor was considered the core component of the scale. For example, Clinical Scale 6 resulted in a Demoralization factor, a factor with items assessing non-self-referential distrust and cynicism, and a third factor that contained items related to self-referential persecutory ideas. In the end, 12 sets of items emerged related to Demoralization and 11 sets of items related to major components measured by the respective Clinical Scale (Ben-Porath & Tellegen, 2008/2011).

Step three. The third step in the development of the RC scales consisted of developing a set of seed scales to represent the 12 recognized Clinical Scale core components (Ben-Porath & Tellegen, 2008/2011; Tellegen et al., 2003). To develop a

set of seed scales that would be statistically consistent yet representative, repeated analysis and refinement occurred in five steps. Items from all the Clinical Scales were selected for a particular seed scales if the item initially demonstrated its highest loading on the respective Clinical Scale core factor and lacked a high Demoralization loading. Next, most overlapping items were removed. Provisional seed scales were then created and items with item-scale correlations of less than .20 were removed. A second set of provisional seed scales was created and items were removed that did not demonstrate the highest average correlation with their seed scale across the four samples. Finally, the remaining 99 items formed the third and final set of 11 seed scales. The seed scale for Demoralization was created by removing four items that were only weakly correlated with the provisional scale.

Step four. In the final step of RC scale development, nine scales were constructed to represent demoralization (RCd) and the eight Clinical Scale areas, Hs (RC1), D (RC2), Hy (RC3), Pd (RC4), Pa (RC6), Pt (RC7), Sc (RC8), and Ma (RC9; Ben-Porath & Tellegen, 2008/2011). Since the RC scales were developed to measure core dimensions of psychopathology, RC scales were not constructed for Clinical Scales 0 (Si) and 5 (Mf; Tellegen et al., 2003). Tellegen et al. (2003) then conducted correlations between all of the 567 items on the MMPI-2 and the seed scales. Items with higher average absolute correlations to a specific seed scale when compared to their average absolute correlation to any of the other seed scales were provisionally assigned to that specific seed scale. A given item was only assigned to a specific seed scale if it had adequate convergent and discriminant properties for the target seed scale.

In further refinement, Scales RC7 and RC9 were examined to enhance the core of these scales and some items were removed (Tellegen et al., 2003). The internal consistencies of the scales were assessed and one item was removed based on its influence and the relevant alpha coefficients in the four samples. Finally, RC1, RC2, RC4, RC6, RC7, and RC8 were correlated with relevant external criterion and a small number of items were reassigned for scales RC3, RC6, and RC8. There were no suitable criterion measures for correlations with RC3 and RC9.

Psychometric properties of the RC Scales. The psychometric properties of the RC scales were investigated in several archival data sets, including men and women from the MMPI-2 normative group, a community mental health outpatient center, an inpatient psychiatric hospital, and male inpatients at a Veterans Administration Medical Center (Tellegen et al., 2003). Since the RC scales were created to improve upon the psychometric properties of the Clinical Scales, a majority of the psychometric research focused on comparing the scales.

The RC scales produced Cronbach's alphas ranging from .62 to .89 in the normative sample, .77 to .93 in a the community mental health sample, .82 to .95 in the inpatient sample, and .83 to .93 in the VAMC sample (Tellegen et al., 2003). Overall, the RC scales demonstrated comparable or greater internal consistencies in relation to the Clinical Scales. Test-retest reliabilities ranged from .74 to .88, with the exception of .62 for RC6. The developers noted that the lower test-retest reliability of RC6 may be related to its restricted variance in the samples.

Intercorrelations between RC and Clinical Scales were high, with the exception of RC3 and Clinical Scale 3 (Tellegen et al., 2003). The developers noted that this

correlation was expected due to the very heterogeneous nature of Clinical Scale 3. The correlations also tended to be higher between the RC and Clinical Scales in the clinical samples, due to increased variance. Overall, RCd correlated higher with the Clinical Scales than other RC scales, indicating that the first factor of Demoralization was noticeably removed from the RC scales. Interestingly, the correlation between RCd and RC9 increased slightly when compared to the correlation of RCd and Clinical Scale 9. The developers note that this may be due to the more focused nature of RC9 on the affective state of hypomania relative to the heterogeneous content of Clinical Scale 9. It is also important to note that the correlation between RCd and the other RC scales is not zero and thus some Demoralization component remains in the RC scales. With a few exceptions, the RC scales demonstrate less intercorrelation amongst themselves compared to the Clinical Scales.

To assess the convergent and discriminant validity of scores on the RC scales compared to the Clinical Scales, correlations were calculated between those scale scores and scores on a clinician-rated measure called the Patient Description Form (Graham, Ben-Porath, & McNulty, 1999) available for the outpatient sample (Tellegen et al., 2003). Scores on RC1, RC2, RC4, RC6, RC7, and RC8 and the Clinical Scales were correlated with variables extracted from medical records in the inpatient sample. RCd could not be compared to a related Clinical Scale and was instead examined for correlations to external criterion. Based on these correlations, RCd appeared most associated with depression and to a lesser extent anxiety. With the exception of RC6 and RC8, the aforementioned RC scales achieved greater or comparable convergent validity in all four samples. RC6 and RC8 did not demonstrate convergent validity in

the outpatient sample, likely related to the restricted population, but showed substantially increased convergent validity in the inpatient samples. All of the assessed RC scales achieved greater discriminant validity across the samples, apart from RC2, which demonstrated comparable discriminant validity in the inpatient sample compared to Clinical Scale 2.

RC3 and RC9 were not able to be examined in this way based on the lack of available criterion variables (Tellegen et al., 2003). The developers pointed out that a comparison of RC3 and Clinical Scale 3 would likely not be meaningful because RC3 represents only a portion of the dimensions assessed by Clinical Scale 3. They recommended more research on these two scales to help clarify the scales' convergent and discriminant validity.

External validity was further examined with regards to differences in the scales' ability to predict external criterion measures. To this end, each criterion was regressed on the best three RC and Clinical Scale predictors for that particular scale, as determined by a forward entry method. The RC scales demonstrated similar or improved prediction of the criterion variables relative to the Clinical Scales across a range of characteristics and psychopathology in all four samples. Specifically, the RC and clinical scales were similar in predicting internalizing psychopathology but the RC scales achieved better prediction of externalizing symptoms. Discriminant validity was examined by comparing correlations between each RC Scale and its corresponding Clinical Scale and external criterion variables that should not conceptually be strongly correlated with each targeted construct.

Concluding comments. While the RC scales represent an achievement in improving the psychometric functioning of the MMPI-2, they were not developed to be the sole means of profile interpretation (Ben-Porath & Tellegen, 2008/2011). Additional scales were needed to assess for dimensions originally captured in the Clinical Scales but not in the related RC Scale, clinically important characteristics not assessed by the RC Scales (e.g., suicidal ideation, fears), and facets assessed by *Mf* and *Si*. In fact, the RC Scales were actually the beginning of a massive initiative to revise the entire measure with a goal of improving the overall psychometric properties, enhancing efficiency, and improving construct validity (Ben-Porath, 2012).

The MMPI-2-RF. Based on the need for more diverse, yet psychometrically sound scales, the MMPI-2-RF (Tellegen & Ben-Porath, 2008/2011) was developed. In developing the MMPI-2-RF the authors report that their goal was to examine the MMPI-2 items and “identify potential targets for additional substantive scale construction that would result in a comprehensive set of scales yielding an efficient and exhaustive assessment of the most salient, clinically relevant variables measurable with the MMPI-2 item pool” (Tellegen & Ben-Porath, 2008/2011, p. 5). The MMPI-2-RF was built upon the foundation of the RC scales, as the same statistical techniques that resulted in the RC Scales (described above) were used to develop other scales on the MMPI-2-RF (Ben-Porath, 2012). The relevant item areas were factor analyzed, seed scales were created, and items were added from across the MMPI-2 item pool (Ben-Porath, 2012). While keeping the external correlates of the scales in consideration, the resulting scales were examined and tailored for maximum reliability, discriminant validity, and meaningfulness.

The resulting MMPI-2-RF is both theory-based and empirically informed and demonstrates strong psychometric properties (Ben-Porath, 2012; Tellegen & Ben-Porath, 2008/2011). The MMPI-2-RF is a more concise measure as well; reducing the item pool from 567 to 338 items (Ben-Porath & Tellegen, 2008/2011). The resulting MMPI-2-RF contains nine Validity Scales: VRIN-r; TRIN-r (both discussed above); Infrequent Responses (F-r; responses infrequent in the general population); Infrequent Psychopathology Responses (Fp-r; responses infrequent in psychiatric populations); Infrequent Somatic Responses (Fs; responses infrequent in medical patient populations); Symptom Validity (FBS; somatic and cognitive complaints associated with high levels of overreporting); Response Bias Scale (RBS; non-credible memory complaints); Uncommon Virtues (L-r; rarely endorsed moral attributes or activities); and Adjustment Validity (K-r; declarations of good psychological adjustment associated with high levels of under-reporting; Ben-Porath & Tellegen, 2008/2011). Three Higher-Order (H-O) Scales are also included on the MMPI-2-RF, including Emotional/Internalizing Dysfunction (EID; mood and affect problems); Thought Dysfunction (THD; disordered thinking difficulties); and Behavioral/Externalizing Dysfunction (BXD; problems related to under-controlled behavior). The RC scales (discussed above) remain intact in the MMPI-2-RF.

The MMPI-2-RF introduces twenty three Specific Problem Scales, discussed at length below. Finally, the MMPI-2-RF presents two Interest Scales, Aesthetic-Literary Interests (AES; literature, music, and theater interests) and Mechanical-Physical Interests (MEC; interests in fixing and building things, the outdoors, and sports). Harkness and McNulty (2007) revised the PSY-5 Scales for the MMPI-2-RF,

which include Aggressiveness-Revised (AGGR-r; instrumental, goal-directed aggression); Psychoticism-Revised (PSYC-r; disconnection from reality); Disconstraint-Revised (DISC-r; under-controlled behavior); Negative Emotionality/Neuroticism-Revised (NEGE-r; anxiety, insecurity, worry, and fear); and Introversion/Low Positive Emotionality-Revised (INTR-r; social disengagement and anhedonia). An additional Validity Scale, the Response Bias Scale (RBS), was added in 2011 (Ben-Porath, 2012; Ben-Porath & Tellegen, 2008/2011).

MMPI-2-RF Specific Problem (SP) Scales. Since the present study focuses on the MMPI-2-RF's Specific Problem (SP) Scales, a more thorough discussion of the SP Scales is warranted. The SP scales were developed to highlight characteristics included in or related to, yet not exclusively or saliently addressed by one of the RC scales (Ben-Porath, 2012). However, the SP scales do not serve an adjunctive role and should be interpreted independently of scores on the related RC scale (Ben-Porath & Tellegen, 2008/2011). Based on conceptual considerations and empirical analyses, four sets SP Scales were developed, including Somatic/Cognitive, Internalizing, Externalizing, and Interpersonal scales.

The Somatic/Cognitive SP scales assess symptoms related to physical and cognitive symptoms (Ben-Porath, 2012). Their interpretation should rest on the results of the Fs and FBS-r validity scales, which indicate possible over-reporting of somatic and cognitive symptoms (Ben-Porath & Tellegen, 2008/2011). Elevated scores on Fs and FBS-r may not indicate intentional over-reporting, as such item endorsements may be related to a genuine medical condition. However, in the case of a somatoform disorder and Fs and FBS-r scores of 100T or more, the items endorsed on the

Somatic/Cognitive Scales can provide distinct information regarding symptoms.

Attention to health information will aid in the interpretation of these scales.

The first Somatic Cognitive scale, the Malaise (MLS) scale consists of eight items and assesses a general sense of poor health and physical debilitation (Ben-Porath & Tellegen, 2008/2011). More specific complaints of poor appetite, nausea, and upset stomach are measured by the five items on the Gastrointestinal Complaints (GIC) scale. In the absence of extra-test health information indicating a related medical condition, the symptoms may be related to stress. The Head Pain Complaints (HPC) scale, which consists of six items, indicates complaints of head and neck pain. The Neurological Complaints (NUC) scale consists of ten items and measures reports of dizziness, weakness, and involuntary movement. An elevation of this scale may warrant neuropsychological or neurological evaluation. Finally, memory difficulties, problems concentrating, and confusion is assessed by the ten items of the Cognitive Complaints (COG) scale.

The Internalizing SP scales assess dimensions of two RC Scales, RCd and RC7 (Ben-Porath & Tellegen, 2008/2011). The Suicidal/Death Ideation (SUI), Helplessness/Hopelessness (HLP), Self-Doubt (SFD), and Inefficacy (NFC) Scales measure various aspects or correlates of RCd. The Stress/Worry (STW), Anxiety (ANX), Anger Proneness (ANP), Behavior-Restricting Fears (BRF), and Multiple Specific Fears (MSF) Scales assess aspects of RC7. The correlations between the scales that assess facets related to a RC scale are expectedly high. Nevertheless, each of the Internalizing Scales has demonstrated unique empirical correlates.

The first of nine Internalizing Scales is the Suicidal/Death Ideation (SUI) scale, which contains five items assessing for suicidal ideation or acts (Ben-Porath & Tellegen, 2008/2011). Particularly noteworthy, a raw score of one on SUI will produce an elevated score. Obviously, an elevation on this scale warrants a thorough suicide risk assessment. The Helplessness/Hopelessness (HLP) scale consists of five items and high scores indicate that the test taker feels overwhelmed and incapable of making changes in life. HLP is one of the scales with critical items on the MMPI-2-RF and thus any items keyed true will be printed in the Score Report. The Self-Doubt (SFD) scale, a four item scale, assesses for lack of confidence and feelings of uselessness. The Inefficacy (NFC) scale consists of nine items and measures beliefs about being incapable of coping with stress or making decisions. Preoccupation with disappointments and specific worries is assessed by the seven item Stress/Worry (STW) scale.

Another Internalizing Scale, the Anxiety (AXY) scale is a five item scale that evaluates pervasive anxiety, including intrusive ideation, sleep problems, and posttraumatic stress. An elevated AXY scale does not mean that the test taker has experienced a traumatic event (part of the criterion for a diagnosis of Posttraumatic Stress Disorder) but instead is highly indicative of a posttraumatic stress reaction if the person has experienced a traumatic event. AXY items were not endorsed very often by the normative sample and thus a raw score of two results in an elevated score. Based on the item content of the AXY scale, it is a critical scale and endorsed items will print on the Score Report. The Anger Proneness (ANP) scale contains seven items assessing tendencies to become easily upset and impatient. ANP correlates

involve more negative emotional experience and expression of anger rather than aggressive acting-out behavior. The Behavior-Restricting Fears (BRF) scale contains nine items and assesses fears restricting behavior in and out of the home. Finally, distinct fears of animals and acts of nature are evaluated by the nine items of the Multiple Specific Fears (MSF) scale. In addition, test takers with elevated MSF scores will likely avoid taking risks.

The Externalizing SP scales relate to RC4 and RC9 and include scales assessing adolescent conduct problems, substance abuse, aggression, and activation (Ben-Porath & Tellegen, 2008/2011). The Externalizing scales can be used to clarify elevations on RC4 and RC9 and as previously mentioned, should be interpreted independent of RC Scale elevations. The Juvenile Conduct Problems (JCP) and Substance Abuse (SUB) Scales assess components of RC4. Aggression (AGG) and Activation (ACT) measure areas of RC9. The Juvenile Conduct Problems (JCP) scale, a six item scale, assesses undesirable school conduct, stealing, and negative peer influence. An elevated JCP Scale score can be associated with juvenile delinquency and current acting out behavior. However, if JCP is the only elevated behavioral dysfunction scale, the test taker may have a history of juvenile conduct problems but may no longer engage in such behaviors.

The second Externalizing scale, the Substance Abuse (SUB) scale consists of seven items measuring past or current substance abuse. A test taker with a known history of substance abuse who does not produce an elevated SUB score may be in denial regarding his/her abuse. SUB is another scale with critical items and thus endorsed items will print out on the Score Sheet. The Aggression (AGG) scale

contains nine items that measure physically aggressive behavior. An elevation on AGG may indicate a history of interpersonal violence and abusiveness. Based on its content, AGG is another scale deemed to have critical items. The final Externalizing Scale, Activation (ACT) contains eight items and measures excessive excitation and energy level, mood swings, and limited sleep. An elevated ACT score may indicate a hypomanic or manic episode but substance-induced activation should also be considered.

While all of the scales on the MMPI-2-RF have implications for interpersonal functioning, the Interpersonal SP scales place a range of interpersonal functioning at the forefront (Ben-Porath & Tellegen, 2008//2011). The Family Problems (FML) scale's ten items measure negative family experiences, past, present, or both. The Interpersonal Passivity (IPP) scale assesses unassertive, passive, submissive behavior. A low FML score indicates a conflict-free family environment. The Interpersonal Passivity (IPP) scale contains ten items that describe unassertive, submissive behavior, failure to assert oneself, the lack of strong opinions, and not liking to take charge. A low score on the IPP scale indicates beliefs that one has leadership ability but likely is perceived by others as domineering or self-centered.

Another Interpersonal Scale, the Social Avoidance (SAV) scale contains ten items and evaluates avoidance of social situations and social introversion. Alternatively, low SAV scores may indicate that the test taker enjoys social situations and is outgoing. Interestingly, an elevated SAV score paired with a non-elevated Shyness (SHY) score designates that the social avoidance is perhaps more linked to an avoidant personality style rather than social anxiety (particularly if SFD and NFC are

elevated as well). SHY, a seven item scale, assesses for social anxiety, including being easily embarrassed and feeling uncomfortable around other people. Given other information, an elevated SHY score may indicate a social phobia. A low SHY score indicates the lack of social anxiety and a normal range of personality characteristics. However, paired with other elevations, a low SHY score may be indicative of psychopathic tendencies or conversion disorders. Finally, the Disaffiliativeness (DSF) scale contains six items and measures a dislike of people, lack of close relationships, and preference to being alone. If the DSF scale is extremely elevated (score of 100T or more), the test taker may meet criteria for schizoid personality disorder.

A History of Test Bias Research

The issue of bias in testing has a long history in psychological assessment literature. Cole (1981) discusses the issue of test bias as emerging from social concern with equality. Such concern has then led to questioning a variety of other issues in social life and policy, of which psychological testing may or may not have an impact. While Cole (1981) outlines a number of different types of test bias, a more recent article (Millsap, 1997) condenses past literature on test bias and identifies the two most distinguishable and recently researched forms of test bias, predictive and measurement bias. Of note, early research into test bias often simply examined score differences between groups.

Measuring test bias via the prediction of external variables. Predictive bias can be seen when a test leads to systematic inaccuracies in the prediction of an external variable based on group membership (Millsap, 1997). Predictive bias is typically investigated in one of two ways. One way to examine the possibility of

predictive bias involves investigating whether the predictor systematically under- or overpredicts the criterion variable for the different groups (Anastasi & Urbina, 1997; Nunnally & Burnstein, 1994). This form of test bias, commonly referred to as Intercept Bias, was introduced by Cleary (1968). It is typically investigated using moderated multiple regression. In this method, a series of regression analyses are conducted and the resulting change in R^2 is examined (Mattern & Patterson, 2013). The first model uses just the criterion and predictor variables and the second model adds group membership as a criterion variable. If the R^2 change after adding the group membership variable is significant, the test is reported to demonstrate intercept bias.

Another way to assess for predictive bias involves examining the slope of the regression line between the predictor and criterion variables for different groups, known as assessing for slope bias (Anastasi & Urbina, 1997; Nunnally & Burnstein, 1994). This occurs when there is a difference in the magnitude of the correlation between the predictor and criterion for the different groups and suggests a bias in the prediction accuracy across the range of predictor scores (Arbisi et al., 2002). In this case, an interaction term is created between the group membership variable and the predictor variable (Mattern & Patterson, 2013). The interaction term is then added to the model that already contains the predictor and group membership variable. Slope bias is said to exist when the addition of the interaction terms results in a significant change in R^2 .

Measuring test bias via measurement bias. Measurement bias involves systematic inaccuracies in the data a test provides about a characteristic or latent

variable based on group membership (Millsap, 1997). Put another way, measurement bias is present if two people from different groups are indistinguishable on the latent variable but produce different scores on the test measuring that latent variable. This is an internal type of bias and does not require the use of any external criterion variables. Testing for measurement invariance involves confirmatory factor analysis, both of which are described in more detail below.

Test Bias Research on the MMPI/MMPI-2 with Minority Populations

As previously mentioned, the norms for the original MMPI were based on Caucasian visitors to the University of Minnesota hospital (Handel & Ben-Porath, 2000). The sample was from a rural background with an average of eight years of education. Multicultural issues were almost completely ignored in the early years after the MMPI's publication but eventually research began examining questions of culture with regard to the normative sample. Generally speaking, research began exploring the question of test bias by focusing on mean score differences and evolved into examining external correlates.

The majority of multicultural research on the MMPI/MMPI-2 has concentrated on the differences between African American and Caucasian samples (Handel & Ben-Porath, 2000). Since the current research focuses on evaluation of the Internalizing SP scales in African American and Caucasian samples, related research will be more thoroughly explored in later sections. Instead, this section will briefly outline the history and current state of MMPI research with other minority populations, including Hispanic Americans, Native Americans, and Asian Americans. Importantly, this

author was unable to find any research examining the presence of test bias in MMPI-2-RF in Hispanic Americans, Native Americans, or Asian Americans.

Hispanic Americans. First, and of significant importance, research with Hispanic Americans is difficult to interpret based on the heterogeneity of the people categorized as Hispanic Americans and the potential confound of language proficiency. Greene (1987) examined 10 published empirical studies examining differences in MMPI scale scores between Hispanic and Caucasian groups. Results indicated that although significant differences existed, there was no pattern to the differences. Campos (1989) found that Hispanics consistently score four T-score points higher on the L scale when compared to Caucasians. However, given the limited information, results did not indicate that the MMPI's predictive ability for job performance was impacted.

A number of studies have demonstrated that although differences exist in scores, characteristics and profiles are often similar between Hispanic and Caucasian psychiatric samples with the same diagnoses (Velasquez, Callahan, & Carrillo, 1989; Velasquez, Callahan, & Carrillo, 1991). For example, Velasquez and Callahan (1990a) investigated MMPI scale score differences between Hispanic and Caucasian populations with alcoholism. Results indicated that although the Hispanic sample scored significantly lower on Scales 4, 5, and 0 when compared to Caucasians, their profile patterns were similar. In another study, Velasquez and Callahan (1990b) reported similar findings with Hispanic and Caucasian patients diagnosed with schizophrenia.

In yet another study of the MMPI, groups of male Hispanic and Caucasian patients diagnosed with schizophrenia, major depression, or antisocial personality disorder were compared (Velasquez, Callahan, & Young, 1993). After statistical correction, only a few differences emerged. The Hispanic patients with schizophrenia scored higher on scale 1 when compared to the Caucasian patients with schizophrenia. For the groups diagnosed with major depression, the Hispanic sample scored lower than the Caucasian sample on scale 5. No significant differences were found between Hispanics and Caucasians in the antisocial personality disorder groups.

With regards to the MMPI-2, limited research is available (Graham, 2006). An official Spanish-language translation of the MMPI-2 is available, which may contribute to the lack of research comparing Hispanics and Caucasians on the English language MMPI-2. However, Graham (2006) examined the normative sample's scores for Hispanics and Caucasians that is presented in the MMPI-2 manual (Butcher et al., 1989). First, Graham (2006) noted that given the geographic locations from which the data was collected, it is probably more accurate to classify the sample as Mexican-American. Although differences existed between Hispanic and Caucasian men, none of these differences exceed five T-score points. When comparing Hispanic and Caucasian women, scale score differences of more than five T-score points emerged for scales F, 1, 4, 7, 8, and 9. However, neither the men or women groups were matched for age or education.

Research has reported differences between Hispanic and Caucasian college students on particular validity and clinical scales but again, none of these differences were greater than five T-score points (Hall, Bansal, & Lopez, 1999; Whitworth &

McBlaine, 1993; Whitworth & Unterbrink, 1994). However, differences between the samples of more than five T-score points were found on two of the MMPI-2 Content Scales, Family Problems (FAM) and Cynicism (CYN; Whitworth and Unterbrink, 1994). Velasquez, Ayala, & Mendoza (1998) completed a review of more than 170 studies exploring the MMPI in Hispanic populations and reported higher scores for Hispanic samples on some MMPI/MMPI-2 scales. However, a number of the studies were unpublished and thus difficult to assess and did not provide the data needed to explore the meaning of the results.

Interestingly, research differs with regard to the impact of acculturation on MMPI-2 scores. Some results have indicated that higher L scores are associated with lower acculturation (Canul & Cross, 1994), while other research has demonstrated no relationship between acculturation and MMPI-2 scores (Lessenger, 1997). In all, the research on the MMPI-2 with Hispanic Americans is limited and does not allow for adequate conclusions. That being said, Graham (2006) recommends considering that moderate elevations may be a result of acculturation and interpreting the L scale with care.

Native Americans. A review of seven studies comparing MMPI scores of Native Americans and Caucasians demonstrated that while Native Americans tended to score higher on some of the clinical scales, no pattern emerged in the differences (Greene, 1987). A very early study conducted by Arthur (1944) found more similarities than differences between groups of Native American and Caucasian young adults and college students. A study of native and nonnative Alaskan college students

yielded one scale difference that was greater than 5 T-score points; scale 5 was higher in native woman when compared to nonnative woman (Herreid & Herreid, 1966).

A number of studies have examined the MMPI scores of Native Americans with alcoholism to other populations (Graham, 2006). Although early research concluded that Native Americans with alcoholism have more deviant MMPI scores when compared to Caucasians with alcoholism (Klein, Rozytko, Flint, & Roberts, 1973), other studies have found comparable scores between the groups (Page & Bozlee, 1982; Venn, 1988) with Caucasians scoring higher in one study on scales 4 and 5 (Uecker, Boutilier, & Richardson, 1980). Notably, two studies found no difference between the groups on the MacAndrew Alcoholism Scale (MAC-r; Page & Bozlee, 1982; Uecker, Boutilier, & Richardson, 1980). However, Lapham et al. (1995) found that a higher percentage of Native Americans with their first DWI offense elevated the MAC-r when compared to Caucasians with their first DWI offense. Graham (2006) points out that no data concerning alcohol use/abuse between the groups was available and thus we are not sure whether this finding reflects test bias or underlying real world differences.

Research evaluating the scores of Native Americans on the MMPI-2 is sparse. In examining the MMPI-2 manual's normative sample, which contained 77 Native Americans, Graham (2006) points out that Native American men scored more than five T-score points higher on scales F and 4 when compared to Caucasian men. When comparing Native American and Caucasian women in the normative sample, score differences of more than five T-score points emerged on scales F, 1, 4, 5, 6, and 8.

Again, the data does not allow for evaluation of whether these differences reflect test bias or real world differences.

Two recent studies examined the mean T-scores of a different Native American samples compared to the MMPI-2 normative standard of T-score = 50 and found clinically significant differences on a range of Clinical, Harris-Lingos, Supplemental, and Content Scales (Lacey, 2004; Prewett, 2012). Interestingly, 14% and 33% of the variance in MMPI-2 scores was accounted for by the linear combination of assessed demographic variables (i.e., age, gender, level of education, socioeconomic status, languages spoken, and cultural identification). In the latter study, the standard deviation of the Native American test taker's mean T-scores overlapped with the MMPI-2 normative standard. The former study did not report standard deviations or standard errors of the Native American test taker's T-scores.

A large scale study compared the MMPI-2 Validity, Clinical, Content, and Supplementary scales of 535 Southwestern and 297 Plains Native Americans with the MMPI-2 normative sample (Robin, Greene, Albaugh, Caldwell, & Goldman, 2003). Surprisingly, no differences were found between the two Native American tribes. However, several differences were evident in comparing the scores of the combined Native American sample with the normative sample. Native Americans scored more than 5 T-score points higher on scales L, F, 4, 8, 9, five content scales, and the two alcoholism scales. As a follow-up to this study and using the same data, Greene, Robin, Albaugh, Caldwell, and Goldman (2003) examined correlations between the MMPI-2 scores and measures of symptoms and behaviors. Results indicate that the majority of the MMPI-2 scales correlated with the expected measures. This indicates

that the differences noted in MMPI-2 scale scores may be more related to real world differences in the symptoms or characteristics and not test bias. Notably, the revised MAC-r scale was not appropriately correlated with other measures of substance problems, which provides further support to apprehensions about its use with Native Americans (Greene et al., 2003).

A more recent study compared Eastern Woodland Oklahoma (EWO), Southwest Plains Oklahoma (SWPO) Native Americans, and the MMPI-2 normative sample (Pace et al., 2006) on MMPI scale scores. Results indicate that only differences in the F scale were clinically significant between the two Native American groups. Clinically significant differences were found in six Clinical Scales in comparing the mean T-scores of the SWPO tribe to the normative standard T-score and clinically significant differences emerged in one Clinical Scale when comparing the mean T-scores of the EWO tribe to the MMPI-2 normative standard.

In further analysis, EWO tribe test takers with low education scored clinically significantly higher on the L scale than EWO tribe test takers with higher education (Pace et al., 2006). In the EWO tribe sample, low acculturation test takers demonstrated clinically significantly higher scores on scale F and 8 when compared to their highly acculturated counterparts. While differences existed in mean T-score scores between the two Native American groups and the normative group, it seems that such differences may reflect differences in symptomology, behavior, and characteristics related to culture. This does not dismiss the need for careful consideration of MMPI-2 scores in Native American groups, particularly with the

evidence that education and acculturation may affect scores, but also does not provide evidence that the MMPI-2 is biased in the assessment of Native Americans.

Using the same EWO tribe sample, Hill, Pace, and Robbins (2010) examined the difference in item endorsement between the tribe and the MMPI-2 normative sample. Using item analysis and a conservative alpha, results indicated that 27 of the 113 items examined were endorsed significantly more and 3 of the 113 items were endorsed significantly less in the EWO tribe group when compared to the normative sample.

More research is needed with Native American populations and the MMPI-2/MMPI-2-RF, particularly with regard to examining predictive bias and comparing scale scores to related external characteristics. Overall, Graham (2006) states that clinicians should expect Native Americans to score moderately high on a number of MMPI-2 scales, reflective of cultural differences. However, T-scores above 65 on the Clinical and Content Scales should be interpreted the same in Native Americans and Caucasian test takers. Based on the above research, interpretation of the revised MAC-r scale should be done so cautiously with Native American test takers.

Asian Americans. As with other minority groups, research with Asian Americans is difficult to interpret due to the heterogeneity of populations labeled Asian American and potential language proficiency confounds. Sue and Sue (1974) compared the MMPI scores of Chinese and Japanese and non-Asian students from a psychiatric center and found that the Asian sample scored higher on scales L, F, 1, 2, 4, 6, 7, 8, and 0. Another study found that Chinese and Japanese college students living in Hawaii had higher scores on scale 2 when compared to Caucasians (Marsella,

Sanborn, Kameoka, Shizuru, & Brennan, 1975). Other studies have also found differences in MMPI scores between the groups (Lee, Cheung, Man, & Hsu, 1992; Kwan, 1999), while others have found these differences to be small, not clinically meaningful, or accounted for by other variables (e.g., diagnoses; Greene, 1987; Tsushima & Onorato, 1982; Tsushima & Stoddard, 1990). However, Graham (2006) notes that the most consistent finding is that Asian Americans score meaningfully higher on scale 0, suggestive of a higher degree of social introversion.

Asian Americans were not well represented in the MMPI-2's normative data. As such, some have questioned the applicability of such norms to Asian Americans (Kwan, 1999). Some research has uncovered statistically significant differences between MMPI-2 Validity, Clinical, and Supplementary Scale scores of Chinese American and foreign Chinese students when compared to Caucasian students but noted that while some scores were in the moderately elevated range, none of the scores were in the clinically pathological range (Robens, 1992; Stevens, Kwan, & Graybill, 1993; Telander, 1999). Some research has pointed to acculturation as a factor potentially influencing MMPI score differences of Asian Americans (Okazaki & Sue, 1995; Tsai & Pike, 2000; Sue, Keefe, Enomoto, Durvasula, & Chao, 1996). A more recent study investigated differences in Asian American and Caucasian personal injury or compensation litigation test takers and found no significant T-score differences related to race on five Validity Scales (Tsushima & Tsushima, 2009).

Graham (2006) recommends that clinicians expect moderate elevations (T-scores between 50 and 60) on the MMPI-2 scales when testing an Asian American client. Such elevations are likely more the product of stress or level of acculturation

rather than psychopathology. T-scores above 65, however, should be interpreted as usual.

Of important note, the research discussed above has focused on minority populations within America. The MMPI-2 has been translated into 21 different languages and the MMPI-2-RF has been translated into four different languages (University of Minnesota Press, 2011). Research is ongoing regarding the reliability and validity of translated MMPI-2/MMPI-2-RFs.

Test Bias Research on the MMPI/MMPI-2 with African American Populations

As previously mentioned, the majority of multicultural research on the MMPI and MMPI-2 has been focused on differences between African Americans and Caucasians (Handel & Ben-Porath, 2000). This section will expand upon the history and current state of research examining potential test bias in the MMPI/MMPI-2 in African American populations.

MMPI research. Greene (1987) summarized the MMPI research to date examining MMPI performance of African American samples. While the specific studies will be discussed in more detail below, Greene concluded that no consistent pattern of differences can be seen across the studies in particular populations (e.g., inpatient, non-patients, forensic, etc.).

Harrison and Kass (1967) examined mean T-score and item differences in African-American and Caucasian pregnant women from a socioeconomically underprivileged area around Boston City Hospital. Such comparison demonstrated significant differences in T-scores between the groups on the scales Cannot Say (CNS), F, 1, 8, and 9. Of the 550 items on the original MMPI, this study found that

213 items discriminated between the groups at a .05 significance level. In comparing scores of rural and isolated African Americans to Caucasian samples, African-Americans scored higher on scales F, 4, 5, 6, 7, 8, 9, and 0 (Gynther, Fowler, & Erdberg, 1971). However, African American and Caucasian groups receiving welfare for dependent children did not differ on MMPI scores (McGill, 1980).

Ball (1960) found that when compared to Caucasian high school students, African American high school students scored higher on Scales F, 1, 8, and 0. Further research has found differences on Scales F, L, and Content Scale CYN between low income African American and Caucasian adolescents (Moore & Handal, 1980). Caucasian students scored higher on Scales K and CYN. Along the same lines, McDonald and Gynther (1962) found significant differences on multiple scales between African American and Caucasian high school students. Interestingly, they found differences in multiple comparisons of ethnicity and gender (e.g., African American men and Caucasian men, African American women and Caucasian women) and even between the two genders, combining the ethnic groups.

Research has shown that demographic variables, such as age, sex, education, institutional differences, and socioeconomic level, affect African American's performance on the MMPI (Butcher, Ball, & Ray, 1964). Even while controlling for these variables, differences between the groups remained in scales L, 6, and 9. In another study that controlled for such variables, African-Americans scores higher than Caucasians on Clinical Scale 9 while Caucasians scored higher on Clinical Scale 2 and 6 (King, Carroll, & Fuller, 1977). However, the latter study did not find any significant differences and all scores fell within the normal range. Controlling for

socioeconomic status did not eradicate the mean T-score differences in a sample of African American and Caucasian high school students (McDonald & Gynther, 1962). In this study, African Americans scored higher on multiple scales when compared to Caucasians but men also tended to score higher on multiple scales when compared to women, ethnicity aside.

When controlling for gender, age, residence, employment, education, marital status, socioeconomic status, and hospital status, no differences were found on MMPI scales, items, high-points, or elevations between African American and Caucasian psychiatric patients (Bertelson, Marks, & May, 1982). While Davis (1975), Davis and Jones (1974), and Davis, Beck, and Ryan (1973) found different MMPI scores based on diagnoses and education in an inpatient population, no differences in the scales investigated emerged related solely to ethnicity. Further, Miller, Wertz, and Counts (1961) found demographic factors to account for more variance in MMPI scores than ethnicity.

An interesting study compared the MMPI scores African Americans and Caucasians upon admission to an inpatient psychiatric hospital, at discharge, and at an 18-month follow-up visit (Genthner & Graham, 1976). While differences existed between the groups at admission, these disappeared at discharge and 18-months post-hospitalization, suggesting that the groups do not respond differently to treatment. In examining external correlates of the F scale between African American and Caucasian inpatients, researchers found that African American and Caucasian inpatients did not significantly differ on the scale and the scale measures similar characteristics in both groups (Smith & Graham, 1981). This study even attempted to create an alternate

MMPI F scale based on profiles of non-patient African Americans but the scale did not relate to external correlates.

Conversely, another study found differences both with controlling and not controlling for socioeconomic status on Scales F, 6, 8, and 9 between African American and Caucasian inpatients (Butcher, Braswell, & Raney, 1983). While controlling for demographic variables, Costello, Fine, and Blau (1973) found that African American women in a psychiatric hospital scored higher on a number of scales when compared to Caucasian women. African American men scored higher on only the F scale relative to Caucasian men. Another study found that while differences in scale scores between hospitalized African Americans and Caucasians were not significant, African American participants were overrepresented in the small subsample that produced extreme elevations (Liske & McCormick, 1976). Other research has found differences in African-American and Caucasian profiles and code types in psychiatric populations (Costello, Tiffany, & Gier, 1972; Miller, Knapp, & Daniels, 1968). The earlier of this research found similar mean profiles but differences in elevations on scales 5 and 8 and 1-8/8-1 and 2-7/7-2 code types (Miller, Knapp, & Daniels, 1968). Costello, Tiffany, and Grier (1972) found that African Americans tended to elevate more scales than Caucasians. The most common code type for African Americans was 8-6 and 2-4, while Caucasians produced more 2-7 and 4-7 codes.

A very early study found differences between young African American and Caucasian inmates on Scales 5 and 9 (Caldwell, 1953) while another found no differences in similar groups (Stanton, 1956). In examining MMPI scores in

individuals being assessed for competency to stand trial, Cooke, Pogany, and Johnson (1974) found that although African Americans were assessed as having greater psychopathology when compared to Caucasians, MMPI scores did not differ significantly. Costello, Fine, and Blau (1973) found no differences in the MMPI scores of African American and Caucasian prison inmates. Holland (1979) found that incarcerated African Americans tended to score higher on Scales F, 8, and 9 when compared to their Caucasian counterparts. When controlling for socioeconomic status, African American inmates and forensic patients only scored higher on Clinical Scale 9 relative to Caucasian inmates and forensic patients (Flanagan & Lewis, 1969; Holcomb & Adams, 1982). Differences on Scales K, 3, and 9 remained between the groups when controlling for education and occupation (McCreary & Padilla, 1977). Other research has also highlighted the importance of controlling for such variables (Rosenblatt & Pritchard, 1978).

While looking at the difference in MMPI scores of African American and Caucasian inmates with a history of recidivism compared to those without such a history, scales differences emerged across groups (Ingram, Marchioni, Hill, Caraveo-Ramos, & McNeil, 1985). When controlling for age, IQ, and socioeconomic status, African Americans without a history of recidivism scored significantly higher than the other three groups. African Americans with a history of recidivism scored higher on the F Scale than both groups without a history of recidivism.

In comparing the MMPI scores of African American and Caucasian men and women residents of a substance abuse program, results indicate that Caucasian participants scored higher on Scales 1, 3, 7, and 0 while African Americans

participants scored higher on the L scale (Patalano, 1978). Sutker, Archer, and Allain (1978) found that Caucasians scored higher on scales F, 2, 6, and 7 when compared to African American in a residential drug abuse treatment program. Along the same lines, a study comparing African American and Caucasian men and women from two different substance abuse treatment centers found consistently higher elevations across scales for the Caucasian sample when compared to the African American sample (Sutker, Archer, & Allain, 1980). In fact, the only differences in elevations occurred for African American women on scale 5 and for one group of African American men on Clinical Scale 9.

However, when controlling for demographic variables, no clinically meaningful differences emerged in test scores between African Americans and Caucasians with alcohol abuse (Patterson, Charles, Woodward, Roberts, & Penk, 1981). Yet, in controlling for similar confounding variables, other research demonstrated that African Americans score lower on Scales 2, 3, 4, and 7 when compared to Caucasians seeking treatment for polysubstance abuse (Penk et al., 1982). Similarly, African Americans tended to score lower on scales F, 2, 4, 7, 8, and 0 when compared to Caucasians seeking treatment for heroin addiction when controlling for such variables (Penk, Woodward, Robinowitz, & Hess, 1978). Higher, but not clinically significantly higher, scores on scales 2 and 7 have also been observed in Caucasians in drug abuse treatment relative to their African American counterparts (Weiss & Russakoff, 1977).

Interestingly, one study found no differences between the MMPI scores of male African American and Caucasian with alcoholism but found that the MMPI may

have difficulty detecting alcoholism in African Americans (Walters, Greene, & Jeffrey, 1984). Since the code type most associated with alcoholism was a 2-4/4-2 combination, the researchers were surprised when only the Caucasian group obtained this pair of elevations. In fact, only the Caucasian group obtained significantly more elevations on Clinical Scale 4 when compared to the African American and Caucasian control group.

Some of the earliest research on the MMPI examined mean T-score differences between African-American and Caucasian veterans admitted to a Wisconsin Veterans Affairs Hospital for tuberculosis (Hokanson & Calden, 1960). Significant differences were found between the groups on scales L, F, 4, 5, 8, and 9. However, the differences were interpreted as socioeconomic experiences rather than being the result of test bias. Millsap (2011) provided an example of measurement invariance using MMPI data collected from African American and Caucasian adolescents from 1964 to 1965. The example examined an Assertiveness factor scale, created based on factor analysis and not in regular use, and found different item functioning in the two groups.

Overall, research on differential MMPI scores between African American and Caucasian populations varies greatly. Some research points to greater scores for African American samples, while other finds no meaningful differences. Some research, particularly with substance abuse populations, demonstrates higher scores for Caucasian samples. However, numerous methodological issues plague this research. Greene (1987) outlines a host of methodological problems prominent in such research. First, some studies do not adequately report participants' demographic characteristics and settings. He also outlines problems and inconsistencies in the research with

regard to assessing membership in and identification with a particular ethnic group.

Other issues include not excluding invalid protocols, inappropriate analysis, and using insufficient sample sizes. Moderator variables, the type of scores analyzed, and effect sizes are often neglected.

Greene's (1987) most salient point involves empirical correlates. While mean score or item differences may exist between the groups, such differences do not necessarily automatically equate with test bias. Such differences instead may simply reflect underlying group differences in symptoms or setting (Archer, Griffin, & Aiduk, 1995). Indeed, Prichard and Rosenblatt (1980) discussed the difficulties of relying solely on mean score differences in examining test bias. The issue of statistical significance also comes into play when discussing mean T-score differences (Greene, 1987). T-score differences of less than five points are not likely to be clinically meaningful. However, such differences may still be statistically significant. In reviewing the aforementioned research, it is clear that few studies examined empirical correlates when investigating test bias on the MMPI.

A more recent study illustrated the ability to assess for measurement bias, rather than using mean T-scores to examine group differences, in homogenous and heterogeneous scales of the MMPI (Waller, Thompson, and Wenk, 2000). While a more technical discussion of measurement bias and measurement invariance follows, it is important to note that measurement bias and measurement invariance research uses latent variables in addition to observed variables and has the ability to provide estimates of and constrain latent variables. Although the authors used more advanced statistical techniques, including Item Response Theory to evaluate for potential

differential item functioning, the study used MMPI data collected between 1964 and 1965 to illustrate the analysis. Results demonstrate evidence of differential item functioning, or bias at the item level, on an average of 38% of the items on Clinical Scales 1, 2, 3, 4, 6, 7, 8, 9, 0 and Validity Scales L, F, and K. However, the authors pointed to the fact that differential item functioning may or may not produce bias in the respective scales. Since no bias was found amongst scales in this analysis, the differential item functioning may not be important to scale interpretation.

MMPI-2 research. While the MMPI-2 is a revision of the MMPI, continuity was a main objective. Thus, the aforementioned studies on the MMPI can still more or less be evaluated as they may apply to the MMPI-2. The differences between MMPI-2 scores of African American and Caucasian populations remained a major area of research. For example, Hall, Bansal, and Lopez (1999) undertook a meta-analysis of 25 MMPI and MMPI-2 studies examining test bias between African American and Caucasian test takers from multiple settings. For African American males, results point to higher scores on Scales L, F, K, 1, 7, 8, and 0 and lower scores on scales 2, 3, 4, 5, and 9 relative to Caucasian men. African American women demonstrated higher scores on Scales L, F, 1, 2, 4, 5, 6, 7, and 8 but lower scores on scales K, 2, 3, and 9 when compared to Caucasian women. However, the aggregate effect sizes for both men and women were small. Also, this study is obviously plagued by some of the issues faced by earlier research and outlined above (i.e., statistical versus clinical significance, lack of external correlates) as well as varied study procedure (i.e., all studies did not control for demographic variables).

In a study using the MMPI-2 normative sample, African American men were found to score higher on Clinical Scale 8 relative to Caucasian men while African American women scored higher on Scales 4, 5, and 9 (Timbrook & Graham, 1994). However, all of the mean differences were less than 5 T-score points, indicating that the findings are likely not clinically meaningful. In examining external correlates, researchers used partner provided ratings given during the MMPI-2 normative group test administration. Mean error scores were computed comparing African American and Caucasian men and women for the scales with external correlates, scales, 2, 4, 7, 9, and 0. While no significant differences emerged between African American and Caucasian male's error scores, the authors note that a general pattern of negative error scores indicating minor underprediction can be seen in the male African American group. When comparing African American and Caucasian women, a significant difference in error of prediction emerged wherein Clinical Scale 7 underpredicted partner ratings of anxiety for the African American group of women. No other comparisons were statistically significant and the general pattern of negative error scores also indicated slight underprediction of ratings in the African American women group.

Frueh, Smith, & Libet (1996) compared raw scale scores of male African American and Caucasian veterans seeking outpatient treatment for posttraumatic stress disorder at a Veterans Affairs Hospital. Results indicate that African Americans scored statistically significantly higher on the F-K index and scales 6 and 8. Conversely, a later study examining test bias using a similar sample of male African American and Caucasian veterans seeking outpatient treatment for posttraumatic stress

disorder did not find any statistically or clinically significant differences between the groups (Frueh, Gold, de Arellano, & Brady, 1997). It is important to note that neither of these studies employed external correlates, so the presence or lack of score differences may or may not be attributable to test bias or differences in psychopathology.

To assess predictive bias and separate the mean score differences versus greater psychopathology issue, researchers have used external criterion variables. One such study used the Record Review Form, which provides a range of external variables obtained from admission summaries, mental status exams, and discharge summaries (Arbisi et al., 2002). In men, 32 comparisons between scales and these external variables demonstrated bias. Nonetheless, all produced small effect sizes. Interestingly, overprediction for African American men was only noted for the comparison of Clinical Scale 2 and being on antidepressants, Clinical Scale 8 and being on antidepressants, Clinical Scale 9 and a bipolar disorder diagnosis, and the Content Scale DEP (Depression) and being on antidepressants. For women, 12 comparisons demonstrated bias. Overprediction for African American women was only noted for the comparison of Clinical Scale 4 and an Axis II diagnosis, Clinical Scale 9 and a bipolar diagnosis, and the Supplementary Scale APS (Addiction Potential) and an Axis II diagnosis. However, it is important to note that all of the other comparisons that demonstrated bias (i.e., 28 comparisons in men and 9 in women) evidenced underprediction of psychopathology in the African American participants.

Also using external criterion variables, researchers assessed predictive bias in the MMPI-2 Clinical and RC scales in a community mental health outpatient population (Castro et al., 2008). The external variables in this study came from a brief application and interview. Mean T-score comparisons revealed significantly higher scores for African Americans on Clinical Scale 1 and RC Scales 1, 3, 6, and 8. All but one of the differences was greater than five T-score points. Regressions using the F scale, Clinical Scales 1, 4, and 8, and RCd, RC1, RC4, and RC8 were performed. Only these scales could be used based on the available external criterion. This analysis did not find any evidence of predictive bias related to ethnicity.

Using a varied sample of African American and Caucasian clients at an outpatient community health center, McNulty and colleagues (1997) compared mean T-score differences and correlations to external criterion variables between African American and Caucasian populations. Solely focusing on clinically meaningful differences in T-scores, African American men scored higher on the L scale when compared to Caucasian men and African American women scored lower on the Content Scale LSE relative to Caucasian women. External correlates were provided in the form of the patient description form, a therapist-rating scale. No differences between the groups in the comparisons of the scales and patient description form ratings were noted.

Mean scale differences were also explored in a sample of African American and Caucasian veterans residing in an inpatient facility (Munley, Morris, Murrary, & Baines, 2001). No statistically or clinically significant differences were found between the scores of the two groups with regards to the Validity or Clinical Scales.

A statistically significant multivariate effect was found in comparing the Supplementary Scale scores of the two groups but no significant univariate effect emerged. However, African American participants tended to score higher on Clinical Scales FRS (Fears), BIZ (Bizarre Mentation), CYN, and ASP (Antisocial Practices) relative to their Caucasian counterparts. All but the ASP scale differences were clinically meaningful with T-score differences greater than five points.

Schinka, Lalone, & Greene (1998) used a subsample of the MMPI-2 normative sample and two inpatient samples to investigate the role demographic variables, including ethnicity, have on MMPI-2 scores. Using multiple linear regression, results indicate that demographic variables contribute less than 10% of the incremental score variance on the Validity and all but one Clinical Scales. More than 10% of the score variance on Clinical Scale 5, Content Scale FRS and ASP, and five Supplementary Scales was attributed to demographic variables. It is important to note, however, that the majority of variance related to the demographic variables was influenced by gender.

The MMPI-2 scores of African Americans and Caucasians has also been examined in forensic populations. In comparing such groups who were assessed for a court-ordered forensic evaluation, Ben-Porath, Shondrick, and Stafford (1995) found that African American participants produced clinically significantly higher scores on Content Scales CYN and ASP relative to their Caucasian counterparts. Nevertheless, it remains unclear whether these differences represent test bias or underlying differences in psychopathology between the groups.

Fortunately, predictive bias has also been investigated in this population employing external variables obtained from a forensic assessment (Gironda, 1999). African American men were found to have meaningfully higher scores on Scales Fp, Clinical Scale 9, Content Scales FRS, BIZ, ASP, and Supplementary Scale MAC-r (MacAndrew Alcoholism Scale-revised) relative to Caucasian men. African American women had clinically meaningfully higher scores on Fp, Clinical Scale 5, FRS, CYN (Cynicism), ASP, and Supplementary Scale AAS (Addiction Potential Scale) compared to Caucasian women. In comparing the scale scores to external criterion variables, three out of 47 comparisons demonstrated test bias. Clinical Scale 8 and psychosis were more highly correlated in the African American population, while APS and collateral report of substance abuse and APS and chemical treatment were more highly correlated in the Caucasian sample.

In line with the push toward external correlate and predictive bias research, Monnot and colleagues (2009) examined such issues in male African American and Caucasian veterans seeking or engaged in substance abuse treatment. The external variable was diagnosis as measured by structured interviews. While differences were noted in 14 scales, meaningful mean T-score differences (T-score difference greater than five points) were only demonstrated for Clinical Scale 9 and RC9. However, results indicate a pattern of predictive bias concerning diagnoses across scales. Of the 46 comparisons that demonstrated intercept bias, all but one overpredicted diagnosis for African Americans either across the range of test scores or for higher test scores. The authors note that since these findings are clearly different from those reported by

Arbisi et al. (2002), evaluation of test bias should continue in various populations and settings.

MMPI-2 research. A recent unpublished thesis examined the predictive bias of the MMPI-2-RF's RC, H-O, SP, and PSY-5 scales in African American and Caucasian college students (McBride, 2013). Statistically significant mean T-score differences were found across ethnicity on several scales, including THD, RC3, RC6, MSF, DSF, SUB, MEC, and DISC-r. However, a step-down hierarchical multiple regression analysis only demonstrated predictive bias in 8 of the 39 analyses. Underprediction of criteria scores for African Americans was found for RC8 while overprediction of criteria scores for African Americans was found for RC4, RC7, RC9, and ACT. However, incremental changes in R2 for these scales produced less than small effect sizes and did not support any evidence of predictive bias in the examined scales.

Establishing Measurement Invariance

Measurement invariance, as applied in psychometrics, is a concept that an item (or any variable) relates to a latent variable (i.e., construct) in the same way across groups (Millsap, 2011). For example, measurement invariance is achieved if the items on a depression inventory measure the latent variable of depression in the same way in men and women. Measurement invariance can be assessed using Multiple Indicator Multiple Causes (MIMIC) modeling (Kim, Yoon, & Lee, 2012). MIMIC modeling is a special case of Structural Equation Modeling (SEM) where categorical covariates are added to a measurement model to examine their effect on the latent variable (Schumacker & Lomax, 2012). In the measurement model a confirmatory factor

analysis is undertaken where indicators are regressed upon one or more latent variables. The structural model additionally regresses the latent variable on one or more observed covariates to examine latent mean differences across groups (Kim, Yoon, & Lee, 2012; Schumacker & Lomax, 2012). Taken a step further, differential item functioning can be evaluated by regressing indicators on these categorical covariates. In an attempt to explain the conceptual underpinnings of MIMIC modeling, an outline of the underlying techniques and rationale is provided below. Since MIMIC modeling involves confirmatory factor analysis (CFA), a brief introduction to CFA is warranted.

Confirmatory factor analysis. CFA is similar to exploratory factor analysis (EFA) in that the goal is to find latent factors that are able to account for the variance and covariance of a set of observed indicators (Brown, 2006). In this way, CFA is a SEM procedure. CFA is also theory-driven, as all parts of the CFA must be pre-specified. CFA produce estimates of model parameters, including factor loadings, error variances, and factor variances (discussed below). Such model parameters are obtained using a fitting function (most often the Maximum Likelihood estimator) which attempts to reproduce the input variance/covariance matrix. This fitting function repeatedly refines the parameter estimates, called iteration, to get increasingly close to this goal. In other words, CFA delivers parameter estimates that are geared at maximizing the probability that the sample and predicted variance/covariance matrix are not statistically significantly different. Goodness-of-fit indices are then examined to evaluate the fit of the model based on whether the solution best represents the observed variances and covariances from the input data.

In CFA, parameters can be free, constrained, or fixed in terms of estimation (Brown, 2006; Muthén, & Muthén, 2009a). When parameter estimates are freed, the analysis attempts to find the values that best reproduce the variance/covariance matrix. Fixed parameters are set by the researcher to equal a certain value (Brown, 2006). For example, a model may propose that an indicator, such as an item of a psychological test that measures a latent variable, only loads on one of two factors in a two factor-hypothesized model. As such, the researcher can set the loading of that indicator to 0 on the second factor to specify the lack of a relationship. This scenario is common in CFA. Fixed parameters are also commonly used to provide relevant scaling of the latent variables. Finally, parameter estimates can be constrained rather than freed or fixed. A constrained parameter estimate is allowed to be any value within a restricted range. For example, a researcher may pre-specify that all factor loadings on a particular latent variable should be equal. In this way, the factor loadings are free to be any value but restricted in the sense that all loadings must be equal.

CFA model parameters. Parameter estimates in CFA, given in completely standardized, partially standardized, and unstandardized forms, typically include factor loadings, error variances, and factor variance (Brown, 2006; Muthén, & Muthén, 2009a). Error covariances, if desired, and factor covariances, if relevant, can also be specified in a model. It is important to note that while exploratory factor analysis (EFA) tends to use completely standardized variables, CFA analysis is usually completed with unstandardized observed and latent variables. The CFA solution can be produced in all three forms. A completely standardized solution fixes factor variances to 1.0 and factor loadings are correlations or standardized regression

coefficients. A partially standardized solution provides the relationship between unstandardized indicators and standardized latent variables. Finally, the parameter estimates are presented in the metric of the indicators in an unstandardized solution.

In an unstandardized solution, factor loadings (λ) are regression slopes of the factor on the indicator and can be interpreted as the expected change in the item for a one unit increase in the latent factor (Brown, 2006). Error variance (δ) is the variance in the indicator not explained by the latent factor and is most often presumed to be measurement error. Finally, factor variances (φ) are the sample variability on the latent factor. In standardized solutions, factor loadings are correlations when items are congeneric or partial regression coefficients when items are not congeneric. Indicators are said to be congeneric when they all load on the same factor. An indicator would not be congeneric if it loaded on more than one factor. Standardized error variances are correlations while standardized factor variances are fixed to 1.00.

A researcher can also specify error covariances, which demonstrate the amount that two indicators covary apart from their relationship to the latent factor (Brown, 2006). Most often, these values are fixed (assuming no or equal error covariance) but there may be expected reasons that two indicators covary apart from their relationship to the factor. For example, Byrne (2012) noted that a high degree of overlap in item content is a type of method effect that can result in residual covariances. Finally, if two or more latent factors are hypothesized, factor covariance may also be specified. Factor covariances estimate the relationship between two latent factors.

The aforementioned parameter estimates are based on the ability to reproduce the input variance-covariance matrix (Brown, 2006). At the foundation of this

analysis is that indicators (and latent variables), are assessed as deviations from their means, which are set to 0. However, this analysis can be adapted to include the analysis of mean structures, including indicator means and standard deviations. In including an analysis of mean structures, CFA parameter estimates attempt to reproduce not only the input variance/covariance matrix but also the observed sample means of indicators. Such an analysis allows for the investigation of the equivalence of indicator intercepts and latent factor means between groups. In line with the other parameters, the indicator intercepts can be constrained and the latent means fixed in CFA models. If indicator intercepts are constrained, latent mean values are meaningless. Thus fixing the mean of the latent factor in one group allows the mean of the latent factor in another group to be directly compared. For example, if group A's latent mean is set to 0 and group B's latent mean is 2.13, group B's average mean is 2.13 higher than group A's mean on the latent factor (construct).

Important to note, CFA can be used as a precursor to SEM in an attempt to outline structural relationships between latent variables (Brown, 2006). SEM models can be measurement models or structural models. Measurement models delineate the number of factors, factor loadings, and error covariances. Alternatively, structural models specify the relationship between latent factors, including latent factor variances, covariances, and means.

Goodness-of-fit indices. The goodness-of-fit indices provide information on how well a solution, based on a specified model, fits or reproduces the input data. The most highly recommended goodness-of-fit indices include χ^2 , the standardized root mean square residual (SRMR), the root mean square error of approximation

(RMSEA), the comparative fit index (CFI), and the Tucker-Lewis index (TLI; Brown 2006). χ^2 difference test is a hypothesis significance test based on the χ^2 distribution. In CFA, a statistically significant χ^2 rejects the null hypothesis that the resultant parameter estimates, and thus specified model, match the sample variance/covariance matrix. Therefore, the researcher is looking for a non-significant χ^2 difference test to conclude that the specified model is a good fit for the data. However, the χ^2 difference test should not be used as the only test of model fit based on its shortcomings. First, in the case of a small sample size of non-normally distributed data, the χ^2 distribution does not apply. Second, it is heavily affected by a large sample size such that larger samples increase the χ^2 value which can lead to an inappropriate rejection of the null hypothesis. Finally, since it is based on the strict equality of the sample and predicted variance/covariance matrices, χ^2 will lead to rejection of the null hypothesis even in cases where a reasonable fit exists.

Similar to χ^2 , SRMR assess the hypothesis that the sample variance/covariance matrix is equitable with the predicted variance/covariance matrix while not taking into account model fit relative to a more restricted model (Brown, 2006). Based on its name, SRMR is a positive value that is based on a square root average of the residual correlation. It is the mean difference between the input matrix correlations and the predicted model correlations. The SRMR can be between 0.0 and 1.0, with values less than or equal to .08 indicating good model fit (Muthén & Muthén, 2009a). However, there has been evidence that SRMR is not ideal for CFA with categorical indicators (Yu, 2002).

The root mean square error of approximation (RMSEA) is another goodness-of-fit index but varies from the aforementioned indices in that it rewards model parsimony (Brown, 2006). A more parsimonious model would have more degrees of freedom and thus less freely estimated parameters than another model. The RMSEA relies on the noncentral χ^2 distribution. This is the distribution of the fitting function (i.e., estimator) for a non-perfect model. As an error estimator, the RMSEA value demonstrates whether a model fits reasonably well in the population which is a less stringent hypothesis than other indices. It is also not as influenced by sample size as other indices. A perfect model fit would be represented by a RMSEA value of 0.0 and although the upper limit of the value is limitless, upper limits usually do not exceed 1.0. A good model fit would be represented by RMSEA values less than or equal to 0.06 (Brown, 2006; Muthén & Muthén, 2009a).

The last two recommended goodness-of-fit indices, the comparative fit index (CFI) and the Tucker-Lewis index (TLI) assess the fit of a hypothesized model against a nested, more specified model (Brown, 2006). This nested, more specified model usually has the indicator covariances fixed to zero, thus indicating no relationship between indicators. Essentially, the CFI and TLI are comparing the fit of a given model to a very restricted model and thus are more likely to provide values indicating good model fit when compared to the aforementioned fit indices. The CFI also uses the noncentral χ^2 distribution for a non-perfect fitting model. The TLI also favors parsimonious models and compares a given model against a more restrictive model. While the CFI can range from 0.0 to 1.0, the TLI is non-normed and thus can produce

values that are larger or smaller. However, for both indices, values at or higher than 0.95 indicating a good model fit (Brown, 2006; Muthén & Muthén, 2009a).

Of important note, all of the above recommended values indicating good model fit have been researched on continuous indicators using the Maximum Likelihood (ML) estimator (Brown, 2006). Since this estimator is inappropriate for use with categorical variables, less stringent cut-off values have been used with categorical indicators (Ketterer, 2011). It is also crucial to note that goodness-of-fit indices should only be one portion of evaluating the fit of a model. A researcher must also consider a particular solution with regards to areas of localized strain (areas of the specified model that are not appropriately reproduced) and interpretability and strength (Brown, 2006). With regards to the latter, special attention should be paid to any Heywood cases (out-of-range parameter estimates) and whether the direction and size of the results correctly portrays the pre-specified model. Further, interpretability of the factors should be considered.

Modification indices. Modification indices allow for further evaluation of the model based on particular relationships in the solution (Brown, 2006). Modification indices can be calculated for each fixed and constrained parameter in the model, indicating the approximate amount the model χ^2 would decrease if the parameter were freed. The modification indices in a good-fitting model should be under 4.00 (Brown, 2006; Jaccard & Wan, 1996).

Similar to model χ^2 and standardized residuals, modification indices are influenced by large sample sizes (Brown, 2006). In such a case the large modification index may point to the need to freely estimate a model parameter when in actuality the

freely estimated parameter, when applied, is not meaningful. To remedy this, expected parameter change (EPC) values are provided for each modification index in some statistical programs. EPC values indicate the amount the particular parameter is expected to increase or decrease if freely estimated. EPC values can be unstandardized, standardized, or completely standardized (Mplus provides all three). Unstandardized EPC values are on the scale of the observed measures and thus completely standardized EPC values are more meaningful and more frequently used. EPC values, the size and direction, should be used in combination with modification indices when employing a large sample.

Brown (2006) notes that while modification indices and EPC values may prompt freeing parameters, researchers need to be careful only to do based on sound reasoning (i.e., research or theoretical bases). Research has noted the downfalls and misspecifications that can arise from revising a model solely based on modification indices and trivial EPCs (MacCallum, 1986; Silvia & MacCallum, 1988). It is also important to note that multiple high modification indices may be decreased by freeing only one of the parameters (Brown, 2006). Thus, only one parameter should be freed at a time in subsequent analysis. Researchers should start by freeing the parameter with the largest modification index and EPC first, if justified by theory or research and the parameter can be interpreted (Jöreskog, 1993). If there is not a compelling reason to free the parameter with the largest modification index and EPC, researchers should move to the parameter with the second largest modification index, etc.

CFA with categorical variables. The above outlined information on CFA is based on linear CFA, which is meant for continuous variables (Kim, & Yoon, 2011;

Millsap, 2011). CFA, or any CFA- based approach, with categorical or dichotomous variables, involves a change in the input matrix, variables, and interpretation. First, rather than the sample variance/covariance matrix being as input (as is done with linear CFA), the analysis is conducted on a correlation matrix (Brown, 2006). In the case of dichotomous indicators, as in this study, a tetrachoric correlation matrix serves as the input data.

Based on an approach described by Muthén and Asparouhov (2002) and used in Mplus (Muthén & Muthén, 2011; the statistical programming used in this study), CFA with categorical data can be conducted using latent continuous response variables, y^* (Muthén & Muthén, 2009b). In this approach, y^* is the amount of a latent and continuous construct (e.g., personality, intelligence, psychopathology, etc.) needed to endorse (in the case of a dichotomous indicator) a particular observed indicator (Brown, 2006). For example, if evaluating a psychopathy test using this approach, y^* would represent the particular amount of psychopathic behavior needed to endorse an item indicating the presence of psychopathic behavior. Thus, this approach assumes that constructs could be measured on a more appropriate and specific scale rather than by simple yes-no or true-false responses (Brown 2006). Tetrachoric correlations between y^* variables are then used as the sample input data.

The initial dichotomous variables are associated with the y^* variables through threshold parameters, that is, the point on the y^* variable wherein the threshold is exceeded and the indicator (i.e., item) is endorsed (Brown, 2006; Muthén & Asparouhov, 2002; Muthén & Muthén, 2009b). Thresholds essentially cut the underlying y^* variables into ordered categories (Finney & DiStefano, 2013). Thresholds are the

point where the test taker's response "moves" from one category to another. Put another way, a threshold parameter for a test with dichotomous responses (i.e., true-false, yes-no) is the point on the y^* variable where a test taker chooses the affirmative response (i.e., yes, true) indicating the presence of the underlying construct. Items with three levels of responses (e.g., 0, 1, 2) would have two threshold parameters, one for each level of the construct that can be endorsed (e.g., 0 to 1 and 1 to 2; Brown, 2006).

Since the metric of the y^* variables is arbitrary, the mean is set to zero and the standard deviation is set to one (Finney & DiStefano, 2013). A threshold is thus a z-score that corresponds to the cumulative area under the curve to the left of the category. For example, imagine a 0 or 1- (incorrect-correct) scored test of intelligence that asks test takers to solve math problems mentally. For examples sake, this test was administered to a group of college students. One particular item was found to have a threshold parameter of 1.46. This threshold parameter indicates that a correct response is triggered when the college students are 1.46 standard deviations above the mean on the underlying y^* variable for the item.

Thus, the 0-1 scored item is the observed variable assessing incorrect-correct on this item. The y^* variable turns this dichotomous observed variable into a latent and continuous construct representing the 0-1, incorrect-correct, mathematical ability on this item. Threshold parameters are the point on the y^* variable where the response changes from incorrect to correct and can be interpreted as z-scores.

The thresholds are used to compute latent correlations between the y^* variables, tetrachoric correlations in the case of the current analysis, and used as input for estimating the model parameters (Finney & DiStefano, 2013).

MIMIC modeling. MIMIC modeling, also referred to as CFA with covariates, is one of two forms of multiple group CFA (Brown, 2006). Both forms of multiple group CFA serve a means of assessing measurement invariance (Brown, 2006; Kim, Yoon, & Lee, 2012). In the MIMIC modeling approach, measurement invariance is tested by regressing the latent factor(s) and indicators onto dummy-coded covariates that denote group membership (Brown, 2006). MIMIC modeling begins with finding a valid CFA measurement model on the full sample, merging groups. The second step involves adding the dummy-coded covariates representing group membership to the model in order to assess their direct effects on the latent factor and any chosen indicators. A single input matrix is used that contains variances and covariances (or tetrachoric correlations, for dichotomous variables) of the latent factor and observed covariates. Of note, the latent factor in a MIMIC model is endogenous rather than exogenous, meaning that it is a dependent variable and caused by one or more other variables in the model (in this case the dummy-coded covariate). Some statistical programming requires latent-Y specification in such cases.

A significant direct effect of the observed covariate on the latent factor points to group differences on latent means which is commonly referred to as population heterogeneity (Brown, 2006). This result demonstrates that the latent factor means vary at different levels of the covariate (i.e., varies based on group membership) and indicates population heterogeneity. By the same token, a significant direct effect of

the observed covariate on an indicator signifies group differences on the indicator's intercept (or threshold parameter, for categorical variables), or measurement noninvariance. Put another way, this direct effect means that when the latent factor is held constant, the mean of the chosen indicator (or the threshold parameter, probability of endorsing the item) varies at different levels of the covariate (i.e., varies based on group membership), pointing to differential item functioning.

Differential item functioning (DIF) points to different measurement properties of an item based on group membership, holding any group mean differences constant (Woods, Oltmanns, and Turkheimer, 2011). An item demonstrating DIF is noninvariant because part of whether it is endorsed is based on group membership, not levels of underlying traits. There are two types of DIF, uniform and non-uniform (Walker, 2011). Uniform DIF occurs when items functioning differently in a uniform fashion at all levels of the latent trait. DIF is said to be non-uniform when items only function differently at certain levels of the latent trait (e.g., at extreme scores).

MIMIC models can be tested with or without a hypothesis regarding invariance (Brown, 2006). In an exploratory approach to MIMIC modeling, all direct effects between the covariate and indicators are set to zero. Modification Indices are then examined for significant direct effects. Freely estimating these direct effects in exploratory MIMIC modeling would result in an underidentified model.

MIMIC models test the invariance of factor means and indicator intercepts (or threshold parameters; Brown, 2006). However, factor means and indicator intercepts are not estimated in the analysis. Indicator means are also not included in the input matrix. Instead, group mean differences in factor means and indicator intercepts are

provided by parameter estimates of direct effects where factor and indicator means are zero. In unstandardized terms, the direct effect of the covariate on the latent factor can be interpreted as the difference in latent means between the groups. Since MIMIC models only test the invariance of factor means and indicator intercepts/thresholds, it assumes that all other measurement and structural parameters are equal across levels of the covariates.

Advantages and disadvantages to MIMIC modeling. To discuss the advantages and disadvantages of MIMIC models, it is important to briefly discuss the other form of multiple group CFA used for assessing measurement invariance, Multiple-Group Confirmatory Factor Analysis (MGCFA; Brown, 2006). Assessing for measurement invariance using MGCFA involves specifying increasingly restrictive CFA models in different groups and examining the model fit indices to determine whether the more restrictive model is a worse fit than the less restrictive model. Separate input matrices are used for each group. The researcher is testing for different forms and causes of noninvariance as these CFA models becoming increasingly restrictive. Indeed, the ability to test all aspects of measurement invariance and population heterogeneity is an advantage of MGCFA when compared to MIMIC models.

By the same token, MIMIC models have three main advantages over MGCFA (Brown, 2006). First, MIMIC models have smaller sample size demands. Given that MGCFA analyzes multiple measurement models (depending on the number of groups), it is no surprise that large samples are needed to allow for adequate power in each separate CFA. On the other hand, MIMIC models only require one CFA and do

not need as large of a sample (overall and/or in each group). Second, MIMIC models are more parsimonious when dealing with more than two groups. Conducting separate CFA in three or more groups can become complex based on specifying model parameters across groups. MIMIC models allow for multiple dummy-coded covariates.

CHAPTER III

RATIONALE OF THE PRESENT STUDY

The goal of the current study is to examine the measurement invariance of the MMPI-2-RF Internalizing Specific Problem Scales in African American and Caucasian men. African Americans are included in the MMPI-2-RF normative sample and research with the MMPI-2 has only examined prediction invariance. For the most part, these studies have shown no bias for many scales, and some scales have shown small to moderate predictive bias in some studies. It is important to note that these small/moderate differences would not likely affect clinical interpretation significantly, if at all. No published studies have examined measurement invariance for the any of the MMPI-2-RF scales, including the SP Scales.

In a thorough introduction to the theory, application, and use of measurement invariance, Millsap (2011) outlines the continued need to assess for measurement bias in psychological tests given the long history of such research. First, early research on test bias employed then current and upcoming research techniques later shown to have fundamental flaws. While some of these methods have been improved, more appropriate approaches are less often used due to computing and/or software demands and a general lack of awareness of such techniques.

The history of test bias with the MMPI/MMPI-2/MMPI-2-RF began with the evaluation of mean T-score differences and evolved into assessing for predictive bias via correlation and regression. While the problem with solely examining differences in mean scores with a goal of elucidating test bias has been reviewed (i.e., different mean scores may reflect underlying group differences rather than bias), problems also arise

in attempting to assess for test bias using regression and correlation (Millsap, 2011). Historically, researchers have claimed that a lack of differences when comparing groups in correlations or regressions between a test and an external criterion indicated the lack of meaningful test bias. However, research has shown that a test may produce identical regressions across groups, but still be a biased measure (Borsboom, Romeijn, & Wicherts, 2008; Millsap, 2007). Therefore, as noted by Millsap (2007), it is important to evaluate both prediction invariance via multiple regression and measurement invariance using methods such that confirmatory factor analysis.

The examination of measurement invariance for the MMPI-2-RF scales provides a much needed advance in the statistical analysis of possible measurement bias. This author only knows of two previous analyses examining measurement invariance in the MMPI/MMPI-2. The first of these, described above, was provided as an example of measurement invariance and used forensic adolescent data from African Americans and Caucasians collected in the 1960s and analyzed scales that are not in use (Millsap, 2011). The second, most recent, and most comprehensive of these examines the measurement invariance of the English language and Korean MMPI-2 RC Scales in Korean and American normative samples (Ketterer, 2011). However, Ketterer's (2011) analysis is an unpublished doctoral dissertation. Thus, the current study is the first to examine measurement invariance in MMPI-2-RF specific scales and measurement invariance in any MMPI-2-RF scales in an adult African American and Caucasian sample.

In the present study, the measurement invariance of the MMPI-2-RF Internalizing SP scales is examined. The SP Scales were chosen because they are

more likely to be unidimensional than other MMPI-2-RF Scales given their narrow focus. Given that most studies of measurement invariance typically focus on a single scale (e.g., Culhane et al., 2009), examining an entire set of MMPI-2-RF scales is an ambitious undertaking. Moreover, this study examines the measurement invariance of the MMPI-2-RF Internalizing SP Scales in an amalgamated sample of African American and Caucasian men and attempt to replicate the results in an inpatient sample of African American and Caucasian men. This present analysis is meant to build upon previous test bias research, but also advance this research, by providing the first assessment of measurement invariance in the MMPI-2-RF specific to African American and Caucasian men. However, since this study is an initial step in furthering this research, future studies should investigate measurement invariance in various scales in multiple difference populations, including setting-specific samples.

This study investigates the measurement invariance of the MMPI-2-RF SP scales using MIMIC modeling. Initially, the study aimed to assess measurement invariance in these scales using Multiple-Group Confirmatory Factor Analysis (MGCFA; discussed previously in the literature review). However, since MGCFA involves running separate CFAs for each group, the sample sizes for African Americans in both the amalgamated Pearson and psychiatric inpatient data were too small for adequate power. Thus, MIMIC modeling and DIF was chosen as an alternative means of assessing measurement invariance based on the ability to use the entire sample in the CFA with group as a covariate, thus meeting sample size requirements. As previously noted, MIMIC modeling has some limitations in comparison to MGCFA.

Because no literature on the topic of measurement invariance in the MMPI-2-RF Scales is available, there is no way to generate hypotheses on the likely nature and extent of any measurement noninvariance. Given that this is the first study, any findings suggestive of measurement noninvariance will need to be replicated in additional studies.

CHAPTER IV

METHODOLOGY

Participants

Pearson sample. The Pearson sample used in this study was requested from Pearson Assessments' archival data (NCS Pearson, 2008-2014). The data were requested as MMPI-2-RF protocols, starting with the most current protocols working backward in date for a satisfactory number of protocols. Protocols were requested for clinical outpatient test takers; however, Pearson reported that they do not have data on setting for the MMPI-2-RF protocols. MMPI-2 protocols were under consideration for use, as data were collected on setting for those protocols but Pearson did not have ethnicity data for MMPI-2 protocols. As such, the Pearson sample is an amalgamated sample of protocols from African American and Caucasian test takers. The provided data from Pearson included age, gender, ethnicity, and raw MMPI-2-RF data (338 items). The initial data consisted of 3,407 protocols from 309 African American men and 3,098 Caucasian men.

Invalid protocols were removed based on validity criteria of Cannot Say (CNS-r) ≥ 15 , Variable Response Inconsistency (VRIN-r) and True Response Inconsistency (TRIN-r) ≥ 80 , Infrequency Responses (F-r) = 120, Infrequent Psychopathology Responses (Fp-r) > 99 , and Uncommon Virtues (L-r) ≥ 80 . The number of valid and invalid protocols by scale and in total can be found in Table 1. For comparison, in the Pearson sample 27.18 percent of protocols from African American and 11.07 percent of protocols from Caucasian test takers were removed based on the validity criteria. In the inpatient sample, 55.92 percent of protocols from African American and 29.98

percent of protocols from Caucasian test taskers were removed based on the validity criteria. After removal of invalid protocols, a total sample of 2,980 valid protocols remained, 225 from African American men and 2,755 Caucasian men. This sample size is adequate, as research has demonstrated that a large sample ($n > 400$) is needed for adequate power in CFA (Meade & Bauer, 2007) and the current analysis is done collapsing across group. The final sample of African American and Caucasian men had a mean age of 37.99 years with a standard deviation of 20.89 years. The median age of the entire sample was 37 years old.

Table 1

Number of Valid (and Invalid) Protocols by Validity Scale for Each Sample

Scale	Pearson		Inpatient	
	African-American	Caucasian	African-American	Caucasian
None Removed	309 (0)	3098 (0)	304 (0)	1778 (0)
CNS-r	306 (3)	3083 (15)	298 (6)	1757 (21)
VRIN-r	299 (10)	3064 (34)	267 (37)	1722 (56)
TRIN-r	396 (13)	3053 (45)	261 (43)	1643 (135)
F-r	284 (25)	3013 (83)	195 (109)	1461 (317)
Fp-r	292 (17)	3009 (89)	205 (99)	1591 (187)
L-r	268 (41)	2943 (155)	295 (9)	1720 (58)
All Validly Scales	225 (84)	2755 (343)	134 (170)	1245 (533)

Table 1 Continued

Note. CNS-r refers to the Cannot Say Scale, VRIN-r refers to the Variable Response Inconsistency Scale, TRIN-r refers to the True Response Inconsistency Scale, F-r refers to the Infrequent Responses Scale, Fp-r refers to the Infrequent Psychopathology Responses Scale, and L-r refers to the Uncommon Virtues Scale.

Psychiatric inpatient sample. The inpatient data were archival and obtained from Kent State University with Paul Arbisi's permission. The provided data contained protocols of inpatient populations from the Minneapolis VAMC (61.40 percent of the sample) and the Hennepin County Medical Center (HCMC; 38.60 percent of the sample). A subset of the same data set was used in previous test bias research (Arbisi et al., 2002) and subsets of the data were also used in the validation of the RC scales. Additionally, this sample was used as a validation sample for the MMPI-2-RF (Tellegen & Ben-Porath, 2008/2011). The data provided contained information on age, ethnicity, war veteran status, branch of the military, hospitalization length, and raw MMPI-2 data (567 items). The initial data consisted of 2,082 protocols from 304 African American men and 1,778 Caucasian men.

Again, invalid protocols were removed based on validity criteria of Cannot Say (CNS-r) ≥ 15 , Variable Response Inconsistency (VRIN-r) and True Response Inconsistency (TRIN-r) ≥ 80 , Infrequency Responses (F-r) = 120, Infrequent Psychopathology Responses (Fp-r) > 99 , and Uncommon Virtues (L-r) ≥ 80 . The number of valid and invalid protocols by scale and in total can be found in Table 1. After removal of invalid protocols, 134 valid protocols from African American men

and 1,245 valid protocols from Caucasian men remained for a total sample of 1,379 combined valid protocols. Again, this sample size is ample for adequate power in CFA.

The final sample of inpatient African American and Caucasian men had a mean age of 42.91 years and a standard deviation of 14.50 years. The median age of the combined sample was 42.00 years old. African American and Caucasian men in the inpatient sample had an average hospitalization stay of 20.99 days and median hospitalization stay of 15 days. The majority of the veterans from the VAMC sample were Vietnam veterans (27.80 percent of the valid combined sample), followed by post-Vietnam veterans, World War II veterans, veteran status unknown, Korean veterans, Persian Gulf veterans, Post-Korean veterans, and World War I veterans. Of the veterans that reported their previous military affiliation, most of the veterans reported serving in the Army, followed by the Navy, Marines, and Air Force. The demographics of the inpatient sample are presented in Table 2.

Table 2

Demographic Information of the Inpatient Sample from the Minneapolis Veterans Affairs Medical Center (VAMC) and Hennepin County Medical Center (HCMC)

Demographic	N	Mean (SD) or Percentage
Age	1379	42.91 (14.50)
Site	1379	
VAMC		61.40%

Table 2 Continued

Demographic	N	Mean (SD) or Percentage
HCMC		38.60%
Length of Hospitalization	1378	20.99 (20.77)
World War I Veteran	1095	0.90%
World War II Veteran	1102	9.90%
Korean Veteran	1103	6.70%
Vietnam Veteran	1107	27.80%
Post-Korean Veteran	962	4.10%
Post-Vietnam Veteran	1105	10.90%
Persian Gulf Veteran	1103	4.40%
Veteran Status Unknown	1098	8.50%
Branch of the Military	867	
Army		33.30%
Navy		15.30%
Marines		7.80%
Air Force		6.50%
Unknown		37.10%

Instruments

MMPI-2-RF. Since the MMPI-2-RF is described in detail above, the current section will provide a brief overview and more thoroughly discuss the measure's psychometric properties. The MMPI-2-RF is a 338 item true-false measure of personality and psychopathology (Ben-Porath & Tellegen, 2008/2011). It is intended to be a broad assessment instrument for use in a variety of settings. The instrument consists of 50 scales, described above, that measure a range of psychopathology and personality dimensions. The MMPI-2-RF can be hand scored, computer scored on-site using a software system, or mailed to Pearson for scoring. The resulting Score Report delivers raw and standard T-scores for each scale. Item level information, specifically critical items and unscorable responses, is also provided in the Score Report. The test administrator can also request for the relevant group data to be plotted along with a specific test taker's scores. The Interpretive Report provides an interpretation of the scores in addition to information available in the Score Report. The interpretative statement, which can also be provided along with the scale that produced the statement, is based on external correlates as well as item content.

Psychometric properties of the MMPI-2-RF. The psychometric properties of the MMPI-2-RF scales were investigated in several archival data sets, including men and women from the MMPI-2 normative group, a community mental health outpatient center, an inpatient psychiatric hospital, and male inpatients at a VAMC (Tellegen & Ben-Porath, 2008/2011). Apart from VRIN-r and TRIN-r, the Validity Scales (i.e., F-r, Fp-r, Fs, FBS-r, RBS, L-r, and K-r) produced Cronbach's alphas ranging from .39 to .69 in the normative sample, .53 to .85 in the community mental health sample, .47 to

.87 in the inpatient sample, and .54 to .87 in the VAMC sample. In all samples, VRIN-r and TRIN-r produced alphas ranging from .16 to .41, which the test developers point out that it is not surprising since their item content was not designed to assess a particular content area but rather random and fixed response patterns.

Cronbach's alphas for the three H-O Scales ranged from .69 to .88 in the normative sample, .79 to .94 in the community mental health sample, .81 to .95 in the inpatient sample, and .84 to .93 in the VAMC sample (Tellegen & Ben-Porath, 2008/2011). For the Interpersonal Scales, alpha coefficients ranged from .43 to .78 in the normative sample, .57 to .85 in the community mental health sample, .61 to .86 in the inpatient sample, and .61 to .85 in the VAMC sample. The Interest Scales alphas ranged from .49 to .67 across the four samples. Finally, the PSY-5 Scales achieved alphas ranging from .69 to .78 in the normative sample, .70 to .85 in the community mental health sample, .73 to .88 in the inpatient sample, and .75 to .86 in the VAMC sample. In the normative sample, test-retest reliabilities for the Validity Scales ranged from .52 (TRIN-r) to .84 (K-r). Test-retest reliabilities ranged from .71 to .91 for the H-O Scales, .60 to .88 for the Interpersonal Scales, and .76 to .93 for the PSY-5 Scales. The Interest Scales produced test-retest reliabilities of .86 to .92 for AES and MEC, respectively.

To assess the validity and comparability of VRIN-r and TRIN-r, researchers examined whether protocols could be identified in which varying amounts of the original responses were replaced with either random or fixed responses (Handel, Ben-Porath, Tellegen, & Archer, 2007). Results indicated that both scales were able to detect such responding. In comparing VRIN-r and TRIN-r to their MMPI-2

counterparts, the revised scales on the MMPI-2-RF appeared to perform as well or better than their predecessors. Generally, intercorrelations between the MMPI-2-RF and MMPI-2 over-reporting Validity Scales (i.e., F-r, Fp-r, Fs, and FBS-r) are high in simulated samples (Tellegen & Ben-Porath, 2008/2011). In fact, the correlation between FBS-r and FBS for personal injury test takers and test takers instructed to simulate head injury was .96. The two under-reporting scales, L-r and K-r have been demonstrated to appropriately detect underreporting in simulated samples and samples where underreporting may be expected (e.g., legal cases) and are highly correlated with their MMPI-2 counterparts.

Intercorrelations between the 42 major scales of the MMPI-2-RF were correlated with the 103 main MMPI-2 scales (Tellegen & Ben-Porath, 2008/2011). The majority of the main MMPI-2 scales were demonstrated to correlate with at least one major MMPI-2-RF scale in expected relationships. However, since the MMPI-2-RF is not meant to be an exact continuation of the MMPI-2 and major changes in scales occurred, such correlates are not expected to be extremely high. In examining the three H-O scales, which meant to be overarching domains, expected correlations emerged. For example, the THD Scale correlated .74 with RC6, .87 with RC8, and .95 with PSYC-r in the normative sample. As in the MMPI-2, of all the Validity Scales, F-r is the most highly related to the major MMPI-2-RF scales. Since test takers with a high level of psychopathology also tend to elevate F-r, this correlation is not surprising. However, the test developers point out that the correlations between F-r and the main scales are generally lower than those found between F and the main MMPI-2 scales. Also, convergent validity was evidenced in scales that conceptually

should not be related producing small correlations. For example, the AGG Externalizing Specific Problem Scale correlated .00 with RC2, .14 with GIC, .12 with SHY, and .07 with MSF.

Psychometric properties of the SP scales. In assessing the reliability of SP scales, Cronbach's alphas and test-retest correlations were examined (Tellegen & Ben-Porath, 2008/2011). Cronbach's alphas for the Somatic/Cognitive SP Scales ranged from .52 to .69 in the normative sample, .74 to .83 in the community mental health sample, .71 to .84 in the inpatient sample, and .74 to .82 in the VAMC sample. The Internalizing SP Scales produced alpha coefficients ranging from .34 to .73 in the normative sample, .48 to .82 in the community mental health sample, .61 to .84 in the inpatient sample, and .57 to .80 in the VAMC sample. Finally, the Externalizing SP Scales achieved alphas ranging from .56 to .66 in the normative sample, .59 to .75 in the community mental health sample, .71 to .77 in the inpatient sample, and .71 to .75 in the VAMC sample. In the normative sample, the Somatic/Cognitive, Internalizing, and Externalizing Scales demonstrated test-retest reliabilities ranging from .54 to .82, .65 to .85, and .77 to .87, respectively.

Intercorrelations for the SP scales demonstrated correlations in expected directions (Tellegen & Ben-Porath, 2008/2011). For example, the Somatic/Cognitive Scales correlate highly with FBS and RC1 and only slightly with RC9 and JCP; the Internalizing scales demonstrate a strong relationship to F-r and EID but are only slightly related to RC4 and AGG; and the Externalizing scales were negatively correlated with K-r, SAV, and IPP but were related to BXD and RC4. In comparing the SP Scales to the MMPI-2 Clinical Scales, more expected correlations emerge. For

example, HPC, MLS, and NUC are most highly correlated with Clinical Scale 1 and 3, EID with Clinical Scale 2 and 7, and ACT with Clinical Scale 9. Some correlations were observed between the SP and Clinical Scales that were not expected.

Statistical Analyses

Data preparation. To prepare for the analyses, the MMPI-2 protocols for the VAMC/HCMC sample were transformed to MMPI-2-RF protocols. All protocols were scored to examine validity criteria and invalid protocols were removed. Separate data files were then created for each scale containing only the items on the respective scale. Protocols with any missing responses on items were removed. Some scales are keyed all true while other scales are keyed a mixture of true and false (e.g., responding “false” is endorsing the symptom/item). All data were recoded so that a keyed response was coded a one and an unkeyed response coded a zero. Finally, Mplus 7.2 input files were created for each Internalizing SP Scale by creating text (.txt) files from the SPSS files.

Model specification and analysis. One factor solutions were the baseline models for each of the Internalizing SP Scales, for a total of nine separate measurement models in both the Pearson (amalgamated) and inpatient (VAMC/HCMC) data. Single factor solutions were chosen based on previous research in the development of the MMPI-2-RF (Tellegen & Ben-Porath, 2008/2011). A more complex nine factor solution was considered for use in each sample as a means to investigate the nine Internalizing Specific Problem Scales as interrelated factors. However, ultimately individual one factor solutions were decided upon since

the Specific Problem Scales are all interpreted individually and are not dependent upon each other or any other MMPI-2-RF scales for elevation.

As a first step in the analysis, one factor models were analyzed and examined for model fit for each of the nine Internalizing SP Scales using Mplus 7.2 (Muthén and Muthén, 1998-2012). The latent variable was scaled using the marker indicator approach, which involves fixing the metric of the latent factor to be the same as one of the indicators. The marker indicator approach is the default in Mplus.

The default parameterization for CFA with categorical indicators, used in this study, is delta parameterization. In this approach, y^* is scaled by fixing variances to 1.0 for all of the indicators (Brown, 2015). Therefore, unlike CFA with continuous variables, the residual variances of categorical indicators are not identified and thus not a part of the model. Measurement errors of the CFA with categorical indicators are also not free parameters. Delta and theta are similar parameterizations of a CFA and produce identical goodness-of-fit indices and nested model results. Theta parameterization is used less frequently and includes the indicator error variances as part of the CFA model but fixes the error variances to all have the sample value.

The model was estimated using the Mplus default for categorical indicators, the weighted least-square mean variance (WLSMV) estimator. The WLSMV estimator affords weighted least square estimates via robust standard errors, a diagonal weight matrix, and mean- and variance- adjusted χ^2 (Brown, 2006).

As a second step in the analysis, the baseline one factor CFA model was analyzed in terms of modification indices and item content related to the potential need to allow correlated error terms of the indicators. After examining each

Internalizing SP scale's modification indices in each sample and reviewing the item content, some indicator error terms were allowed to correlate if such correlations made statistical and substantive sense. Before allowing indicator error terms to correlate, modification indices pointing to improved model fit with indicator correlated error terms was necessary in both the amalgamated and inpatient sample. Next, the item content was examined for similarity of item wording or overall meaning. For example, indicator error terms were allowed to correlate for MMPI-2-RF items 93 and 164 in both samples for the SUI scale. The items are copyrighted by the University of Minnesota Press and cannot be reproduced. Instead, a list of the indicator error terms that were allowed to correlate and a general description of the respective items can be found in Table 3. The University of Minnesota Press approved the broad item descriptions provided in this dissertation.

Table 3

Correlated Indicator Error Terms for MMPI-2-RF items in the CFA Models for Both the Outpatient and Inpatient Sample by Scale

Scale	MMPI-2-RF Items	Description
SUI	93 with 164	both active suicidal ideation
HLP	none	
SFD	89 with 232	both occasional self-doubt
NFC	152 with 198	both specific to difficulties
STW	73 with 167	both specific to nervousness

Table 3 Continued

Scale	MMPI-2-RF Items	Description
AXY	79 with 289	both related to nighttime
ANP	134 with 293	both related to quick temper
BRF	none	
MSF	54 with 151	both related to storms

Note. MMPI-2-RF refers to the Minnesota Multiphasic Personality Inventory-2-Restructured Form, CFA refers to Confirmatory Factor Analysis, SUI refers to Suicide/Death Ideation, HLP refers to Helplessness/Hopelessness, SFD refers to Self-Doubt, NFC refers to Inefficacy, STW refers to Stress/Worry, AXY refers to Anxiety, ANP refers to Anger Proneness, BRF refers to Behavior Restricting Fears, and MSF refers to Multiple Specific Fears.

The third step in the analysis involved analyzing goodness-of-fit indices in the baseline CFA model, some scales with correlated error terms. When a satisfactory model fit was found for the scale, the fourth step in the analysis was to add the dummy-coded covariate of ethnicity to the model. Fifth in the analysis, model fit and direct effects of the covariate on the latent variable were examined. In the sixth step, to test for differential item functioning (DIF), paths were added from the covariate to each of the indicators constrained to zero (assuming no direct effects). When modification indices pointed to the need to freely estimate a specific path from the

covariate to an item, the path with the highest modification index was freed first and the model was re-estimated.

In the seventh and final step, modification indices were examined and paths between the covariate and relevant items were freed until no significant modification indices remained (over 4.0). Goodness-of-fit indices were examined as the last step when no significant modification indices remained. The aforementioned process for examining measurement invariance using MIMIC modeling is described in a short course video and handout on the Mplus website (Muthén & Muthén, 2009b).

Goodness-of-fit indices. In evaluating model fit for each scale in both samples, RMSEA, CFI, and TLI were consulted. As previously mentioned, these indices demonstrate how well a solution fits or reproduces the input data. While these fit indices were described in more detail earlier under the literature review, RMSEA values less than 0.06 and CFI and TLI values of more than 0.95 indicate a good model fit (Brown, 2006). Of note, such cut-off values have been found in research using maximum likelihood (ML) estimation and research with WLSMV estimation is more limited (Ketterer, 2011). As such, less stringent cut off values with regard to these goodness-of-fit indices may need to be employed due to the use of categorical indicators (and estimation method). χ^2 values and significance was also noted but not relied upon as heavily as the other fit indices due to the previously mentioned shortcomings of the test. Model fit was also examined in terms of factor loadings, modification indices, and the presence of any out-of-range (Heyward) cases.

CHAPTER V

RESULTS

Descriptive Statistics

Descriptive statistics for the Pearson and inpatient sample, divided by ethnicity, can be found in Tables 4 and 5, respectively. For the Pearson sample (Table 4), mean and median scores were similar across scales for African American and Caucasian men, with the exception of BRF and MSF. Cohen's d values point to an almost medium effect size for the difference between MSF mean scores in African American and Caucasian men and a small effect size for the difference between BRF for African American and Caucasian men. All other effect sizes for scale mean differences between African American and Caucasian men were less than small (under .20). For the inpatient sample (Table 5), all mean and median raw scores were fairly consistent across African Americans and Caucasians in the sample. The largest differences in mean scores across the inpatient sample can be seen in the SUI and MSF scales, both demonstrating small effect sizes.

Table 4

Descriptive Statistics and Reliability Coefficients for the Internalizing Specific Problem Scales for the Pearson Sample by Ethnicity

Scale	African- Americans				Caucasians				
	M	Mdn	SD	α	M	Mdn	SD	α	d
SUI	49.30	45.35	11.62	0.75	50.58	45.35	12.76	0.70	0.10
HLP	48.95	40.48	10.96	0.59	49.18	40.48	12.26	0.71	0.02
SFD	48.78	41.83	9.95	0.76	50.82	41.83	12.10	0.85	0.18
NFC	50.69	47.65	11.31	0.80	48.46	47.65	11.75	0.82	-0.19
STW	50.51	47.39	11.37	0.68	50.88	47.39	12.13	0.73	0.03
AXY	53.14	44.02	13.87	0.67	51.63	44.02	13.36	0.73	-0.11
ANP	49.22	46.80	11.48	0.81	47.64	46.80	11.25	0.83	-0.13
BRF	49.95	42.74	10.29	0.52	47.49	42.74	8.92	0.55	-0.24
MSF	47.63	45.62	8.38	0.68	44.23	45.62	6.90	0.63	-0.44

Note. SUI refers to Suicide/Death Ideation, HLP refers to Helplessness/Hopelessness, SFD refers to Self-Doubt, NFC refers to Inefficacy, STW refers to Stress/Worry, AXY refers to Anxiety, ANP refers to Anger Proneness, BRF refers to Behavior Restricting Fears, and MSF refers to Multiple Specific Fears. M refers to mean, Mdn refers to Median, SD refers to standard deviation, α refers to alpha coefficient, and d refers to Cohen's d. Cohen's d was calculated by subtracting the mean scores of African Americans from the mean scores of Caucasians. N for African Americans = 225 for mean, median and standard deviation, n for alpha coefficients varied by scale between 222 – 225; N for Caucasians = 2,755 for mean, median and standard deviation, n for

Table 4 Continued

alpha coefficients varied by scale from 2,736 – 2,755. Unrounded, untruncated T-scores were used to obtain the mean, median, and standard deviation descriptive statistics and raw data were used for the reliability analyses.

Table 5

Descriptive Statistics and Reliability Coefficients for the Internalizing Specific Problem Scales for the Inpatient Sample by Ethnicity

Scale	African- Americans				Caucasians				
	M	Mdn	SD	α	M	Mdn	SD	α	d
SUI	75.78	78.61	24.42	0.77	69.75	65.97	22.97	0.75	-0.25
HLP	59.84	59.74	15.91	0.73	59.34	59.74	14.86	0.67	-0.03
SFD	59.69	65.11	12.56	0.78	60.63	65.11	12.32	0.76	0.07
NFC	57.78	58.17	12.45	0.80	57.07	54.13	12.28	0.78	-0.06
STW	56.56	57.44	10.63	0.49	58.35	57.44	11.96	0.64	0.15
AXY	62.35	59.37	14.77	0.47	62.19	59.37	16.52	0.64	-0.01
ANP	55.06	54.61	11.10	0.69	54.65	54.03	12.04	0.76	-0.04
BRF	54.96	55.77	13.05	0.59	53.95	55.77	11.99	0.52	-0.08
MSF	50.72	50.95	9.77	0.74	47.75	45.62	8.97	0.72	-0.32

Note. SUI refers to Suicide/Death Ideation, HLP refers to Helplessness/Hopelessness, SFD refers to Self-Doubt, NFC refers to Inefficacy, STW refers to Stress/Worry, AXY

Table 5 Continued

refers to Anxiety, ANP refers to Anger Proneness, BRF refers to Behavior Restricting Fears, and MSF refers to Multiple Specific Fears. M refers to mean, Mdn refers to Median, SD refers to standard deviation, α refers to alpha coefficient, and d refers to Cohen's d . Cohen's d was calculated by subtracting the mean scores of African Americans from the mean scores of Caucasians. N for African Americans = 134 for mean, median and standard deviation, n for alpha coefficients varied by scale between 130- 134; N for Caucasians = 1,245 for mean, median and standard deviation, n for alpha coefficients varied by scale between 1,232- 1,242. Unrounded, untruncated T-scores were used to obtain the mean, median, and standard deviation descriptive statistics and raw data were used for the reliability analyses.

In the Pearson sample, Cronbach's alpha internal reliability coefficients ranged from 0.52 (BRF) to 0.81 (ANP) in the African American sample and 0.55 (BRF) to 0.85 (SFD) in the Caucasian sample. The majority of the Cronbach's alpha coefficients for each scale were similar across the African American and Caucasian sample, with the exception of HLP and SFD. In the inpatient sample, Cronbach's alpha coefficients ranged from 0.47 (AXY) to 0.80 (NFC) in the African American population and 0.52 (BRF) to 0.78 (NFC) in the Caucasian population. Again, a number of the scale's Cronbach's alphas were similar in the African American and Caucasian samples, with the exceptions of STW and AXY.

When comparing across ethnicities and Pearson/inpatient samples, a number of alpha coefficients remain comparable. However, for inpatient African American test takers, the alpha coefficients are lower for the STW, AXY, and ANP scales and higher for the HLP scale compared to the Pearson sample of African American test takers. For inpatient Caucasian test takers, alpha coefficients for STW and AXY scales are lower while MSF's alpha coefficient is higher when compared to the Pearson sample of Caucasian test takers. Specific Problem Scale intercorrelations by sample and ethnicity can be found in Table 6 and Table 7.

Table 6

Specific Problem Scale Correlations by Ethnicity for the Pearson Sample

	SUI	HLP	SFD	NFC	STW	AXY	ANP	BRF	MSF
SUI		.37**	.44**	.28**	.34**	.31**	.29**	.26**	.16**
HLP	.51**		.58**	.54**	.57**	.54**	.47**	.42**	.36**
SFD	.46**	.65**		.59**	.62**	.61**	.49**	.37**	.30**
NFC	.38**	.62**	.73**		.62**	.56**	.64**	.47**	.38**
STW	.36**	.57**	.67**	.61**		.58**	.59**	.51**	.34**
AXY	.45**	.57**	.62**	.58**	.59**		.45**	.57**	.35**
ANP	.38**	.49**	.56**	.58**	.58**	.52**		.44**	.19**
BRF	.29**	.41**	.46**	.48**	.45**	.54**	.39**		.36**
MSF	.08**	.11**	.17*	.22**	.23**	.19**	.15**	.28**	

Table 6 Continued

Note. African American ($N = 225$) correlations in the upper diagonal. Caucasian ($N = 2755$) correlations in the lower diagonal. Raw scale totals were used in the analysis. SUI refers to Suicide/Death Ideation, HLP refers to Helplessness/Hopelessness, SFD refers to Self-Doubt, NFC refers to Inefficacy, STW refers to Stress/Worry, AXY refers to Anxiety, ANP refers to Anger Proneness, BRF refers to Behavior Restricting Fears, and MSF refers to Multiple Specific Fears. *Correlations significant at 0.01 level; **Correlation significant at <0.01 level.

Table 7

Specific Problem Scale Correlations by Ethnicity for the Inpatient Sample

	SUI	HLP	SFD	NFC	STW	AXY	ANP	BRF	MSF
SUI		.59**	.55**	.32**	.44**	.44**	.16*	.02	.00
HLP	.48**		.55**	.41**	.32**	.39**	.06	.11	.14
SFD	.47**	.55**		.58**	.55**	.54**	.29**	.09	.11
NFC	.31**	.49**	.62**		.57**	.43**	.39**	.33**	.31**
STW	.34**	.46**	.57**	.59**		.44**	.41**	.15	.24**
AXY	.39**	.39**	.44**	.45**	.53**		.41**	.33**	.29**
ANP	.22**	.29**	.37**	.42**	.46**	.42**		.17	.23**
BRF	.06*	.17**	.19**	.36**	.31**	.38**	.31**		.32**
MSF	.01	.12**	.10**	.25**	.22**	.19**	.16**	.43**	

Table 7 Continued

Note. African American ($N = 134$) correlations in the upper diagonal. Caucasian ($N = 1,245$) correlations in the lower diagonal. Raw scale totals were used in the analysis. SUI refers to Suicide/Death Ideation, HLP refers to Helplessness/Hopelessness, SFD refers to Self-Doubt, NFC refers to Inefficacy, STW refers to Stress/Worry, AXY refers to Anxiety, ANP refers to Anger Proneness, BRF refers to Behavior Restricting Fears, and MSF refers to Multiple Specific Fears. *Correlations significant at 0.05 level; **Correlation significant at 0.01 level.

Population Heterogeneity and Differential Item Functioning (MIMIC Models)

Suicidal/Death Ideation (SUI).

Pearson sample. The factor loadings, thresholds, and model fit indices of the baseline CFA model for the SUI scale are presented in Table 8. For the Pearson sample, factor loadings varied from 0.57 to 0.90. Thresholds for this sample ranged from -1.82 to -1.30. The model fit indices for the baseline CFA, presented at the bottom of Table 8, indicate a good model fit. The error terms for Items 93 and 164 were allowed to covary based on a modification index of 23.75 and review of item content similarity. The standard estimated value for the residual covariance of items 93 and 164 was 0.77 ($p < .01$). There was only a slight change in model fit indices, namely RMSEA decreased from 0.04 to 0.02, when the aforementioned error terms were free to correlate.

Table 8

Suicidal/Death Ideation (SUI) Baseline CFA Factor Loadings, Thresholds, and Model Fit Indices for the Pearson and Inpatient Samples

Item	Pearson Sample ($N = 2,966$)		Inpatient Sample ($N = 1,362$)	
	Factor Loading	Threshold	Factor Loading	Threshold
93	0.83	-1.37	0.85	0.81
120	0.90	-1.82	0.80	-0.01
164	0.89	-1.72	0.87	0.40
251	0.57	-1.76	0.60	0.95
334	0.81	-1.30	0.70	0.38
CFI	0.99		0.99	
TLI	0.99		0.99	
RMSEA	0.02		0.06	
χ^2	7.41 ($p = .12$)		22.22 ($p < .01$)	

Note. CFI = Comparative Fit Index. TLI = Tucker-Lewis Index. RMSEA = Root Mean Square Error of Approximation.

No group differences on latent mean SUI scores were found ($\beta = 0.16$, $SE = 0.13$, $p = 0.22$). Fit indices changed slightly, with only RMSEA decreasing from 0.02 to 0.01, with the addition of the ethnicity covariate in the model. No statistically significant differential item functioning was found when paths were freed from ethnicity to each indicator. Figure 1 shows the partially standardized estimates of the

final SUI MIMIC model with lack of differential item functioning for the Pearson sample. Please note that the indicators on the right side of the all of the figures in this document represent latent continuous response variables, not the initial dichotomous test items, for Minnesota Multiphasic Personality Inventory-2-RF items.

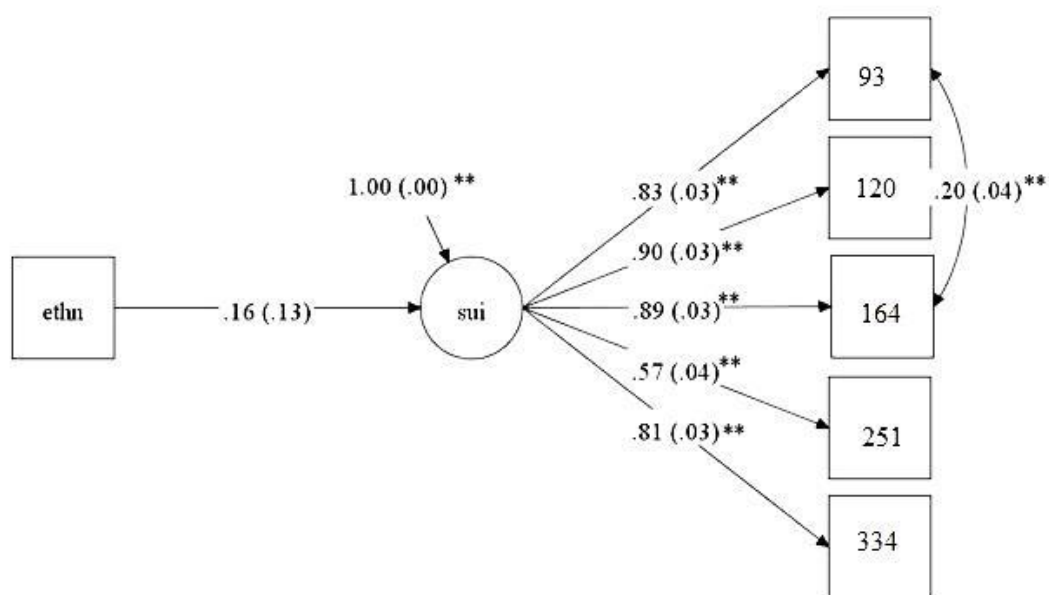


Figure 1. The MIMIC model for the Suicidal/Death Ideation (SUI) Scale in the Pearson sample. All estimates are partially standardized and standard errors are in parenthesis following the estimates. Ethn refers to Ethnicity. Sui refers to Suicidal/Death Ideation. **Estimates significant at 0.01 level.

Inpatient sample. The baseline CFA model for the SUI scale's factor loadings, thresholds, and model fit indices are also presented in Table 8. For the inpatient

sample, factor loadings varied from 0.60 to 0.87. Thresholds for the inpatient sample ranged from -0.01 to 0.95. The model fit indices for the baseline CFA in the inpatient sample indicate a good model fit. Based on a modification index of 34.17 and item content similarity, residuals for items 93 and 164 were allowed to covary. The standard estimated value for the residual covariance of items 93 and 164 was 0.24 in the inpatient sample ($p < .01$). There was a slight change in model fit indices, namely RMSEA decreased from 0.09 to 0.06, when the aforementioned error terms were free to correlate.

In the inpatient sample, African American men scored 0.33 standard scores higher on the latent variable of suicidal/death ideation than Caucasian men ($\beta = 0.33$, $SE = 0.11$, $p < 0.01$). Again, RMSEA decreased from 0.06 to 0.05 with the addition of the ethnicity covariate in the model. CFI and TLI's values did not change. Holding Suicidal/Death Ideation constant, African American men had a higher probability of endorsing item 251, related to a secret suicide attempt, when compared to Caucasian men ($\beta = 0.32$, $SE = 0.11$, $p = 0.01$). No further differential item functioning was found for inpatient men on the SUI scale. The final MIMIC model, including a path pointing to differential item functioning, can be seen in Figure 2.

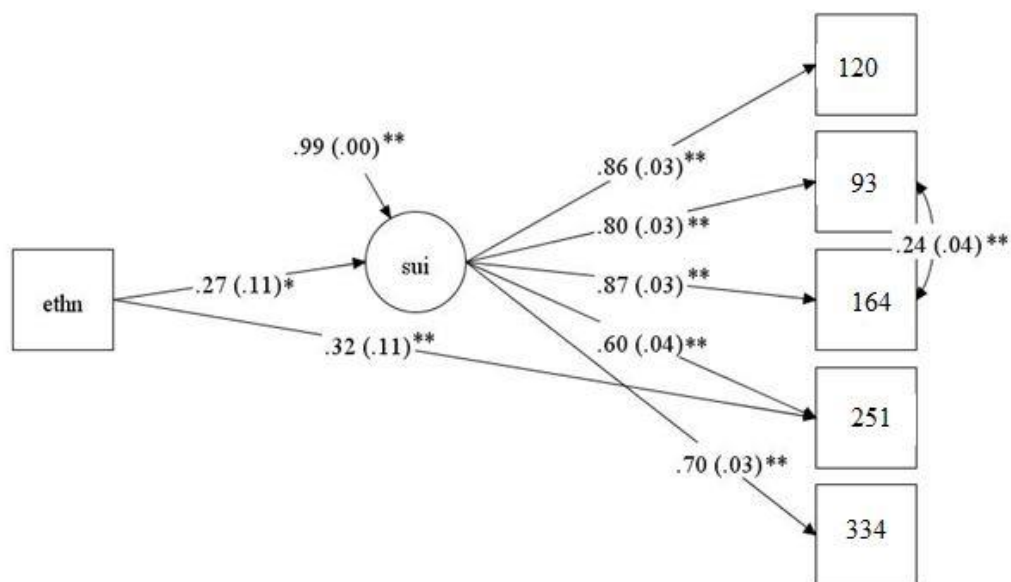


Figure 2. The MIMIC model for the Suicidal/Death Ideation (SUI) Scale in the inpatient sample. All estimates are partially standardized and standard errors are in parenthesis following the estimates. Ethn refers to Ethnicity. Sui refers to Suicidal/Death Ideation. *Estimates significant at 0.05 level; **Estimates significant at 0.01 level.

Helplessness/Hopelessness (HLP).

Pearson sample. Factor loadings, thresholds, and model fit indices for the baseline CFA model of the HLP scale is presented in Table 9. For this sample, factor loadings varied from 0.63 to 0.94. The HLP item thresholds for the sample ranged from 0.56 to 1.25. The model fit indices for the baseline CFA in the Pearson sample indicate a good model fit. A review of modification indices and item content in both samples did not point to the need to allow indicator error term correlations.

Table 9

Helplessness/Hopelessness (HLP) Baseline CFA Factor Loadings, Thresholds, and Model Fit Indices for the Pearson and Inpatient Samples

Item	Pearson Sample ($N = 2,962$)		Inpatient Sample ($N = 1,370$)	
	Factor Loading	Threshold	Factor Loading	Threshold
135	0.68	0.91	0.83	0.35
169	0.94	1.25	0.54	0.05
214	0.84	1.19	0.67	0.56
282	0.78	0.93	0.74	0.25
336	0.63	0.56	0.61	0.21
CFI	0.99		0.98	
TLI	0.99		0.97	
RMSEA	0.03		0.06	
χ^2	20.50 ($p < .01$)		27.99 ($p < .01$)	

Note. CFI = Comparative Fit Index. TLI = Tucker-Lewis Index. RMSEA = Root Mean Square Error of Approximation.

No group differences on latent mean HLP scores were found between African American and Caucasian men ($\beta = 0.00$, $SE = 0.03$, $p = 0.96$). With the addition of the ethnicity covariate in the model, RMSEA increased slightly from 0.03 to 0.04 and CFI and TLI did not change. When paths were freely estimated between ethnicity and

the indicators, two items demonstrated differential functioning. Holding Helplessness/Hopelessness constant, Caucasian men had a higher probability of endorsing items 214, related to helplessness about dissatisfaction with life ($\beta = -0.18$, $SE = 0.04$, $p < 0.01$) and 282, related to not feeling able to reach goals ($\beta = -0.10$, $SE = 0.04$, $p = 0.01$), than African American men. No further evidence of differential item functioning was found. The final MIMIC model, including areas of differential item functioning, can be seen in Figure 3.

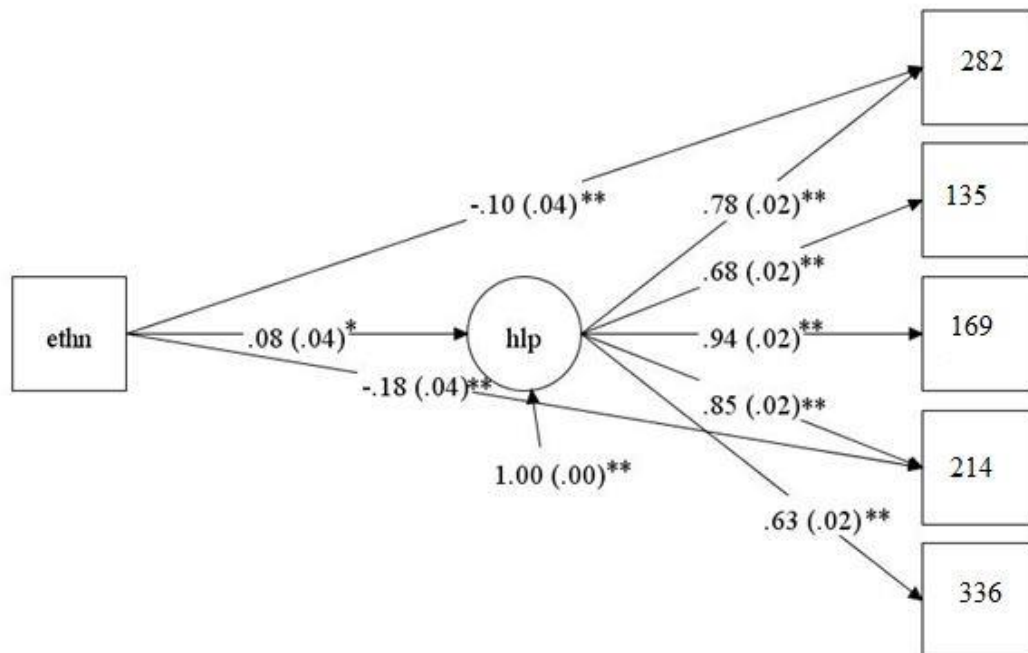


Figure 3. The MIMIC model for the Helplessness/Hopelessness (HLP) Scale in the Pearson sample. All estimates are partially standardized and standard errors are in parenthesis following the estimates. Ethn refers to Ethnicity. Hlp refers to

Figure 3 Continued

Helplessness/Hopelessness. *Estimates significant at 0.05 level; **Estimates significant at 0.01 level.

Inpatient sample. Factor loadings and thresholds for the baseline CFA model of the HLP scale are presented in Table 9. For inpatient men, factor loadings varied from 0.54 to 0.83. Thresholds for the inpatient sample ranged from 0.05 to 0.56. The model fit indices for the baseline CFA in the inpatient sample, also shown in the bottom portion of Table 9, indicate a good model fit. As mentioned in the previous section, no error terms were allowed to correlate based on statistical and practical considerations.

No group differences on latent mean HLP scores were found between African American and Caucasian men ($\beta = 0.05$, $SE = 0.12$, $p = 0.64$). All of the fit indices improved, RMSEA decreasing to 0.04 and CFI and TLI to 0.99 and 0.98, respectively, with the addition of the ethnicity covariate in the model. No statistically significant differential item functioning was found when paths were freed from ethnicity to each indicator. Figure 4 demonstrates the partially standardized estimates for the final HLP MIMIC model for the inpatient sample.

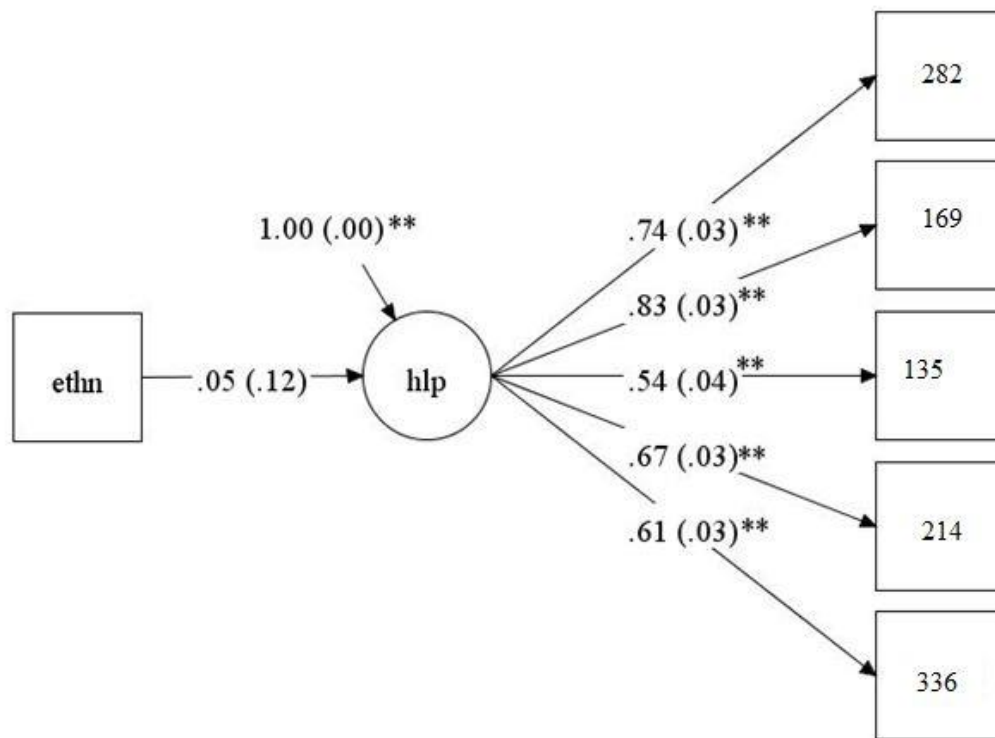


Figure 4. The MIMIC model for the Helplessness/Hopelessness (HLP) Scale in the inpatient sample. All estimates are partially standardized and standard errors are in parenthesis following the estimates. Ethn refers to Ethnicity. Hlp refers to Helplessness/Hopelessness. **Estimates significant at 0.01 level

Self-Doubt (SFD).

Pearson sample. The factor loadings, thresholds, and model fit indices of the baseline CFA model for the SFD scale are presented in Table 10. For the Pearson sample of men, factor loadings were high and varied from 0.87 to 0.94. SFD item thresholds for the sample ranged from 0.40 to 0.70. The model fit indices for the baseline CFA indicate a good model fit in this sample. After review of the

modification indices of the baseline CFA model in both samples, combined with review of similarity in item content, residuals of items 89 and 232 were allowed to correlate. The modification index for items 89 and 232 in the Pearson sample was 34.20. The standard estimated value for the residual covariance of items 89 and 232 was 0.43 in the sample ($p < .01$). Overall, model fit indices improved with the addition of these correlated indicator error terms.

Table 10

Self-Doubt (SFD) Baseline CFA Factor Loadings, Thresholds, and Model Fit Indices for the Pearson and Inpatient Samples

Item	Pearson Sample ($N = 2,976$)		Inpatient Sample ($N = 1,376$)	
	Factor Loading	Threshold	Factor Loading	Threshold
48	0.88	0.69	0.81	-0.16
89	0.87	0.40	0.82	-0.56
232	0.89	0.70	0.81	-0.12
288	0.94	0.71	0.77	0.07
CFI	1.00		0.99	
TLI	1.00		0.99	
RMSEA	0.01		0.07	
χ^2	1.38 ($p = .24$)		8.18 ($p < .01$)	

Note. CFI = Comparative Fit Index. TLI = Tucker-Lewis Index. RMSEA = Root Mean Square Error of Approximation.

Caucasian men scored 0.22 standard scores higher on the latent variable of Self-Doubt than African American men ($\beta = -0.22$, $SE = 0.08$, $p = 0.01$). Model fit indices either improved or stayed the same with the addition of the ethnicity covariate in the model. No significant differential item functioning was found when paths were freed from ethnicity to each indicator for amalgamated sample of African American and Caucasian men on the SFD scale. Figure 5 provides a visual representation of the final MIMIC model for the SFD scale in this sample.

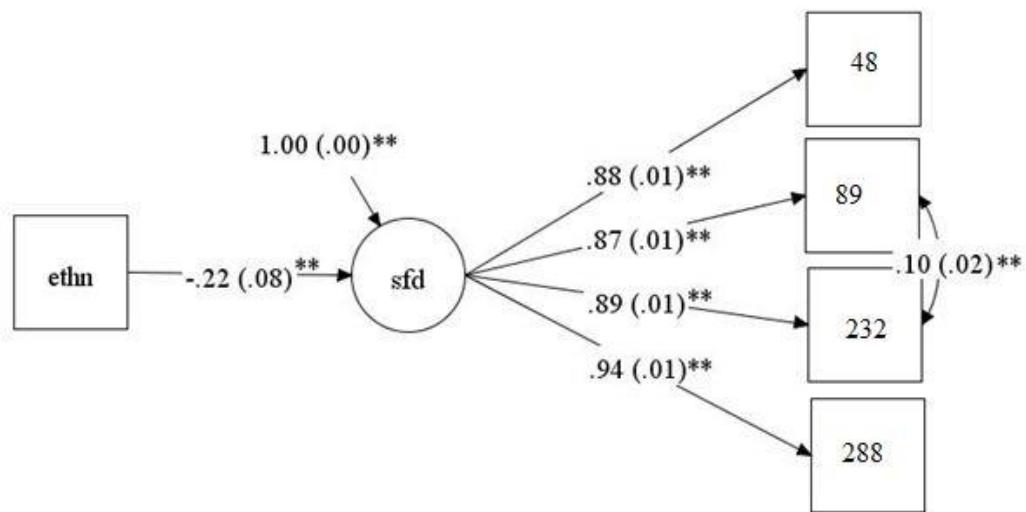


Figure 5. The MIMIC model for the Self-Doubt (SFD) Scale in the Pearson sample.

All estimates are partially standardized and standard errors are in parenthesis following the estimates. Ethn refers to Ethnicity. Sfd refers to Self-Doubt.

**Estimates significant at 0.01 level.

Inpatient sample. Table 10 presents the factor loadings, thresholds, and model fit indices of the baseline CFA model for the SFD scale. For inpatient men, factor loadings ranged from 0.77 to 0.82 and item thresholds varied from -0.56 to 0.07. The model fit indices for the baseline CFA, presented at the bottom of Table 10, indicate an acceptable model fit in this sample. While CFI and TLI indicate a good model fit, RMSEA is slightly high at 0.07 pointing to an acceptable fit. It is worth a reminder that previous research has recommended leniency with model fit indices when using categorical indicators (Ketterer, 2011).

Residuals for items 89 and 232 were allowed to covary based on a modification index of 7.07 and a review of item content. The standard estimated value for the residual covariance of items 89 and 232 was 0.31 ($p < .01$). Overall, model fit indices did not change with the addition of these correlated indicator error terms. No group differences on latent mean SFD scores were found between African American and Caucasian men ($\beta = -0.10$, $SE = 0.11$, $p = 0.39$). Model fit indices improved with the addition of the ethnicity covariate in the model, particularly RMSEA which decreased from 0.07 to 0.03. No significant differential item functioning was found when paths were freed from ethnicity to each indicator for inpatient men on the SFD scale. The final MIMIC model with partially standardized estimates can be seen in Figure 6.

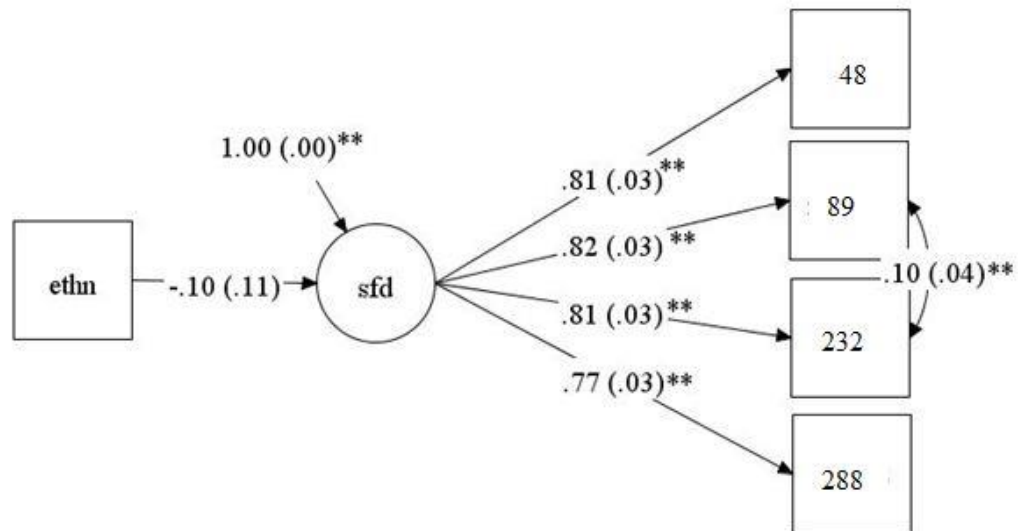


Figure 6. The MIMIC model for the Self-Doubt (SFD) Scale in the inpatient sample.

All estimates are partially standardized and standard errors are in parenthesis following the estimates. Ethn refers to Ethnicity. Sfd refers to Self-Doubt.

**Estimates significant at 0.01 level.

Inefficacy (NFC).

Pearson sample. The NFC scale's baseline CFA model factor loadings and thresholds are presented in Table 11. Factor loadings for the nine NFC items ranged from 0.50 to 0.89 in this sample. Item thresholds for the sample ranged from 0.25 to 0.94. The model fit indices for the baseline CFA, presented at the bottom of Table 11, indicate a good model fit in this sample. Based on a modification index of 16.23 and review of item content similarity, residuals for items 152 and 198 were allowed to correlate. For the Pearson sample, the standard estimated value for the residual

covariance of items 152 and 198 was 0.26 ($p < .01$). The model fit indices did not change with the addition of these correlated indicator error terms.

Table 11

Inefficacy (NFC) Baseline CFA Factor Loadings, Thresholds, and Model Fit Indices for the Pearson and Inpatient Samples

Item	Pearson Sample ($N = 2,962$)		Inpatient Sample ($N = 1,365$)	
	Factor Loading	Threshold	Factor Loading	Threshold
27	0.76	0.49	0.63	-0.10
68	0.50	0.38	0.56	0.11
108	0.80	0.59	0.72	-0.04
152	0.88	0.94	0.73	0.11
198	0.78	0.94	0.72	0.24
229	0.61	0.60	0.53	0.25
271	0.62	0.25	0.40	0.54
274	0.89	0.78	0.84	0.18
324	0.84	0.50	0.80	0.07
CFI	0.99		0.99	
TLI	0.99		0.98	
RMSEA	0.04		0.04	
χ^2	136.44 ($p < .01$)		89.19 ($p < .01$)	

Table 11 Continued

Note. CFI = Comparative Fit Index. TLI = Tucker-Lewis Index. RMSEA = Root Mean Square Error of Approximation.

African American men scored 0.17 standard scores higher on the latent variable of Inefficacy compared to Caucasian men ($\beta = 0.17$, $SE = 0.08$, $p = 0.04$). After addition of the ethnicity covariate in the model, the fit indices increased slightly but still pointed to a good model fit. When paths were freely estimated between ethnicity and the indicators, four items demonstrated differential functioning. Holding Inefficacy constant, African American men had a higher probability of endorsing items 27 ($\beta = 0.32$, $SE = 0.08$, $p < 0.01$) and 68 ($\beta = 0.60$, $SE = 0.09$, $p < 0.01$) when compared to Caucasian men in the Pearson sample. Item 27 relates to difficulty making decisions and thus missing an opportunity and item 68 assesses difficulty taking action in everyday affairs without careful consideration.

Alternatively, controlling for level of Inefficacy, Caucasian men had a higher probability of endorsing items 229 ($\beta = -0.18$, $SE = 0.08$, $p = 0.03$) and 324 ($\beta = -0.36$, $SE = 0.08$, $p < 0.01$) than African American men. Item 229 assesses a test taker's tendency to forego activities if others do not approve and item 324 relates to nervousness in making decisions. No further evidence of differential item functioning was found. The final MIMIC model and differential item functioning partially standardized estimates for the NFC scale can be found in Figure 7 for the Pearson sample.

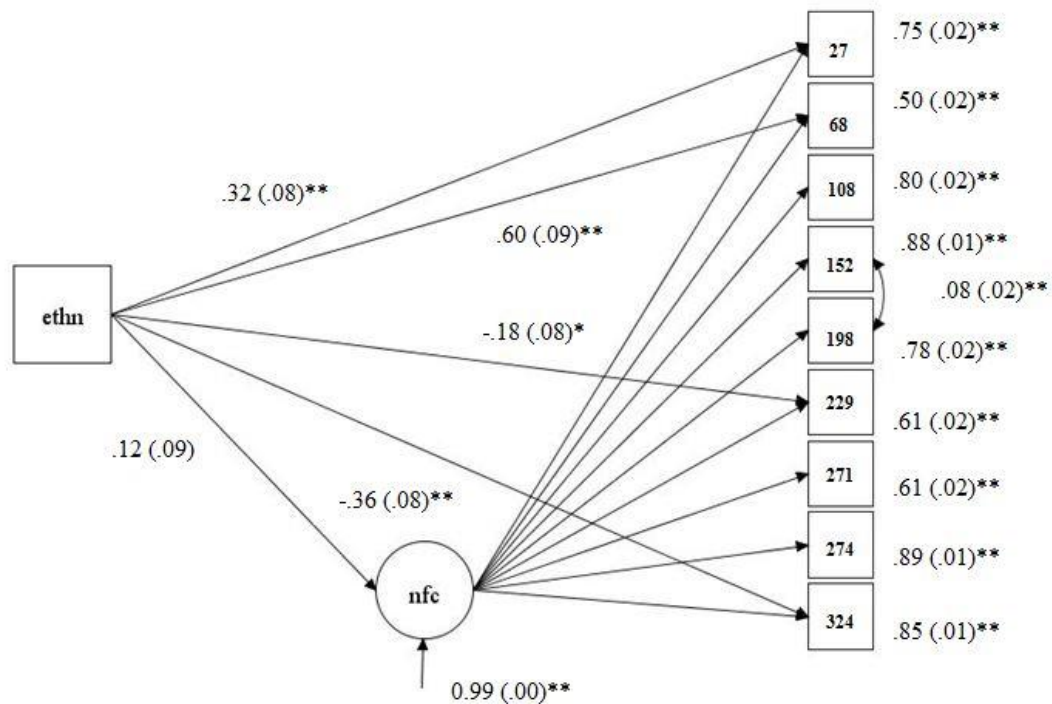


Figure 7. The MIMIC model for the Inefficacy (NFC) Scale in the Pearson sample.

All estimates are partially standardized and standard errors are in parenthesis following the estimates. Estimates for paths between the latent variable and indicators are presented to the right of the indicator for ease of reading. The estimate for the error covariance is presented slightly more to the left of the items for differentiation. Ethn refers to Ethnicity. Nfc refers to Self-Doubt. *Estimates significant at 0.05 level; **Estimates significant at 0.01 level.

Inpatient sample. Baseline CFA model factor loadings and thresholds for the NFC scale are presented in Table 11, with inpatient data on the left. Factor loadings the NFC items ranged from 0.40 to 0.84. Item thresholds for the inpatient sample ranged from -0.10 to 0.54. The model fit indices for the baseline CFA, presented at

the bottom of Table 11, indicate a good model fit in this sample. A modification index of 15.98, combined with review of similarity in item content, pointed to the benefits of allowing the residuals of item 152 and 198 to correlate. The standard estimated value for the residual covariance of items 152 and 198 was 0.25 ($p < .01$). The model fit indices did not substantially change with the addition of these correlated indicator error terms.

No group differences on latent mean NFC scores were found between African American and Caucasian men ($\beta = 0.06$, $SE = 0.11$, $p = 0.61$). Model fit indices either improved slightly or did not change following the addition of the ethnicity covariate in the model. When paths were freely estimated between ethnicity and the indicators, three items demonstrated differential functioning. Holding Inefficacy constant, African American men had a higher probability of endorsing items 27 ($\beta = 0.46$, $SE = 0.10$, $p < 0.01$), 68 ($\beta = 0.32$, $SE = 0.11$, $p < 0.01$), and 108 ($\beta = 0.25$, $SE = 0.10$, $p = 0.01$) when compared to Caucasian men in the inpatient sample. Again, item 27 relates to difficulty making decisions and thus missing an opportunity and item 68 assesses difficulty taking action in everyday affairs without careful consideration. Item 108 assesses giving up on tasks due to lack of self-confidence. No further evidence of differential item functioning was found. Figure 8 visually depicts the final NFC model, including differential item functioning, for the inpatient sample.

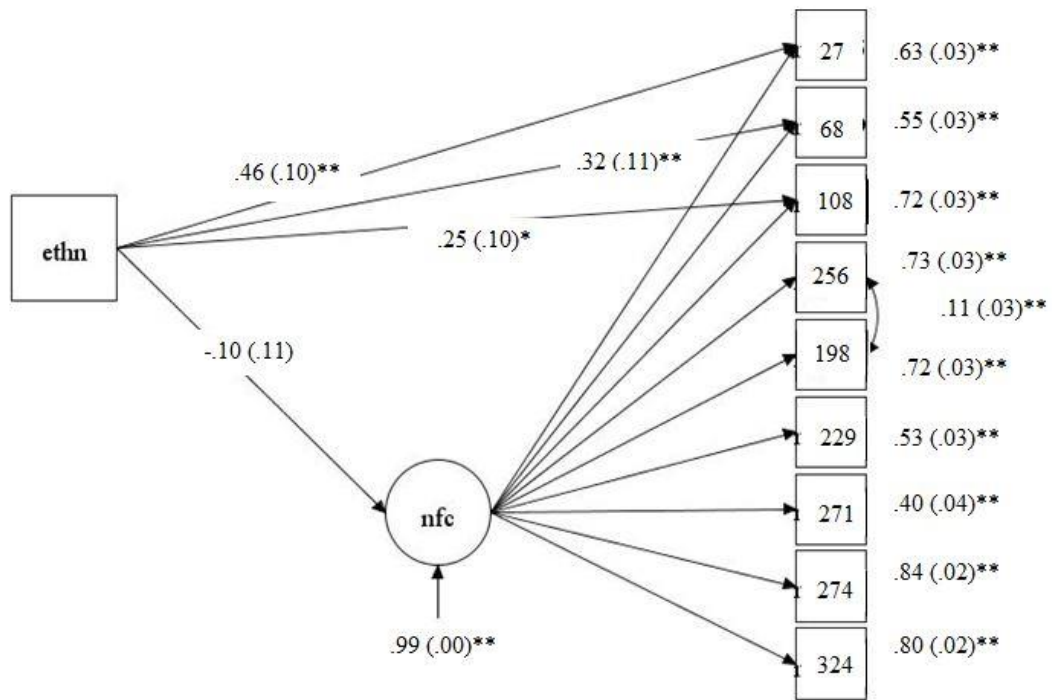


Figure 8. The MIMIC model for the Inefficacy (NFC) Scale in the inpatient sample.

All estimates are partially standardized and standard errors are in parenthesis following the estimates. Estimates for paths between the latent variable and indicators are presented to the right of the indicator for ease of reading. The estimate for the error covariance is presented slightly more to the left of the items for differentiation. Ethn refers to Ethnicity. Nfc refers to Self-Doubt. *Estimates significant at 0.05 level; **Estimates significant at 0.01 level.

Stress/Worry (STW).

Pearson sample. The factor loadings, thresholds, and model fit indices of the baseline CFA model for the STW scale in both samples are presented in Table 12. For this sample, factor loadings the seven STW items ranged from 0.51 to 0.92. Item thresholds for the sample ranged from -0.44 to 0.66. The model fit indices for the baseline CFA, at the bottom of Table 12, indicate a good model fit in this sample. Based on a modification index of 24.56 and review of item content, the residuals for items 73 and 167 were allowed to covary. The standard estimated value for the residual covariance of items 73 and 167 was 0.20 ($p < .01$). The model fit indices demonstrated only minor improvement with the addition of the correlated indicator error terms.

Table 12

Stress/Worry (STW) Baseline CFA Factor Loadings, Thresholds, and Model Fit Indices for the Pearson and Inpatient Samples

Item	Pearson Sample ($N = 2,959$)		Inpatient Sample ($N = 1,371$)	
	Factor Loading	Threshold	Factor Loading	Threshold
29	0.65	0.40	0.64	-0.07
73	0.54	0.66	0.48	0.10
123	0.92	0.46	0.77	-0.10
167	0.69	0.59	0.52	-0.17
224	0.51	0.59	0.37	0.48

Table 12 Continued

Item	Pearson Sample ($N = 2,959$)		Inpatient Sample ($N = 1,371$)	
	Factor Loading	Threshold	Factor Loading	Threshold
234	0.65	-0.44	0.56	-0.84
309	0.70	0.21	0.57	-0.26
CFI	0.99		0.98	
TLI	0.98		0.96	
RMSEA	0.04		0.04	
χ^2	84.04 ($p < .01$)		40.76 ($p < .01$)	

Note. CFI = Comparative Fit Index. TLI = Tucker-Lewis Index. RMSEA = Root Mean Square Error of Approximation.

No group differences on latent mean STW scores were found between African American and Caucasian men in the Pearson sample ($\beta = -0.01$, $SE = 0.08$, $p = 0.94$). Model fit indices did not change with the addition of the ethnicity covariate in the model. When paths were freely estimated between ethnicity and the indicators, two items demonstrated differential functioning. Controlling for level of Stress/Worry, Caucasian men had a higher probability of endorsing items 73 ($\beta = -0.20$, $SE = 0.10$, $p = 0.04$) and 234 ($\beta = -0.23$, $SE = 0.09$, $p = 0.01$) when compared to African American men in the Pearson sample. Item 73 assesses level of nervousness compared to others and item 234 relates to feeling stress and/or pressure. No further evidence of

differential item functioning was evident. The partially standardized estimates of the final STW model can be seen in Figure 9.

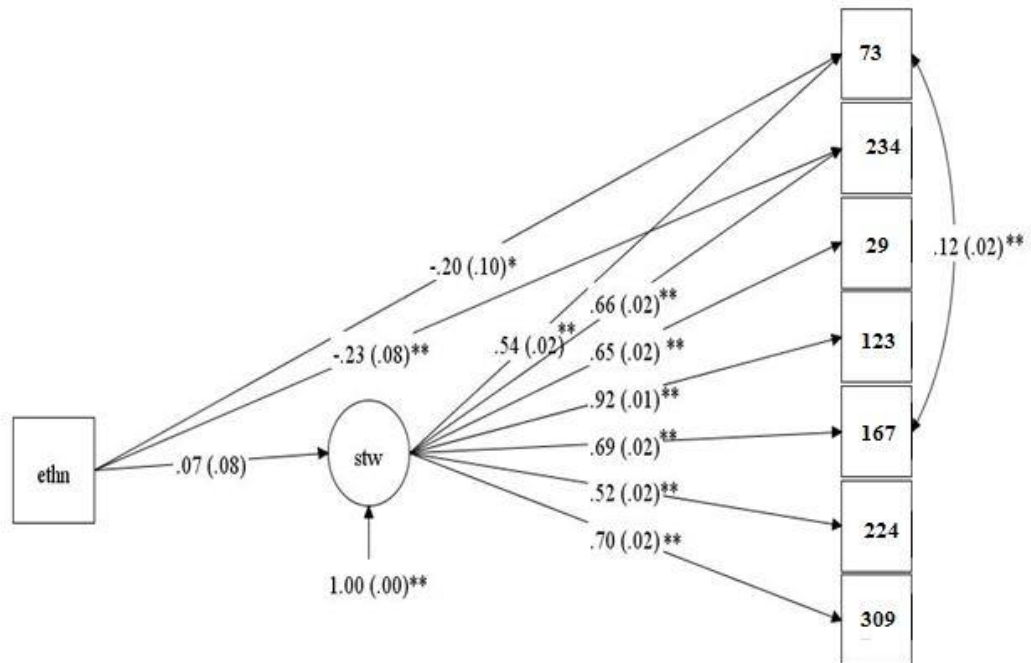


Figure 9. The MIMIC model for the Stress/Worry (STW) Scale in the Pearson sample.

All estimates are partially standardized and standard errors are in parenthesis following the estimates. Ethn refers to Ethnicity. Stw refers to Self-Doubt.

*Estimates significant at 0.05 level; **Estimates significant at 0.01 level.

Inpatient sample. Factor loadings and thresholds for the baseline CFA model of the STW scale are presented in Table 12, with inpatient data on the left. Factor loadings the seven STW items ranged from 0.37 to 0.77. The STW item thresholds

for the inpatient sample varied from -0.84 to 0.48. The model fit indices for the baseline CFA, presented at the bottom of Table 12, indicate a good model fit in this sample. Again, a modification index of 8.38 and review of item content again pointed to the need to allow residuals for items 73 and 167 to correlate. The standard estimated value for the residual covariance of items 73 and 167 was 0.15 for the inpatient sample ($p < .01$). The model fit indices did not substantially change with the addition of these correlated indicator error terms.

No group differences on latent mean STW scores were found between African American and Caucasian men in the inpatient sample ($\beta = -0.12$, $SE = 0.10$, $p = 0.26$). Again, model fit indices did not change substantially with the addition of the ethnicity covariate in the model. Holding level of Stress/Worry constant, African American men had a higher probability of endorsing item 123, related to worry over potential mishaps, than Caucasian men in the inpatient sample ($\beta = 0.34$, $SE = 0.12$, $p < 0.01$). No further evidence of differential item functioning was found. Figure 10 provides the partially standardized estimates for the final STW model in the inpatient sample.

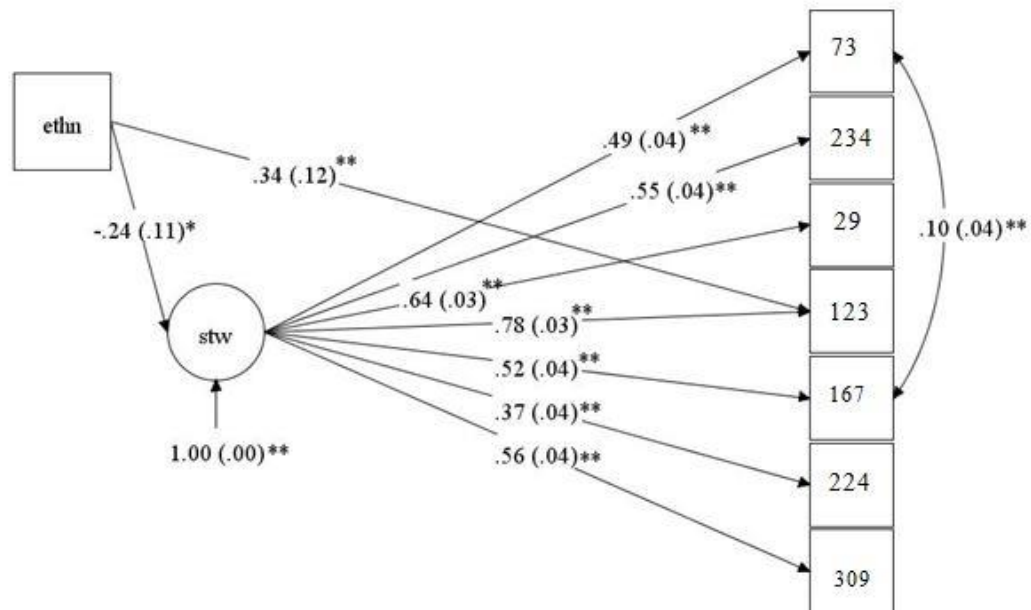


Figure 10. The MIMIC model for the Stress/Worry (STW) Scale in the inpatient sample. All estimates are partially standardized and standard errors are in parenthesis following the estimates. Ethn refers to Ethnicity. Stw refers to Self-Doubt.

*Estimates significant at 0.05 level; **Estimates significant at 0.01 level.

Anxiety (AXY).

Pearson sample. The factor loadings, thresholds, and model fit indices of the baseline CFA model for the AXY scale in both samples are presented in Table 13. For the Pearson sample, factor loadings the five AXY items ranged from 0.67 to 0.91. Item thresholds for this sample varied from 0.75 to 1.62. The model fit indices for the baseline CFA, at the bottom of Table 13, indicate a good model fit in this sample. Residuals for items 79 and 289 were allowed to correlate based on a modification

index of 49.18 and a review of item content similarity. The standard estimated value for the residual covariance of items 79 and 289 was 0.41 ($p < .01$). There was a change in model fit indices, namely RMSEA decreased from 0.06 to 0.02, when the aforementioned error terms were free to correlate.

Table 13

Anxiety (AXY) Baseline CFA Factor Loadings, Thresholds, and Model Fit Indices for the Pearson and Inpatient Samples

Item	Pearson Sample ($N = 2,971$)		Inpatient Sample ($N = 1,372$)	
	Factor Loading	Threshold	Factor Loading	Threshold
79	0.67	1.17	0.50	0.42
146	0.80	1.62	0.61	0.15
228	0.89	0.75	0.68	1.17
275	0.91	1.26	0.80	0.59
289	0.72	1.34	0.62	0.64
CFI	0.99		0.99	
TLI	0.99		0.99	
RMSEA	0.02		0.03	
χ^2	6.50 ($p = .17$)		8.48 ($p = .08$)	

Note. CFI = Comparative Fit Index. TLI = Tucker-Lewis Index. RMSEA = Root Mean Square Error of Approximation.

No group differences on latent mean AXY scores were found between African American and Caucasian men in the Pearson sample ($\beta = 0.15$, $SE = 0.09$, $p = 0.09$). Model fit indices did not change substantially after the addition of the ethnicity covariate in the model. Controlling for level of Anxiety, Caucasian men had a higher probability of endorsing item 228, related to constant anxiety, when compared to African American men in the Pearson sample ($\beta = -0.27$, $SE = 0.10$, $p = 0.01$). No further evidence of differential item functioning was evident. The final AXY model for this sample, including areas of differential item functioning, is visually depicted in Figure 11.

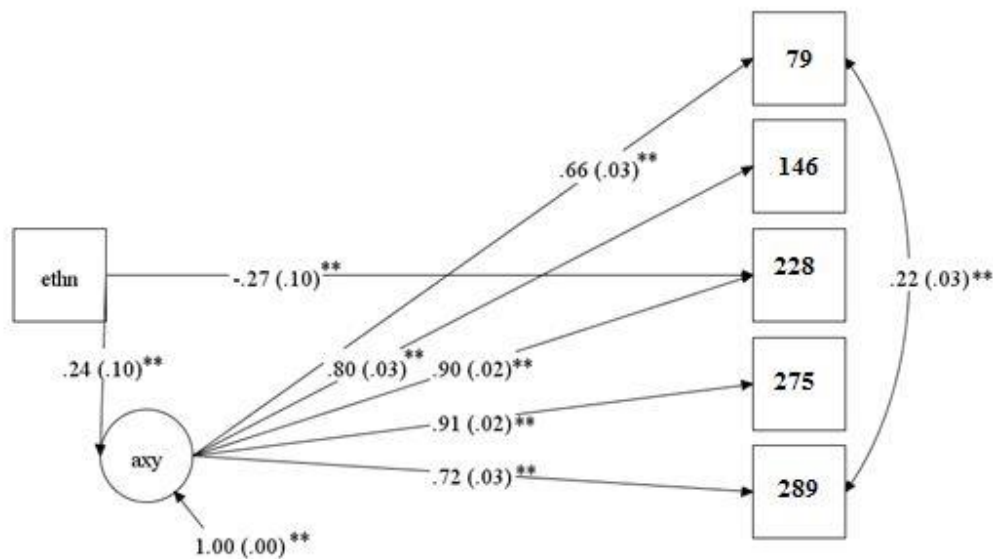


Figure 11. The MIMIC model for the Anxiety (AXY) Scale in the Pearson sample.

All estimates are partially standardized and standard errors are in parenthesis

Figure 11 Continued

following the estimates. Ethn refers to Ethnicity. Axy refers to Anxiety. **Estimates significant at 0.01 level.

Inpatient sample. The baseline CFA model for the AXY scale's factor loadings, thresholds, and model fit indices are presented in Table 13. For the inpatient sample, factor loadings varied from 0.50 to 0.80. Item thresholds for the inpatient sample ranged from 0.15 to 1.17. The model fit indices for the baseline CFA in the inpatient sample indicate a good model fit. Residuals for items 79 and 289 were allowed to covary based on a modification index of 32.21 and review of item content similarity. The standard estimated value for the residual covariance of items 79 and 289 was 0.35 ($p < .01$). Model fit was improved overall with the addition of the aforementioned correlated error terms.

No group differences on latent mean AXY scores were found between African American and Caucasian men in the inpatient sample ($\beta = -0.04$, $SE = 0.11$, $p = 0.72$). No notable changes in model fit indices resulted from the addition of the ethnicity covariate in the model. Controlling for level of Anxiety, African American men had a higher probability of endorsing item 289, related to frequent fear in the night, than Caucasian men ($\beta = 0.27$, $SE = 0.12$, $p = 0.02$). No further differential item functioning was found for inpatient men on the AXY scale. Figure 12 depicts the partially standardized estimates of final AXY MIMIC model for inpatient men.

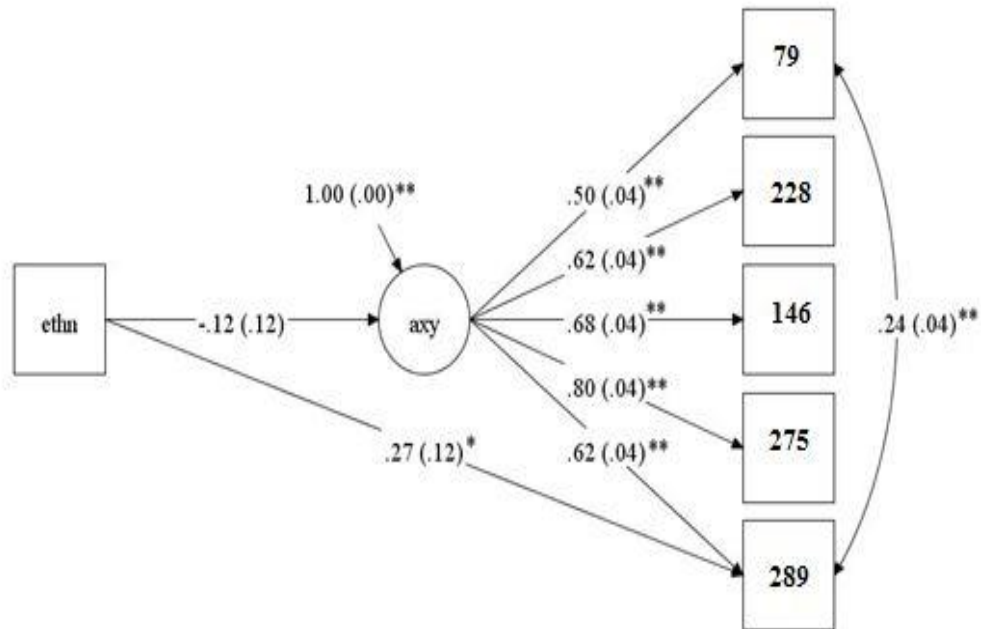


Figure 12. The MIMIC model for the Anxiety (AXY) Scale in the inpatient sample.

All estimates are partially standardized and standard errors are in parenthesis following the estimates. Ethn refers to Ethnicity. Axy refers to Anxiety. *Estimates significant at 0.05 level; **Estimates significant at 0.01 level.

Anger Proneness (ANP).

Pearson sample. The factor loadings and thresholds of the baseline CFA model for the ANP scale in both samples are presented in Table 14. Factor loadings of the ANP items ranged from 0.69 to 0.89 for Pearson sample of African American and Caucasian men. ANP item thresholds for the sample ranged from 0.48 to 1.05. The model fit indices for the baseline CFA, at the bottom left side of Table 14, indicate a

good model fit in this sample. A modification index of 15.56 and review of item content similarities pointed to the benefit of allowing item 134 and 293 residuals to correlate. The standard estimated value for the residual covariance of items 134 and 293 was 0.07 ($p < .01$). The model fit indices demonstrated small improvement with the addition of the correlated indicator error terms.

Table 14

Anger Proneness (ANP) Baseline CFA Factor Loadings, Thresholds, and Model Fit Indices for the Pearson and Inpatient Samples

Item	Pearson Sample ($N = 2,966$)		Inpatient Sample ($N = 1,367$)	
	Factor Loading	Threshold	Factor Loading	Threshold
134	0.89	0.75	0.80	0.23
119	0.82	0.48	0.76	0.05
155	0.69	0.69	0.76	0.05
248	0.82	1.05	0.76	0.63
293	0.86	1.01	0.60	0.37
303	0.76	0.56	0.60	0.20
318	0.83	0.90	0.69	0.15
CFI	0.99		0.99	
TLI	0.99		0.98	
RMSEA	0.04		0.05	
χ^2	68.49 ($p < .01$)		53.95 ($p < .01$)	

Table 14 Continued

Note. CFI = Comparative Fit Index. TLI = Tucker-Lewis Index. RMSEA = Root Mean Square Error of Approximation.

No group differences on latent mean ANP scores were found between African American and Caucasian men in the Pearson sample ($\beta = 0.16$, $SE = 0.08$, $p = 0.05$). Addition of the ethnicity covariate into the model did not produce any notable changes in the model fit indices. When paths were freely estimated between ethnicity and the indicators, three items demonstrated differential functioning. Controlling for level of Anger Proneness, African American men had a higher probability of endorsing items 248 ($\beta = 0.28$, $SE = 0.09$, $p < 0.01$), 303 ($\beta = 0.36$, $SE = 0.08$, $p < 0.01$), and 318 ($\beta = 0.34$, $SE = 0.09$, $p < 0.01$) when compared to Caucasian men in the Pearson sample. Item 248 assesses for a quick temper, item 303 for irritability at disruptions, and item 318 for occasional uncontrollable anger. No further evidence of differential item functioning was evident. The partially standardized estimates of the final ANP MIMIC model, including paths pointing to differential item functioning, can be seen in Figure 13.

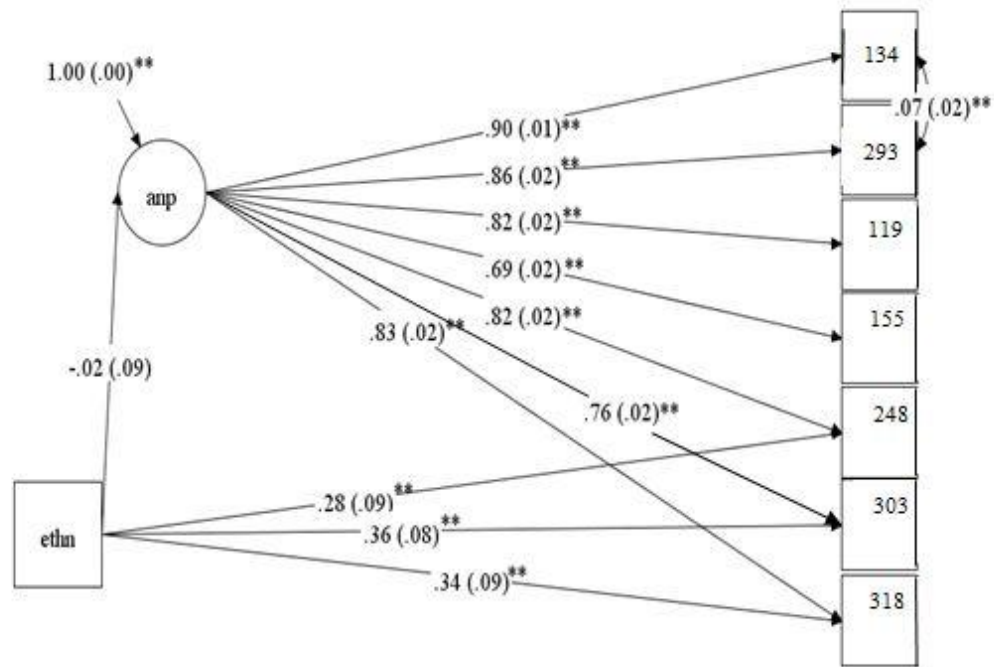


Figure 13. The MIMIC model for the Anger Proneness (ANP) Scale in the Pearson sample. All estimates are partially standardized and standard errors are in parenthesis following the estimates. Ethn refers to Ethnicity. Anp refers to Anger Proneness.

**Estimates significant at 0.01 level.

Inpatient sample. Again, both of the samples' factor loadings, thresholds, and model fit indices of the baseline CFA model for the ANP scale are presented in Table 14. For inpatient men, factor loadings of the seven ANP items ranged from 0.60 to 0.80. Item thresholds for the sample ranged from 0.05 to 0.63. The model fit indices for the baseline CFA, at the bottom right side of Table 14, indicate a good model fit in this sample. A modification index of 19.93 and review of similarities in item content

pointed to the benefit of allowing residuals of item 134 and 293 to covary. The standard estimated value for the residual covariance of items 134 and 293 was 0.31 ($p < .01$). The model fit indices improved overall with the addition of the correlated error terms.

No group differences on latent mean ANP scores were found between African American and Caucasian men in the inpatient sample ($\beta = 0.05$, $SE = 0.10$, $p = 0.64$). With the addition of the ethnicity covariate in the model, RMSEA decreased from 0.05 to 0.04 while CFI and TLI remained the same. Holding level of Anger Proneness constant, Caucasian men had a higher probability of endorsing items 293, related to a tendency to become upset easily, than African American men in the inpatient sample ($\beta = -0.30$, $SE = 0.11$, $p < 0.01$). No further evidence of differential item functioning was evident. The partially standardized estimates of the final ANP model for inpatient men can be seen in Figure 14.

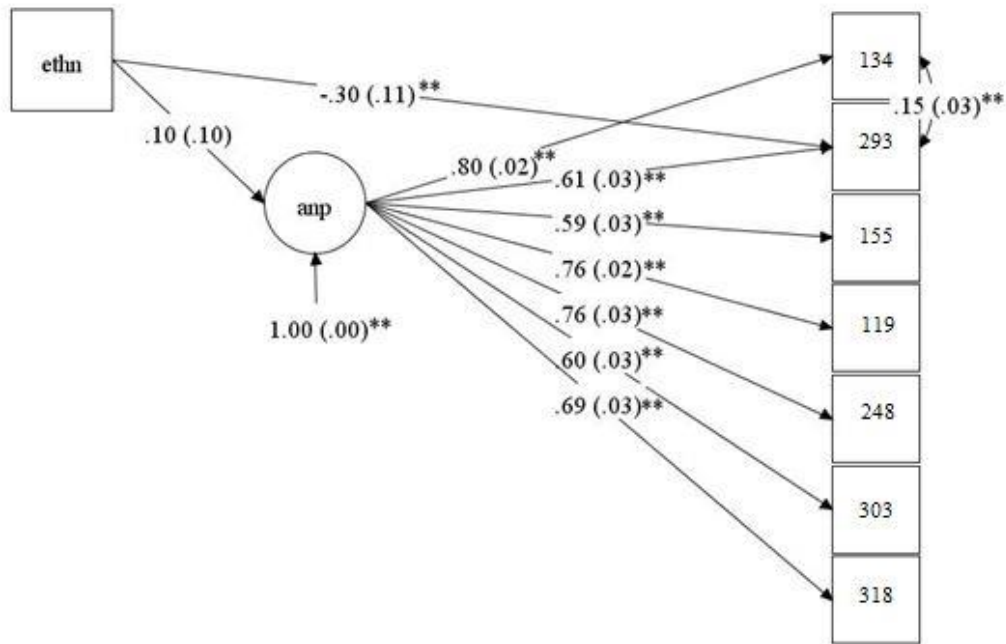


Figure 14. The MIMIC model for the Anger Proneness (ANP) Scale in the inpatient sample. All estimates are partially standardized and standard errors are in parenthesis following the estimates. Ethn refers to Ethnicity. Anp refers to Anger Proneness. **Estimates significant at 0.01 level.

Behavior Restricting Fears (BRF).

Pearson sample. The factor loadings, thresholds, and model fit indices of the baseline CFA model for the BRF scale in both samples are presented in Table 15. Factor loadings of the BRF items ranged from 0.42 to 0.74 for the Pearson sample. The BRF item thresholds for the sample ranged from 1.25 to 2.11. The model fit indices for the baseline CFA indicate a good model fit in this sample. A review of

modification indices and item content in both samples did not point to the need for indicator covariance.

Table 15

Behavior Restricting Fears (BRF) Baseline CFA Factor Loadings, Thresholds, and Model Fit Indices for the Pearson and Inpatient Samples

Item	Pearson Sample ($N = 2,971$)		Inpatient Sample ($N = 1,374$)	
	Factor Loading	Threshold	Factor Loading	Threshold
20	0.61	1.52	0.41	1.02
56	0.67	1.44	0.45	1.04
90	0.66	1.25	0.58	0.73
128	0.43	1.62	0.22	1.05
165	0.66	1.68	0.78	1.38
208	0.46	1.93	0.52	1.50
243	0.52	1.88	0.47	1.15
284	0.74	1.88	0.72	1.38
317	0.68	2.11	0.67	1.78
CFI	0.98		0.99	
TLI	0.97		0.99	
RMSEA	0.02		0.01	
χ^2	51.21 ($p < .01$)		34.76 ($p = .15$)	

Table 15 Continued

Note. CFI = Comparative Fit Index. TLI = Tucker-Lewis Index. RMSEA = Root Mean Square Error of Approximation.

In the Pearson sample, African American men scored 0.44 standard scores higher on the latent variable of Behavior Restricting Fears than Caucasian men ($\beta = 0.44$, $SE = 0.11$, $p < 0.01$). Model fit indices worsened after inclusion of the ethnicity covariate in the model, with RMSEA increasing from 0.02 to 0.03 and CFI and TLI decreasing from 0.97 to 0.92 and 0.97 to 0.90, respectively. After inclusion of the covariate in the model, the CFI and TLI values point to less than optimal model fit while the RMSEA value continues to indicate good model fit. When paths were freely estimated between ethnicity and the indicators, three items demonstrated differential functioning. Controlling for level of Behavior Restricting Fears, African American men had a higher probability of endorsing items 208 ($\beta = 0.57$, $SE = 0.13$, $p < 0.01$) and 243 ($\beta = 0.77$, $SE = 0.12$, $p < 0.01$) when compared to Caucasian men in the Pearson sample. Item 208 assesses fear of using a sharp object and item 243 assesses fear or dislike of dirt.

Holding level of Behavior Restricting Fears constant, Caucasian men had a higher probability of endorsing item 165, related to fear of the dark, than African American men in this sample ($\beta = -0.71$, $SE = 0.22$, $p < 0.01$). No further evidence of differential item functioning was evident. Figure 15 shows the partially standardized

estimates of the final model for the Pearson sample of African American and Caucasian men.

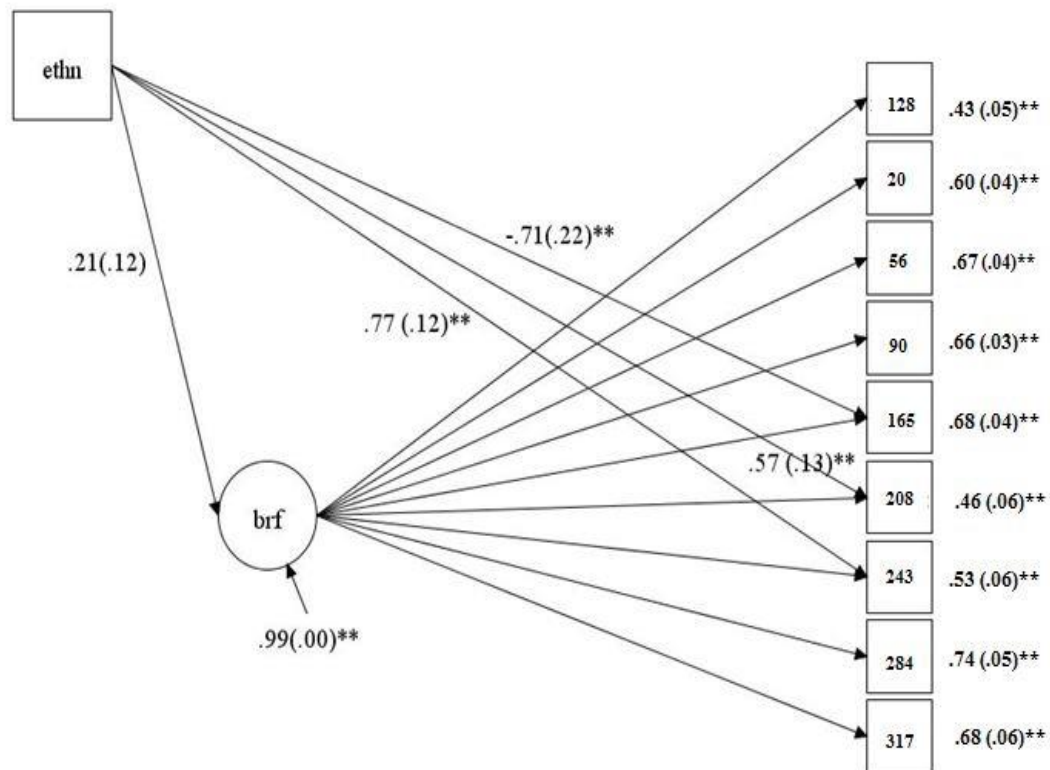


Figure 15. The MIMIC model for the Behavior Restricting Fears (BRF) Scale in the Pearson sample. All estimates are partially standardized and standard errors are in parenthesis following the estimates. Estimates for paths between the latent variable and indicators are presented to the right of the indicator for ease of reading. Ethn refers to Ethnicity. Brf refers to Behavior Restricting Fears. **Estimates significant at 0.01 level.

Inpatient sample. As presented in Table 15, inpatient men demonstrated factor loadings of the BRF items ranging from 0.22 to 0.78. The BRF item thresholds for the sample varied from 0.73 to 1.50. The model fit indices for the baseline CFA, at the bottom right side of Table 15, indicate a good model fit in this sample. Again, review of modification indices and item content in both samples did not point to the need for indicator covariance. No group differences on latent mean BRF scores were found between African American and Caucasian men in the inpatient sample ($b = 0.19$, $SE = 0.13$, $p = 0.15$). Model fit indices changed slightly, with an increase in RMSEA and decrease in both CFI and TLI, with inclusion of the ethnicity covariate in the model. However, the indices continued to point to a good model fit for the data.

Holding level of Behavior Restricting Fears constant, Caucasian men had a higher probability of endorsing items 56, related to anxiety about leaving the house, than African American men in the inpatient sample ($\beta = -0.55$, $SE = 0.17$, $p < 0.01$). No further evidence of differential item functioning was evident. Figure 16 provides a visual representation of the final MIMIC model with differential item functioning for the BRF scale in this sample.

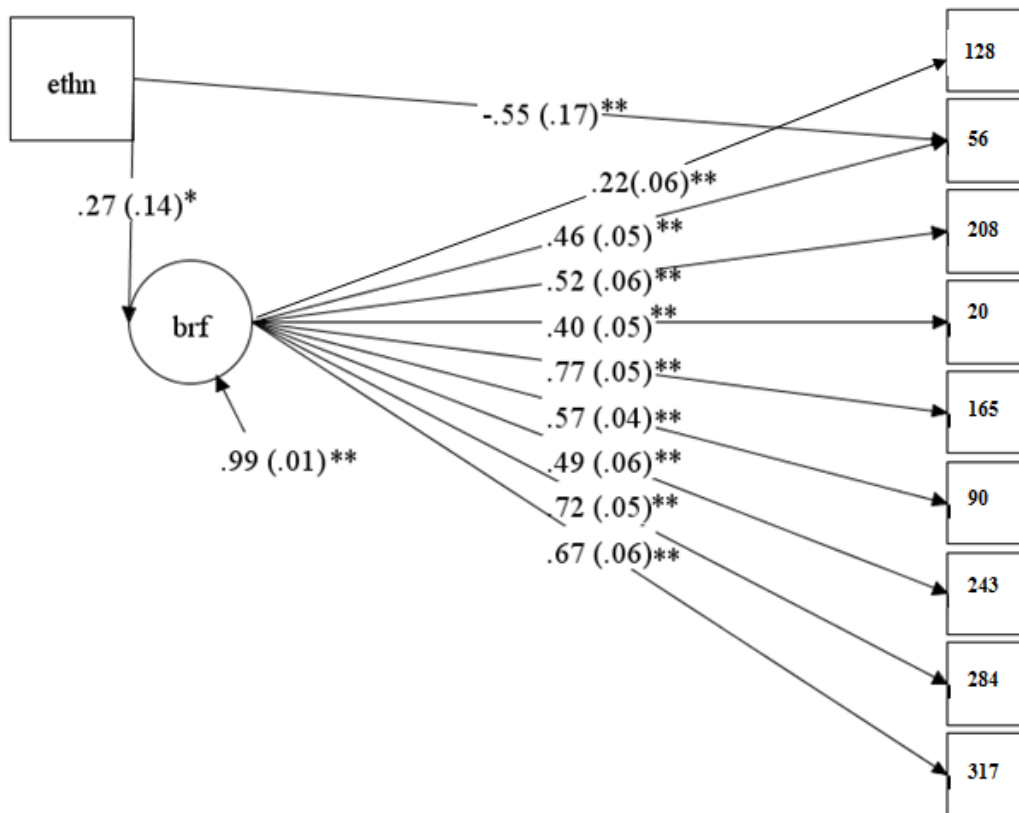


Figure 16. The MIMIC model for the Behavior Restricting Fears (BRF) Scale in the inpatient sample. All estimates are partially standardized and standard errors are in parenthesis following the estimates. Ethn refers to Ethnicity. Brf refers to Behavior Restricting Fears. *Estimates significant at 0.05 level; **Estimates significant at 0.01 level.

Multiple Specific Fears (MSF).

Pearson sample. The factor loadings, thresholds, and model fit indices of the baseline CFA model for the MSF scale for both samples are presented in Table 16.

For the Pearson sample of African American and Caucasian men, factor loadings were

high and varied from 0.43 to 0.64. Item thresholds for the sample ranged from 0.41 to 1.93. The model fit indices for the baseline CFA indicate a good model fit in this sample. Residuals for items 54 and 151 were allowed to correlate based on a modification index of 30.21 and similarity of item content. The standard estimated value for the residual covariance of items 54 and 151 was 0.41 ($p < .01$). The model fit indices improved overall with the addition of these correlated indicator error terms.

Table 16

Multiple Specific Fears (MSF) Baseline CFA Factor Loadings, Thresholds, and Model Fit Indices for the Pearson and Inpatient Sample

Item	Pearson Sample ($N = 2,962$)		Inpatient Sample ($N = 1,376$)	
	Factor Loading	Threshold	Factor Loading	Threshold
82	0.64	0.54	0.65	0.28
115	0.62	0.77	0.64	0.17
184	0.57	0.69	0.60	0.32
220	0.58	0.22	0.60	0.24
286	0.55	1.20	0.54	1.03
54	0.61	1.39	0.73	0.74
151	0.58	1.93	0.66	1.14
258	0.46	0.67	0.63	0.33
320	0.61	0.41	0.55	0.17

Table 16 Continued

	Pearson Sample ($N = 2,962$)	Inpatient Sample ($N = 1,376$)
CFI	0.97	0.95
TLI	0.96	0.93
RMSEA	0.03	0.06
χ^2	119.92 ($p < .01$)	165.15 ($p < .01$)

Note. CFI = Comparative Fit Index. TLI = Tucker-Lewis Index. RMSEA = Root Mean Square Error of Approximation.

In this sample, Caucasian men scored 0.19 standard scores higher on the latent variable of Multiple Specific Fears than African American men ($\beta = -0.19$, $SE = 0.03$, $p < 0.01$). After inclusion of the ethnicity covariate in the model, RMSEA did not change but CFI and TLI decreased from 0.96 to 0.95 and 0.97 to 0.93, respectively. When paths were freely estimated between ethnicity and the indicators, four items demonstrated differential functioning. Controlling for level of Multiple Restricting Fears, African American men had a higher probability of endorsing items 82 ($\beta = 0.10$, $SE = 0.03$, $p < 0.01$) and 184 ($\beta = 0.19$, $SE = 0.03$, $p < 0.01$) when compared to Caucasian men in the Pearson sample. Item 82 relates to fear of snakes and item 184 relates to fear of water.

Also holding level of Multiple Restricting Fears constant, African American men had a higher probability of endorsing items 220 ($\beta = 0.22$, $SE = 0.03$, $p < 0.01$) and 320 ($\beta = 0.13$, $SE = 0.03$, $p < 0.01$) when compared to Caucasian men in the

Pearson sample. Item 220 assesses fear of spiders and item 320 assesses anxiety related to particular animals. No further evidence of differential item functioning was evident. The final MSF model for this sample, including areas of differential item functioning, is visually depicted in Figure 17.

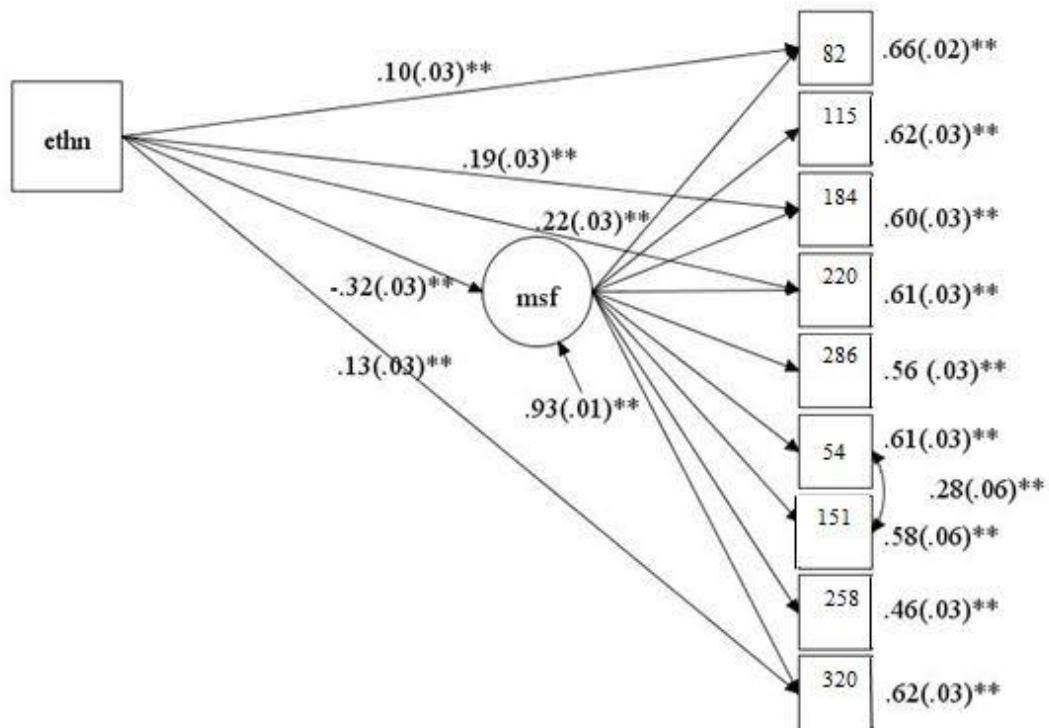


Figure 17. The MIMIC model for the Multiple Specific Fears (MSF) Scale in the Pearson sample. All estimates are partially standardized and standard errors are in parenthesis following the estimates. Estimates for paths between the latent variable and indicators are presented to the right of the indicator for ease of reading. The estimate for the error covariance is presented slightly more to the left of the items for

Figure 17 Continued

differentiation. Ethn refers to Ethnicity. Msf refers to Multiple Specific Fears.

**Estimates significant at 0.01 level.

Inpatient sample. The inpatient sample's factor loadings, thresholds, and model fit indices of the baseline CFA model for the MSF scale are presented on the left side of Table 16. For inpatient men, factor loadings of the nine MSF items ranged from 0.54 to 0.73. Item thresholds for the sample ranged from 0.17 to 1.14. The model fit indices for the baseline CFA, at the bottom right side of Table 16, indicate a decent model fit in this sample. The value for TLI is lower and the value of RMSEA is higher than desired to indicate excellent model fit but as previously mentioned, these cutoffs are guidelines and may need leniency when using categorical indicators (Ketterer, 2011).

Residuals of items 54 and 151 were allowed to covary based on a modification index of 22.03 and similarity of item content. The standard estimated value for the residual covariance of items 54 and 151 was 0.20 ($p < .01$). Overall, the model fit indices did not improve with the addition of the correlated error terms. African Americans men scored 0.40 standard scores higher on the latent variable of Multiple Specific Fears than Caucasian men ($\beta = 0.40$, $SE = 0.11$, $p < 0.01$). After the addition of the ethnicity covariate in the model, CFI and TLI decreased to 0.93 and 0.92, respectively, while RMSEA remained the same.

When paths were freely estimated between ethnicity and the indicators, three items demonstrated differential functioning. Holding level of Multiple Specific Fears constant, African American men had a higher probability of endorsing items 82 ($\beta = 0.34$, $SE = 0.11$, $p < 0.01$), 286 ($\beta = 0.48$, $SE = 0.12$, $p < 0.01$), and 320 ($\beta = 0.46$, $SE = 0.11$, $p < 0.01$) when compared to Caucasian men in the inpatient sample. Item 82 assesses fears of snakes, item 286 asks about fears of mice, and item 320 assesses anxiety about particular animals. No further evidence of differential item functioning was present. Partially standardized estimates of the final MSF model for inpatient men can be found in Figure 18.

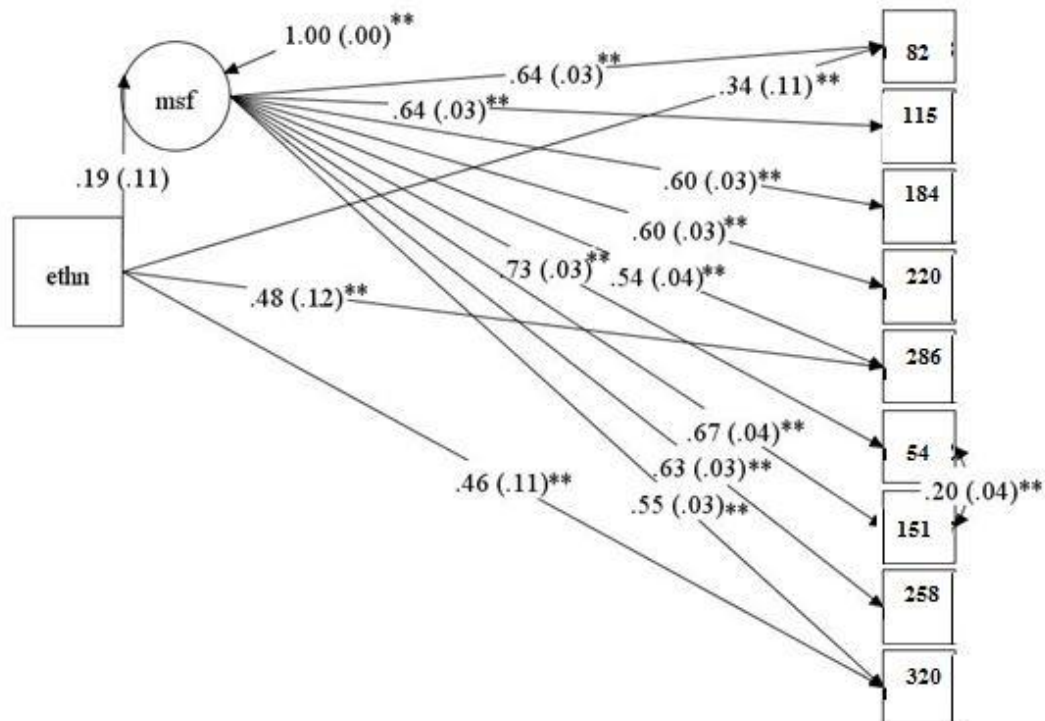


Figure 18. The MIMIC model for the Multiple Specific Fears (MSF) Scale in the inpatient sample. All estimates are partially standardized and standard errors are in

Figure 18 Continued

parenthesis following the estimates. Ethn refers to Ethnicity. Msf refers to Multiple

Specific Fears. **Estimates significant at 0.01 level.

CHAPTER VI

SUMMARY AND DISCUSSION

Overall, when simply examining the differences in raw mean scores between African American and Caucasian men in both samples, the largest differences (per Cohen's d effect size) indicate higher mean scores of African American men. Specifically, African American men in both samples demonstrated higher mean scores on BRF and SUI when compared to Caucasian men, respectively. Both of these differences demonstrated small effect sizes. African American men in both samples had higher mean scores on MSF than Caucasian men, with medium and small effect sizes, respectively.

However, as previously mentioned, it is important to examine the measurement invariance of these scales prior to making decisions about relevant test bias. This study was able to examine the measurement invariance of the Internalizing SP Scales across African American and Caucasian men in both an amalgamated and inpatient sample. Research emphasizes the replicability of any findings and as such, the ability of the current study to investigate measurement invariance in both samples is important. While all findings are reported, the current discussion focuses on findings that were consistent and replicated in both samples.

Baseline Models

Testing measurement invariance via MIMIC modeling begins with an examination of a baseline CFA model, ensuring the appropriateness of the model before the addition of covariates or direct paths between the covariate and items/item thresholds. If the baseline CFA model does not provide a good fit, no further analysis

is warranted. In a total of 18 one factor baseline CFA models, one for each scale in each sample, the majority demonstrated excellent fit. In fact, only four scales, all in the psychiatric inpatient sample, demonstrated less than excellent fit. Even in these cases, all but one of the scales only showed moderate fit in one of the three goodness-of-fit indices examined. For each of these scales, after examining the factor loadings and permitting less stringent cut-off values for the goodness-of-fit indices (based upon recommendations of Kenny and McCoach, 2003 and Ketterer, 2011), the model fit was determined to be good and analysis of measurement invariance continued. One-factor solutions were used because the Internalizing SP Scales were built on the proposition that they are unidimensional scales assessing specific areas not directly or saliently assessed by the RC scales. This proposition of unidimensionality of the Internalizing SP scales was upheld in collapsed samples of African American and Caucasian men in both an amalgamated and inpatient sample. Apart from a single item on the BRF scale, all of the factor loadings of items on all nine Internalizing SP scales were high and considered salient (equal to or higher than .30; Brown, 2015). The factor loading for item 128 on the BRF scale in the inpatient sample was lower than any of the other items at .22.

MIMIC Models

MIMIC modeling involves the addition of a direct path between a covariate and latent factor as a means of assessing for latent mean group differences. In the Pearson sample, African American men had higher latent group means on BRF and NFC scales and lower latent means on MSF and SFD when compared to Caucasian men. As mentioned, one of the strengths of this study is the ability to compare results

in two samples of African American and Caucasian men. In this case, only one of the aforementioned latent mean differences was replicated in the inpatient sample, differences on the MSF scale. However, interestingly the latent mean group difference for the inpatient sample was in the opposite direction of findings in the Pearson sample; inpatient African American men had higher latent mean scores than inpatient Caucasian men on the MSF scale (in comparison to lower latent mean scores compared to Caucasian men in the Pearson sample). Inpatient African American men also demonstrated higher latent mean scores on the SUI scale in comparison to Caucasian men, a finding not seen in the Pearson sample. In both samples, all of the aforementioned latent mean differences between African Americans and Caucasians were .44 or less, pointing to small to medium effect sizes.

With regards to the latent mean differences found on the MSF Scale in both the Pearson and inpatient samples, previous research has found African Americans to score higher on the raw MSF scale when compared to Caucasians (McBride, 2013). It is important to note that the MSF Scale consists of nine items assessing for different fears, including fears of natural elements/weather, animals broadly, and specific animals. As such, this latent mean difference may be related to overall fears and/or specific fears. Previous research has noted cultural differences in endorsement of fears and higher amounts of specific phobias in African Americans compared to Caucasians, particularly related to natural environment, animals, and social phobia (Chapman, Kertz, Zurlage, & Woodruff-Borden, 2008; Chapman, Vines, & Petrie, 2010). While the opposite direction of latent mean MSF scores when comparing African American and Caucasian men in Pearson and inpatient samples is curious and

not well explained by previous research, the overall effect sizes are small to medium in both samples.

Differential Item Functioning

In addition to a path between the covariates and factors, direct paths between covariates and items are included in the model to evaluate differential item functioning (DIF). This allows for an assessment of differential item functioning, or different probability of endorsing (or correctly answering) an item based on group membership, holding the level or performance on the latent variable constant. Items were found to function differently in African American and Caucasian men in the Pearson sample in seven of the nine Internalizing SP scales, with SUI and SFD not showing evidence of DIF. By the same token, DIF was seen in seven of the nine Internalizing SP scales in the inpatient sample, save HLP and SFD scales. Across both the Pearson and inpatient samples, the SFD scale did not evidence any DIF in African American and Caucasian men. Despite evidence of DIF in the same scales across samples, the particular items demonstrating differential functioning were only replicated in the inpatient sample for the NFC and MSF scales.

In the NFC scale, African American men in both samples had a higher probability of endorsing items 27 and 68 than Caucasian men in both samples. The effect sizes seen for these items varied between small and medium, with the largest effect size for item 68 in the Pearson sample. Also, items 82 and 320 on the MSF scale demonstrated a higher probability of being endorsed by African American men in both samples when compared to Caucasian men in both samples, respectively.

Again, effect sizes varied between small and medium with generally small effects in the Pearson sample for both items.

The NFC items that demonstrated differential functioning in both the inpatient and outpatient samples assessed feeling of inefficacy via cognitive roadblocks to completing or initiating tasks. Some of the items on the NFC scale also assess for thinking before acting, but only items 27 and 68 focus on thinking before acting without specific mention of making a decision. The NFC Scale, which measures a test taker's beliefs that he/she is not capable of making decision and dealing with crises, assesses a broad construct of inefficacy or the lack of self-efficacy.

To further explore the cases of DIF seen in both samples, item endorsement probabilities were calculated for item 27 and 68 for African American and Caucasian test takers in the Pearson and inpatient samples. For reference, item 27 had a larger effect in the inpatient sample ($\beta = 0.46$) compared to the Pearson sample ($\beta = 0.32$). On the other hand, item 68 had a larger effect in the Pearson sample ($\beta = 0.60$) compared to the inpatient sample ($\beta = 0.32$). Equation 1 was used to calculate item endorsement probabilities given the factor η_i and covariate x_i , where F is the normal distribution function, τ_j the item threshold, λ_j the unstandardized factor loading of the item, κ_j the unstandardized direct effect of the item on the covariate, and θ the residual variance (Muthén & Muthén, 2009b):

$$P(u_{ij} = 1 | \eta_{ij}, x_i) = 1 - F([\tau_j - \lambda_j \eta_i - \kappa_j x_i] \theta_{jj}^{-\frac{1}{2}}) \quad (1)$$

At the mean of the latent variable of Inefficacy, the probability of endorsing item 27 was 0.38 for African American men in the Pearson sample and 0.21 for Caucasian men in the Pearson sample. Likewise, the probability of endorsing item 27 was 0.75 for African American men in the inpatient sample and 0.53 for Caucasian men in the inpatient sample at the mean of the latent variable of Inefficacy. For item 68, the probability of item endorsement was 0.58 for African American men and 0.31 for Caucasian men in the Pearson sample with latent Inefficacy at its mean. At the mean of latent Inefficacy, African American men in the inpatient sample had a 0.51 probability of endorsing item 68 and Caucasian men had a 0.43 probability of endorsing item 68. A review of these probabilities points to the greatest item endorsement difference between African American and Caucasian men to be on item 27 in the inpatient sample. However, as previously noted, the differential functioning of item 27 in the inpatient sample pointed to a small to medium effect.

Some theories posit that ethnic minorities are likely to have lower self-efficacy based on less access to positive influences, such as history of positive performance, role models, and encouragement (Lent, Brown, & Hackett, 1994, 1996). However, research reviewing over 100 articles looking at motivation in African Americans concluded that results are mixed (Graham, 1994). Within motivation, the study looked at need for achievement, locus of control, and expectancy for future success/self-concept of ability, with the latter being similar to self-efficacy. The study concluded that these aspects of motivation, or self-efficacy, do not appear to be consistently related to ethnicity (Graham, 1994; DeFreitas, 2012).

The two items on the MSF scale that demonstrated differential functioning in African American and Caucasian men in both samples involved animal fears specifically. Of interest, there are a total of four animal-related specific fear items on the MSF scale but the two that functioned differently in the aforementioned samples ask about animals broadly and snakes.

It is important to keep in mind that these items functioned differently in both samples of African American and Caucasian men, holding level of multiple specific fears constant (i.e., despite group latent mean differences on MSF). Thus, research pointing to higher rates of animal-specific fears in African Americans compared to Caucasians is particularly important in this context (Chapman et al., 2008; Chapman et al., 2010). It is unclear why DIF is only seen in only two of the four animal-fear specific items on the MSF scale, but it may be related to these differences being small to medium effects.

Again, item endorsement probabilities for the two differentially functioning items across samples were calculated using equation 1. For reference, item 82 had a larger effect in the inpatient sample ($\beta = 0.35$) compared to the Pearson sample ($\beta = 0.10$). Item 320 also had a larger effect in the inpatient sample ($\beta = 0.46$) compared to the Pearson sample ($\beta = 0.13$). At the mean of the latent variable of Multiple Specific Fears, the probability of endorsing item 82 was 0.60 for African American men in the Pearson sample and 0.46 for Caucasian men in the Pearson sample. Likewise, the probability of endorsing item 82 was 0.51 for African American men in the inpatient sample and 0.34 for Caucasian men in the inpatient sample at the mean of the latent variable of Multiple Specific Fears. For item 320, the probability of item endorsement

was 0.56 for African American men and 0.49 for Caucasian men in the Pearson sample with latent Multiple Specific Fears at its mean.

Again at the mean of latent Multiple Specific Fears, African American men in the inpatient sample had a 0.61 probability of endorsing item 320 and Caucasian men had a 0.39 probability of endorsing item 68. The largest difference in the probability of endorsing an item between African American and Caucasian test takers was seen on item 320 in the inpatient sample. Again, similar to the findings for the NFC scale and previously noted, the differential functioning of item 320 was a small to medium effect.

Implications of Findings

Measurement invariance, specific to this study the assessment of group latent mean differences and differential item functioning, is important to the field of psychological testing, clinical psychology, and psychology in general in its relationship to fairness in testing. In fact, measurement bias, which is heavily related to measurement invariance, is at the heart of the fairness in testing discussion (American Educational Research Association, American Psychological Association, National Council on Measurement in Education [AERA/ APA/ NCME], 2014). Measurement bias, which can be demonstrated via tests of measurement invariance, can lead to inequity in testing. The *Standards for Educational and Psychological Testing* speaks of fairness in terms of accessibility, the opportunity for test takers to accurately depict their level/answer/response on a construct without the influence of construct-irrelevant characteristics. Group differences on latent means, specifically one group scoring higher on a latent variable of a construct measured by a test, points

to areas of needed research to investigate potential measurement invariance, different levels/rates of psychopathology, or other reasons for the difference. If areas of measurement invariance are found and other reasons for score differences ruled out, research may point to potential areas of inequity in testing. Moreover, items demonstrating differential functioning based on ethnicity point to the potential impact of construct-irrelevant characteristics influencing endorsement of items related to a construct, or lack of accessibility based on ethnicity.

Both group differences on latent means and DIF do not necessarily equate to measurement bias. Group differences on a latent mean may point to different meanings of the construct across groups and/or cultural differences in the experience of the construct, rather than differences in the way the construct is measured by the test. On the same note, DIF may be related to different cultural meanings, experiences, or ways of perceiving a particular aspect of a construct (in this case measured by a test item) rather than differences in the way an item measures the construct (AERA/ APA/ NCME, 2014). The determination of measurement bias should be based on review of research indicating whether latent mean differences of DIF may be expected given known cultural differences.

The current study's findings of potential measurement invariance in both the amalgamated and inpatient samples of African American and Caucasian men involved group latent mean differences in one scale and DIF in two scales. Group latent mean differences on the MSF scale were seen in both samples but fell in opposite directions, with African American men in the Pearson sample scoring lower and inpatient African American men scoring higher than Caucasian men in the Pearson and inpatient

samples, respectively. While research points to higher levels of specific fears and phobias in African American samples (Chapman et al., 2008; Chapman et al., 2010), research does not explain these mixed results. The higher latent mean MSF scores of African American men in the inpatient sample is supported by research and thus not considered evidence of measurement bias. The lower latent mean MSF scores of African American men in the Pearson sample does not appear to be supported by research but also was not able to be replicated in the inpatient sample. This finding may be related the nature of the sample in that it is an amalgamated sample of African American men from unknown settings. As such, this finding may be more of a product of the broad sample and further research needs to focus on replication in samples from known settings to assess for possible measurement bias.

Items that demonstrated differential functioning in the NFC scale in both samples assessed inefficacy that arises from feelings that forethought led to inaction or difficulty completing the task. Other items on the NFC scale also assessed inefficacy but tended to focus on crises or decisions. Research and theories on self-efficacy in general and related specifically to ethnic differences are vast and beyond the scope of the current study. Nonetheless, a review of over 100 studies investigating self-efficacy in African Americans concluded that no consistent findings exist (Graham, 1994). Thus, since multicultural research does not appear to help explain differences in self-efficacy, or specifically feelings of inefficacy based on forethought leading to inaction or difficulty completing a task, this area of DIF (items 27 and 68) may represent a specific area of measurement bias on the NFC scale. Again however, it is

important to note that the differences found were small to medium effects and most of the item endorsement probability differences were small.

Items that proved to function differently in both Pearson and inpatient samples of African American and Caucasian men from the MSF scale related to animal-specific fears. While it is curious that only two of the four animal-specific items consistently demonstrated DIF in both samples, cross-cultural research supports the presence of higher rates of animal-specific fears in African American populations (Chapman et al., 2008; Chapman et al., 2010). Thus, this area of DIF is not related to measurement bias but rather underlying traits in the population.

Solutions for measurement non-invariance and DIF. When items are found to be non-invariant based on construct-irrelevant characteristics, several resolution options are available. One option is to delete the non-invariant item (Sass, 2011). Deletion works best with long measures because removal of the item will not greatly impact the measure's psychometric properties. However, this is problematic for widely used scales like the MMPI family of assessments because deletion of the item may make the test no longer comparable to previous versions. Another option is to model the non-invariance into test scoring (Woods, Oltmanns, & Turkheimer, 2009). To accomplish this, the non-invariant items would be estimated separately in groups of interest while invariant items would be estimated the same in both groups. Scores on the test would then be computed from this model that accounts for noninvariance based on group membership. Finally, another means of handling non-invariant items is to assume differences are small and do not influence results greatly (Sass, 2011).

The latter strategy is best for longer measures with only a minority of items demonstrating a small degree of noninvariance.

Of the areas of potential measurement bias, the only measurement invariance that was found in both samples and not adequately explained by previous multicultural research is differential functioning for two items on the nine item NFC Scale. Given the aforementioned strategies to address DIF, the latter solution appears to be the most reasonable. First, the NFC scale is one of the longer Internalizing SP scales with nine items. The NFC scale is also one of nine Internalizing SP scales and one of 23 SP scales (cognitive, internalizing, and externalizing). Also, recall that the SP scales were created as a means to assess areas of psychopathology not directly or saliently measured by the RC Scales. As such, feelings of inefficacy are likely touched upon, albeit not as thoroughly, on one of the RC Scales.

Second, DIF was consistently found in only two of the possible nine items of this scale (i.e., a minority of items). Finally, the probability of African American men endorsing item 27 varied from less than chance to more than chance between the two samples (0.38 in the Pearson sample and 0.75 in the psychiatric inpatient sample), pointing to different manifestations of DIF in the two samples. By the same token, the probability of African American men endorsing item 68 did not vary from chance and was similar in both samples (0.58 in the Pearson sample and 0.51 in the inpatient sample), pointing to a small degree of DIF.

In general, the idea of modifying the MMPI-2-RF based on these results is extremely premature. First, the MMPI-2-RF is used in many settings and contexts and thus it would be necessary to conduct a large number of studies in different settings

and contexts to determine the replicability of the current findings. Second, it is unlikely that the current findings of differential item functioning on one scale with small to medium effect sizes in terms of direct paths would greatly affect clinical interpretation. Any modification to the MMPI-2-RF, whether deletion of items or modeling the DIF into scoring procedures, should only be considered if strong evidence of replicable DIF is found in a large number of studies across various settings and contexts.

Study Limitations and Strengths

The primary limitation of the current study is its exploratory nature. Based on the lack of previous research in the field of measurement invariance and mixed results of previous measurement bias research for the MMPI, hypotheses regarding specific areas of measurement invariance could not be made. However, the ability to explore measurement invariance in the nine Internalizing SP scales in two samples provides a much needed contribution to test bias research on the MMPI-2-RF.

Further, applied examples of measurement invariance with dichotomous items, broadly and specifically using MIMIC modeling, are lacking in the literature. Although MIMIC modeling has been shown to be an appropriate means of assessing for measurement invariance (Brown, 2006; Kim, Yoon, & Lee, 2012; Sass, 2011), real world examples with dichotomous variables are sparse (Woods et al., 2009). As such, the current analyses and interpretation were guided by a limited number of sources combined with the author's decisions in consultation with others (Brown, 2006; Brown, 2015; Ketterer, 2011; Kline, 2013; L. Muthén, 2009; L. Muthén & B. Muthén, 2009b; B. Muthén, 2014; B. Muthén, 2015; Woods et al., 2009). Nonetheless, this

study builds upon the new literature of applied uses of MIMIC modeling to examine the presence of measurement invariance in personality tests. More specifically, this study provides a much needed applied example of MIMIC modeling with dichotomous variables. Broadly, this study helps advance the test bias research within the MMPI family of assessments.

With regards to the chosen analysis, MIMIC modeling has some limitations. For one, MIMIC modeling only assesses for equal latent means and indicator/threshold intercepts and assumes invariance in all other model parameters, including factor loadings, error variances-covariances, and factor variances-covariances (Brown, 2015). Thus, the current analysis assumed equal factor loadings and error and factor variances/covariances across groups. This is clearly a limitation of the current study but the analysis that would have allowed for more detailed invariance testing, multiple-groups CFA, required large sample sizes in all groups because individual CFAs are conducted for the groups separately. Based on the smaller sample size of African American (compared to Caucasian) men in both the samples, MIMIC modeling was the logical choice to adhere to sample size requirements. Despite this limitation, this study provides a first step in assessing for aspects of measurement invariance of the MMPI-2-RF, or any version of the MMPI, in African American and Caucasian samples.

Specific to the design of the study, aspects of the implications are clearly limited. First, the study sought to explore measurement invariance in clinical populations of African American and Caucasian men and women. However, again small sample sizes led the researcher to exclude women from the analysis. While the

research initially aimed to assess clinical samples, the data that were available led to a broader sample in the Pearson data. The Pearson sample protocols came from unknown settings and in general very little information was known about the test takers. The lack of information about the Pearson sample provides a large limitation in interpreting the results from that sample.

Also, this study is obviously limited to comparisons between African American and Caucasian men and did not explore comparisons between other ethnic groups. Finally, the researcher lacked the ability to include other covariates or control for other confounding or contributing variables in the analysis, including but not limited to education level, income, and socioeconomic status (United States Department of Health and Human Services, 2001). It was not possible to include these in the analysis because the archival data used in the study did not contain such information.

Future Directions

The current study provides some bases for future measurement invariance research on the MMPI-2-RF with African American and Caucasian men. First, it provides hypotheses about possible measurement invariance in the Internalizing SP scales with this population. Specifically, research could attempt to replicate findings that were only seen in one of the two samples. Of particular interest, further exploration of group latent mean differences between African Americans and Caucasians on the MSF Scale is needed given the mixed results in the current study. Moreover, it would be interesting to see if the consistent DIF on the MSF Scale found in this study could be replicated in similar samples. Finally, it is important that the

main finding of DIF in two items of the NFC scale be further explored for replicability and degree of differential functioning in various settings, context, and with samples from known settings.

Broadly, future research should address some of the study design limitations of the current study, including gender, setting, ethnicity, and covariate/confounding variable issues. Expanding the current study to include clinical samples of African American and Caucasian women, various ethnicities, and important demographic data will be useful in the generalizability of findings. Also exploring the current results in non-clinical samples would be needed prior to any conclusions about measurement bias. Finally, this study has examined the Internalizing SP scales of the MMPI-2-RF and future research should continue assessing for measurement bias via measurement invariance testing in other MMPI-2-RF scales.

It is also important to note the pitfalls of the comparative nature of the current research and much research on test bias and multi-cultural issues in general. Comparative research in a multi-cultural context occurs when a minority group is compared to Caucasians on any characteristic or construct. Such comparisons can lead to one group being seen as the normative group and the other connoted to be the deviant group. With such comparative research, within-group differences are also ignored. Thus, it will likely also be helpful to examine the MMPI-2-RF scales within cultures as well as between cultures.

REFERENCES

- American Educational Research Association, American Psychological Association, National Council on Measurement in Education [AERA/ APA/ NCME]. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.
- Arbisi, P.A., & Ben-Porath, Y. S. (1995). An MMPI-2 infrequent response scale for use with psychopathological populations: The Infrequency-Psychopathology scale (F_p). *Psychological Assessment*, 7, 424-431.
- Arbisi, P. A., Ben-Porath, Y. S., & McNulty, J. (2002). A comparison of MMPI-2 validity in African-American and Caucasian psychiatric inpatients. *Psychological Assessment*, 14(1), 3-15.
- Archer, R. P., Griffin, R., & Aiduk, R. (1995). MMPI-2 clinical correlates for ten common codes. *Journal of Personality Assessment*, 65, 391-407.
- Arthur, G. (1944). An experience in examining an Indian twelfth-grade group with the Multiphasic Personality Inventory. *Mental Hygiene*, 28, 243-250.
- Ball, J. C. (1960). Comparison of MMPI profile differences among Negro-white adolescents. *Journal of Clinical Psychology*, 16, 1960, 304-307. doi: 10.1002/1097-4679(196007)16:3<304::AID-JCLP2270160323>3.0.CO;2-B
- Ben-Porath, Y. S. (2012). *Interpreting the MMPI-2-RF*. Minneapolis, MN: University of Minnesota Press.

- Ben-Porath, Y. S., & Butcher, J. N. (1989). Psychometric stability of rewritten MMPI items. *Journal of Personality Assessment*, *53*(4), 645-653. doi: 10.1207/s15327752jpa5304_1
- 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., & Sherwood, N. E. (1993). *The MMPI-2 Content Component Scales: Development, psychometric characteristics, and clinical application*. Minneapolis, MN: University of Minnesota Press.
- Ben-Porath, Y. S., Shondrick, D. D., & Stafford, K. P. (1995). MMPI-2 and race in a forensic diagnostic sample. *Criminal Justice and Behavior*, *22*(1), 19-32. doi: 10.1177/0093854895022001002
- Ben-Porath, Y. S., & Tellegen, A. (1995). How (not) to evaluate the comparability of MMPI and MMPI-2 profile configurations: A reply to Humphrey and Dahlstrom. *Journal of Personality Assessment*, *65*(1), 52-58. doi: 10.1207/s15327752jpa6501_4
- 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.
- Bertelson, A. D., Marks, P. A., & May, G. D. (1982). MMPI and race: A controlled study. *Journal of Consulting and Clinical Psychology*, *50*(2), 316-318. doi: 10.1037/0022-006X.50.2.316
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: Wiley.

- Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.
- Brown, T. A. (2015) *Confirmatory factor analysis for applied research (2nd ed.)*. New York, NY: Guilford Press.
- Butcher, J., Ball, B., & Ray, E. (1964). Effects of socio-economic level on MMPI differences in Negro-white college students. *Journal of Counseling Psychology, 11*(1), 83-87. doi: 10.1037/h0046922
- Butcher, J. N., Braswell, L., & Raney, D. (1983). A cross-cultural comparison of American Indian, Black, and White inpatients on the MMPI and presenting symptoms. *Journal of Consulting and Clinical Psychology, 51*(4), 587-594. doi: 10.1037/0022-006X.51.4.587
- 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., Graham, J. R., Williams, C. L., & Ben-Porath, Y. S. (1990). *Development and use of the MMPI-2 Content Scales*. Minneapolis, MN: University of Minnesota Press.

- Butcher, J. N., & Han, K. (1995). Development of an MMPI-2 scale to assess the presentation of self in a superlative manner: The S scale. In J. N. Butcher & C. D. Spielberger (Eds.), *Advances in personality assessment* (Vol. 10, pp. 25-50). Hillsdale, NJ: LEA.
- Byrne, B. M. (2012). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. New York, NY: Routledge.
- Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: the issue of partial measurement invariance. *Psychological Bulletin*, *105*, 456-466.
- Caldwell, M. G. (1953). The youthful male offender in Alabama: A study in delinquency causation. *Sociology and Social Research*, *37*, 236-243.
- Camara, W. J., Nathan, J. S., & Puente, A. E. (2000). Psychological test usage: Implications in professional psychology. *Professional Psychology: Research and Practice*, *31*(2), 141-154. doi: 10.1037/0735-7028.31.2.141
- Campos, L. P. (1989). Adverse impact, unfairness, and bias in the psychological screening of Hispanic peace officers. *Hispanic Journal of Behavioral Sciences*, *11*(2), 122-135. doi: 10.1177/07399863890112002
- Canul, G. D., & Cross, H. J. (1994). The influence of acculturation and racial identity attitudes on Mexican-Americans' MMPI-2 performance. *Journal of Clinical Psychology*, *50*(5), 736-745. doi: 10.1002/1097-4679(199409)50:5<736::AID-JCLP2270500511>3.0.CO;2-Z
- Carle, A. C., Millsap, R. E., & Cole, D. A. (2008). Measurement bias across gender on the Children's Depression Inventory: Evidence for invariance from two latent

variable models. *Educational and Psychological Measurement*, 68(2), 281-303. doi: 10.1177/0013164407308471

Castro, Y., Gordon, K. H., Brown, J. S., Anestis, J. C., & Joiner, T. E. (2008).

Examination of racial differences on the MMPI-2 Clinical and Restructured Clinical scales in an outpatient sample. *Assessment*, 15(3), 277-286.

Chapman, L. K., Kertz, S. J., Zurlage, M. M., & Woodruff-Borden, J. (2008). A

confirmatory factor analysis of specific phobia domains in African American and Caucasian American young adults. *Journal of Anxiety Disorders*, 22, 763-771.

Chapman, L. K., Vines, L., & Petrie, J. (2011). Fear factors: Cross validation of

specific phobia in a community sample of African American adults. *Journal of Anxiety Disorders*, 25, 539-544.

Cheung, G. W., & Rensvold, R. B. (1998). Testing factorial invariance across groups:

A reconceptualization and proposed new method. *Journal of Management*, 25, 1-27.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for

testing measurement invariance. *Structural Equation Modeling*, 9(2), 233-255.

doi: 10.1207/S15328007SEM0902_5

Cole, N. S. (1981). Bias in testing. *American Psychologist*, 36, 1067-1077.

Cook, G., Pogany, E., & Johnston, N. G. (1974). A comparison of blacks and whites

committed for evaluation of competency to stand trial on criminal charges.

Journal of Psychiatry and Law, 2, 319-337.

- Cook, K. F., Kallen, M. A., & Amtmann, D. (2009). Having a fit: Impact of number of items and distribution of data on traditional criteria for assessing IRT's unidimensionality assumption. *Quality of Life Research, 18*(4), 447–460. doi:10.1007/s11136-009-9464-4
- Costello, R. M., Fine, H. J., & Blau, B. I. (1973). Racial comparisons on the Minnesota Multiphasic Personality Inventory. *Journal of Clinical Psychology, 29*(1), 63-65. doi: 10.1002/1097-4679(197301)29:1<63::AID-JCLP2270290124>3.0.CO;2-S
- Costello, R. M., Tiffany, D. W., & Gier, R. H. (1972). Methodological issues and racial (black-white) comparisons on the MMPI. *Journal of Consulting and Clinical Psychology, 38*(2), 161-168. doi: 10.1037/h0032623
- Culhane, S. E., Morera, O. F., Watson, P. J., & Millsap, R. E. (2011). The Bermond-Vorst Alexithymia Questionnaire: A measurement invariance examination among U.S. Anglos and U.S Hispanics. *Assessment, 18*(1), 88-94. doi: 10.1177/1073191110387509
- Culhane, S. E., Morera, O. F., Watson, P. J., & Millsap, R. E. (2009). Assessing measurement and predictive invariance of the Toronto Alexithymia Scale-20 in U.S. Anglo and U.S. Hispanic student samples. *Journal of Personality Assessment, 91*(4), 387-395. doi: 10.1080/00223890902936264
- Davis, W. E. (1975). Race and the differential "power" of the MMPI. *Journal of Personality Assessment, 39*(2), 138-140. doi: 10.1207/s15327752jpa3902_8
- Davis, W. E., Beck, S. J., & Ryan, T. A. (1973). Race-related and educationally-related MMPI profile differences among hospitalized schizophrenics. *Journal*

of Clinical Psychology, 29(4), 478-479. doi: 10.1002/1097-4679(197310)29:4<478::AID-JCLP2270290423>3.0.CO;2-Z

- Davis, W. E., & Jones, M. H. (1974). Negro vs Caucasian psychological test performance revisited. *Journal of Consulting and Clinical Psychology*, 42(5), 675-679. doi: 10.1037/h0037062
- DeFreitas, S. C. (2012). Differences between African Americans and European Americans first-year college students in the relationship between self-efficacy, outcome expectations, and academic achievement. *Review of Educational Research*, 64(1), 55-117.
- Finney, S. J., & DiStefano, C. (2013). Non-normal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (2nd ed.;pp. 269- 301). Charlotte, NC: Information Age Publishing, Inc.
- Flanagan, J., & Lewis, G. (1969). Comparison of Negro and white lower class men on the General Aptitude Test Battery and the Minnesota Multiphasic Personality Inventory. *Journal of Social Psychology*, 78, 289-291.
- Frueh, B. C., Smith, D. W., & Libet, J. M. (1996). Racial differences on psychological measures in combat veterans seeking treatment for PTSD. *Journal of Personality Assessment*, 66(1), 41-53. doi: 10.1207/s15327752jpa6601_3
- Frueh, B. C., Gold, P. B., de Arellano, M. A., & Brady, K. L. (1997). A racial comparison of combat veterans evaluated for PTSD. *Journal of Personality Assessment*, 68(3), 692-702. doi: 10.1207/s15327752jpa6803_14

- Genthner, R. W., & Graham, J. R. (1976). Effects of short-term public psychiatric hospitalization for both Black and White patients. *Journal of Consulting and Clinical Psychology, 44*(1), 118-124. doi: 10.1037/0022-006X.44.1.118
- Gilberstadt, H., & Duker, J. (1965). *A handbook for clinical and actuarial MMPI interpretation*. Philadelphia, PA: W. B. Saunders.
- Gironda, R. J. (1999). Comparative validity of MMPI-2 scores of African-Americans and Caucasians in a forensic diagnostic sample. *Dissertation Abstracts International: Section B: The Sciences and Engineering, 60*(6-B), 2942.
- Graham, J. R. (2006). *MMPI-2: Assessing personality and psychopathology* (4th ed.). New York, NY: Oxford University Press.
- Graham, J. R., Ben-Porath, Y. S., & McNulty, J. L. (1999). *Using the MMPI-2 in outpatient mental health settings*. Minneapolis, MN: University of Minnesota Press.
- Graham, J. R., Timbrook, R. E., Ben-Porath, Y. S., & Butcher, J. N. (1991). Code-type congruence between MMPI and MMPI-2: Separating fact from artifact. *Journal of Personality Assessment, 57*(2), 205-215. doi: 10.1207/s15327752jpa5702_2
- Graham, S. (1994). Motivation in African Americans. *Review of Educational Research, 64*(1), 55- 117. doi: 10.2307/1170746.
- Greene, R. L. (1987). Ethnicity and MMPI performance: A review. *Journal of Consulting and Clinical Psychology, 55*(4), 497-512. doi: 10.1037/0022-006X.55.4.497

Greene, R. L., & Robin, R. W., Albaugh, B., Caldwell, A., & Goldman, D. (2003).

Use of the MMPI-2 in American Indians: II. Empirical correlates.

Psychological Assessment, 15(3), 360-369. doi: 10.1037/1040-3590.15.3.360

Gynther, M. D., Fowler, R. D., & Erdberg, P. (1971). False positives galore: The application of standard MMPI criteria to a rural, isolated, Negro sample.

Journal of Clinical Psychology, 27(2), 234-237. doi: 10.1002/1097-

4679(197104)27:2<234::AID-JCLP2270270225>3.0.CO;2-2

Hall, G. C. N., Bansal, A., & Lopez, I. R. (1999). Ethnicity and psychopathology: A meta-analytic review of 31 years of comparative MMPI/MMPI-2 research.

Psychological Assessment, 11(2), 186-197.

Handel, R. W., & Ben-Porath, Y. S. (2000). Multicultural assessment with the MMPI-

2: Issues for research and practice. In R. H. Dana (Ed.), *Handbook of cross-cultural and multicultural personality assessment* (pp. 229-245). Mahwah,

New Jersey: Lawrence Erlbaum Associates.

Handel, R. W., Ben-Porath, Y. S., Tellegen, A., & Archer, R. P. (2010). Psychometric

functioning of the MMPI-2-RF VRIN-r and TRIN-r scales with varying degrees of randomness, acquiescence, and counter-acquiescence.

Psychological Assessment, 22, 87-95. doi: 10.1037/a0017061

Harkness, A. R., & McNulty, J. L. (2007). An overview of personality: The MMPI-2

Personality Psychopathology Five (PSY-5). In J. N. Butcher (Ed.), *Pathways to MMPI-2 use: A practitioner's guide to test usage in diverse settings*.

Washington, DC: American Psychological Association.

- 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*, 104-114.
- Harkness, A. R., McNulty, J. L., Ben-Porath, Y. S., & Graham, J. R. (1995). *MMPI-2 Personality Psychopathology Five (PSY-5) scales: Gaining an overview for case conceptualization and treatment planning*. Minneapolis, MN: University of Minnesota Press.
- Harrison, R. H., & Kass, E. H. (1967). Differences between Negro and White pregnant women on the MMPI. *Journal of Consulting Psychology, 31*(5), 454-463. doi: 10.1037/h0025012
- Hathaway, S. R., & McKinley, J. C. (1940). A multiphasic personality schedule (Minnesota): I. Construction of the schedule. *Journal of Psychology, 10*, 249-254.
- Hathaway, S. R., & McKinley, J. C. (1942). A multiphasic personality schedule (Minnesota): III. The measurement of symptomatic depression. *Journal of Psychology, 14*, 73-84.
- Hathaway, S. R., & McKinley, J. C. (1943). *The Minnesota Multiphasic Personality Inventory*. Minneapolis, MN: University of Minnesota Press.
- Herreid, C. F., & Herreid, J. R. (1966). Differences in MMPI scores in native and nonnative Alaskans. *The Journal of Social Psychology, 70*(2), 191-198. doi: 10.1080/00224545.1966.9712415

- Hokanson, J. E., & Calden, G. (1960). Negro-white differences on the MMPI. *Journal of Clinical Psychology, 16*, 32-33. doi: 10.1002/1097-4679(196001)16:1<32::AID-JCLP2270160113>3.0.CO;2-R
- Holcomb, W. R., & Adams, N. (1982). Racial influences on intelligence and personality measures of people who commit murder. *Journal of Clinical Psychology, 38*, 793-796.
- Holland, T. R. (1979). Ethnic group differences in MMPI profile pattern and factorial structure among adult offenders. *Journal of Personality Assessment, 43*(1), 72-77.
- Ingram, J. C., Marchioni, P., Hill, G., Caraveo-Ramos, E., & McNeil, B. (1985). Recidivism perceived problem-solving abilities, MMPI characteristics, and violence: A study of black and white incarcerated male adult offenders. *Journal of Clinical Psychology, 41*, 4, 25-432.
- Jaccard, J., & Wan, C. K. (1996). *LISREL approaches to interaction effects in multiple regression*. Thousand Oaks, CA: Sage.
- Jöreskog, K. G. (1993). Testing structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 294-316). Newbury Park, CA: Sage.
- Kenny, D. A. & McCoach, D. B. (2003). Effect of the number of variables on measures of fit in structural equation modeling. *Structural Equation Modeling, 10*, 333-351.
- Ketterer, H. L. (2011). Examining the measurement invariance of the MMPI-2 Restructured Clinical (RC) scales across Korean and American normative

samples. *Dissertation Abstracts International: Section B: The Sciences and Engineering*, 71(7-B), 4532.

Kim, E. S., & Yoon, M. (2011). Testing measurement invariance: A comparison of multiple-group categorical CFA and IRT. *Structural Equation Modeling*, 18, 212-228.

Kim, E. S., Yoon, M., & Lee, T. (2012). Testing measurement invariance using MIMIC likelihood ratio test with a critical value adjustment. *Educational and Psychological Measurement*, 72(3), 469-492. doi: 10.1177/0013164411427395

King, H. F., Carroll, J. L., & Fuller, G. B. (1977). Comparison of nonpsychiatric Blacks and Whites on the MMPI. *Journal of Clinical Psychology*, 33(3), 725-728. doi: 10.1002/1097-4679(197707)33:3<725::AID-JCLP2270330324>3.0.CO;2-W

Kline, J. A., Rozyko, V. V., Flint, G., & Roberts, A. C. (1973). Personality characteristics of male Native American alcoholic patients. *International Journal of the Addictions*, 8(4), 729-732.

Kline, R. B. (2013). Assessing statistical aspects of test fairness with structural equation modeling. *Educational Research and Evaluation*, 19(2-3), 204-222. doi: 10.1080/13803611.2013.767624

Kwan, K. K. (1999). MMPI and MMPI-2 performance of the Chinese: Cross-cultural applicability. *Professional Psychology: Research and Practice*, 30(3), 260-268. doi: 10.1037/0735-7028.30.3.260

- Lacey, K. (2004). The Minnesota Multiphasic Personality Inventory- revised (mmpi-2): Extending American Indian norms. *Dissertation Abstracts International: Section B: The Sciences and Engineering*, 65(2-B), 1062.
- Lapham, S. C., Skipper, B. J., Owen, J. P., Kleyboecker, K., Teaf, D., Thompson, B., & Simpson, G. (1995). Alcohol abuse screening instruments: Normative test data collected from a first DWI offender screening program. *Journal of Studies on Alcohol*, 56(1), 51-59.
- Lee, H. B., Cheung, P. M., Man, H., & Hsu, S. Y. (1992). Psychological characteristics of Chinese low back pain patients: An exploratory study. *Psychology and Health*, 6, 119-128.
- Lees-Haley, P. R., English, L. T., & Glenn, W. J. (1991). A fake bad scale on the MMPI-2 for personal injury claimants. *Psychological Reports*, 68, 203-210.
- Lent, R. W., Brown, S. D., & Hackett, G. (1994). Toward a unifying social cognitive theory of career and academic interest, choice, and performance. *Journal of Vocational Behavior*, 45, 79-122.
- Lent, R. W., Brown, S. D., & Hackett, G. (1996). Career development from a social cognitive perspective. In D. Brown & L. Brooks (Eds.), *Career Choice and Development* (pp. 373-722). San Francisco: Jossey-Bass Publishers.
- Lessenger, L. H. (1997). Acculturation and MMPI-2 scale scores of Mexican American substance abuse patients. *Psychological Reports*, 80(3), 1181-1182.
doi: 10.2466/pr0.1997.80.3c.1181

- Liske, R., & McCormick, R. (1976). MMPI profiles compared for Black and White hospitalized veterans. *Newsletter for Research in Mental Health & Behavioral Sciences, 18*(1), 30-32.
- MacCallum, R. C. (1986). Specification searches in covariance structure modeling. *Psychological Bulletin, 100*, 107-120.
- Marsella, A. J., Sanborn, K., Kameoka, V., Brennan, J., & Shizuru, L. (1975). Cross-validation of self-report measures of depression in different ethnocultural groups. *Journal of Clinical Psychology, 31*, 281-287.
- 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. doi: 10.1037/a0030610
- Mathur, A., Barak, B., Zhang, Y., & Lee, K. S. (2001). A cross-cultural procedure to assess reliability and measurement invariance. *Journal of Applied Measurement, 2*, 241-255.
- McBride, W. F. (2013). *Examination of racial bias on the MMPI-2 Restructured Form among African Americans and Caucasians* (Master Thesis). Retrieved from Online Theses and Dissertations database. (UMI No. 1426707449).
- McCreary, C., & Padilla, E. (1977). MMPI differences among Black, Mexican-American and White male offenders. In R. Nunez (Ed.), *Pruebas psicometricas de la personalidad* [Psychometric personality tests] (pp. 67-84). Mexico City: Trillas.

- McDonald, R. L., & Gynther, M. D. (1962). MMPI norms for southern adolescent Negroes. *The Journal of Social Psychology, 58*(2), 277-282. doi: 10.1080/00224545.1962.9712377
- McDonald, R. L., & Gynther, M. D. (1963). MMPI differences associated with sex, race, and class in two adolescent samples. *Journal of Consulting Psychology, 27*(2), 112-116. doi: 10.1037/h0048549
- McGill, J. C. (1980). MMPI score differences among Anglo, Black, and Mexican-American welfare recipients. *Journal of Clinical Psychology, 36*(1), 147-151. doi: 10.1002/1097-4679(198001)36:1<147::AID-JCLP2270360114>3.0.CO;2-A
- McKinley, J. C., & Hathaway, S. R. (1940). A multiphasic personality schedule (Minnesota): II. A differential study of hypochondriasis. *Journal of Psychology, 10*, 255-268.
- McKinley, J. C., & Hathaway, S. R. (1942). A multiphasic personality schedule (Minnesota): IV. Psychasthenia. *Journal of Applied Psychology, 26*, 614-624.
- McKinley, J. C., & Hathaway, S. R. (1944). A multiphasic personality schedule (Minnesota): V. Hysteria, Hypomania, and Psychopathic Deviate. *Journal of Applied Psychology, 28*, 153-174.
- McNulty, J. L., Graham, J. R., Ben-Porath, Y. S., & Stein, L. A. R. (1997). Comparative validity of MMPI-2 scores of African American and Caucasian mental health center clients. *Psychological Assessment, 9*(4), 464-470. doi: 10.1037/1040-3590.9.4.464

- 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
- Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology, 93*(3), 568-592. doi: 10.1037/0021-9010.93.3.568
- Miller, C. Knapp, S. C., & Daniels, C. W. (1968). MMPI study of Negro mental hygiene clinic patients. *Journal of Abnormal Psychology, 73*(2), 168-173. doi: 10.1037/h0025619
- Miller, C., Wertz, C., & Counts, S. (1961). Racial differences on the MMPI. *Journal of Clinical Psychology, 17*(2), 159-161. doi: 10.1002/1097-4679(196104)17:2<159::AID-JCLP2270170216>3.0.CO;2-S
- Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge.
- Millsap, R. E. (2007). Invariance in measurement and prediction revisited. *Psychometrika, 72*(4), 461-473. doi: 10.1007/s11336-007-9039-7
- Millsap, R. E. (1997). Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychological Methods, 2*(3), 248-260.
- Monnot, M. J., Quirk, S. W., Hoerger, M., & Brewer, L. (2009). Racial bias in personality assessment: Using the MMPI-2 to predict psychiatric diagnoses of African American and Caucasian chemical dependency inpatients. *Psychological Assessment, 21*(2), 137-151.

- Moore, C., & Handal, P. J. (1980). Adolescents' MMPI performance, cynicism, estrangement, and personal adjustment as a function of race and sex. *Journal of Clinical Psychology, 36*(4), 932-936. doi: 10.1002/1097-4679(198010)36:4<932::AID-JCLP2270360417>3.0.CO;2-A
- Munley, P. H., Morris, J. R., Murray, D. A., & Baines, T. C. (2001). A comparison of African-American and White American veteran MMPI-2 profiles. *Assessment, 8*(1), 1-10. doi: 10.1177/107319110100800101
- Muthén, B.O. (1998-2004). *Mplus technical appendices*. Los Angeles, CA: Muthén & Muthén.
- Muthén, B. O. (2014) *Mplus discussion: MIMIC models*. Retrieved from <http://www.statmodel.com/cgi-bin/discus/show.cgi?23/48>
- Muthén, B. O. (2015a) *Mplus discussion: Allowing for correlation among error terms in manifest variables in CFA*. Retrieved from <http://www.statmodel.com/cgi-bin/discus/show.cgi?9/351>
- Muthén, L. K. (2009) *Mplus discussion: Measurement equivalent/invariance*. Retrieved from <http://www.statmodel.com/discussion/messages/9/1666.html?1358891945>
- Muthén, B. O., & Asparouhov, T. (2002). Latent variable analysis with categorical outcomes: Multiple group and growth modeling in Mplus. *Mplus Web Notes, 4*(5). Retrieved from <http://www.statmodel.com/download/webnotes/CatMGLong.pdf>
- Muthén, L. K., & Muthén, B. O. (2009a). *Mplus Short Courses, Topic 1: Exploratory factor analysis, confirmatory factor analysis, and structural equation modeling*

for continuous outcomes [PowerPoint slides]. Retrieved from
http://www.statmodel.com/videos/topic1_sm.shtml

Muthén, L. K., & Muthén, B. O. (2009b). *Mplus Short Courses, Topic 2: Regression analysis, exploratory factor analysis, confirmatory factor analysis, and structural equation modeling for categorical, censored, and count outcomes* [PowerPoint slides]. Retrieved from
http://www.statmodel.com/videos/topic2_sm.shtml

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus User's Guide* (7th Ed.). Los Angeles, CA: Muthén & Muthén.

NCS Pearson. (2008-2014). MMPI-2-RF Protocols, African American and Caucasian men. Unpublished raw data.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

Okazaki, S., & Sue, S. (1995). Cultural considerations in psychological assessment of Asian-Americans. In J. N. Butcher (Ed.), *Clinical personality assessment: Practical approaches* (pp.107-119). New York, NY: Oxford University Press.

Pace, T. M., Robbins, R. R., Choney, S. K., Hill, J. S., Lacey, K., & Blair, G. (2006). A cultural-contextual perspective on the validity of the MMPI-2 with American Indians. *Cultural Diversity and Ethnic Minority Psychology, 12*(2), Apr 2006, 320-333. doi: 10.1037/1099-9809.12.2.320

Page, R. D., & Bozlee, S. (1992). A cross-cultural MMPI comparison of alcoholics. *Psychological Reports, 50*, 639-646.

- Patalano, F. (1978). Personality dimensions of drug abusers who enter a drug-free therapeutic community. *Psychological Reports, 42*, 1063-1069.
- Patterson, E. T., Charles, H. L., Woodward, W. A., Roberts, W. R., & Penk, W. E. (1981). Differences in measures of personality and family environment among Black and White alcoholics. *Journal of Consulting and Counseling Psychology, 49*, 1-9.
- Penk, W. E., Roberts, W. R., Robinowitz, R., Dolan, M. P., Atkins, H. G., & Woodward, W. A. (1982). MMPI differences of Black and White male polydrug abusers seeking treatment. *Journal of Consulting and Clinical Psychology, 50*(3), 463-465. doi: 10.1037/0022-006X.50.3.463
- Penk, W. E., Woodward, W. A., Robinowitz, R., & Hess, J. L. (1978). Differences in MMPI scores of black and white compulsive heroin users. *Journal of Abnormal Psychology, 87*, 505-513.
- Prewett, B. M. (2012). Native Americans and the Minnesota Multiphasic Personality Inventory (MMPI-2): Continuing to extend Native American norms. *Dissertation Abstracts International: Section B: The Sciences and Engineering, 72*(10-B), 6395.
- Pritchard, D. A., & Rosenblatt, A. (1980). Racial bias in the MMPI: A methodological review. *Journal of Consulting and Clinical Psychology, 48*, 263-267. doi: 10.1037/0022-006X.48.2.263
- Robin, R. W., Greene, R. L., Albaugh, B., Caldwell, A., & Goldman, D. (2003). Use of the MMPI-2 in American Indians: I. Comparability of the MMPI-2 between

- two tribes and with the MMPI-2 normative group. *Psychological Assessment*, 15(3), 351-359. doi: 10.1037/1040-3590.15.3.351
- Rosenblatt, A. I., & Pritchard, D. A. (1978). Moderators of racial differences on the MMPI. *Journal of Consulting and Clinical Psychology*, 46(6), 1572-1573. doi: 10.1037/0022-006X.46.6.1572
- 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. doi: <http://dx.doi.org/10.1177/0734282911406661>
- Schinka, J. A., & LaLone, L. (1997). MMPI-2 norms: Comparisons with a census-matched subsample. *Psychological Assessment*, 9(3), 307-311. doi: 10.1037/1040-3590.9.3.307
- Schinka, J. A., LaLone, L., & Greene, R. L. (1998). Effects of psychopathology and demographic characteristics on MMPI-2 scale scores. *Journal of Personality Assessment*, 70, 197-211.
- Schumacker, R. E., & Lomax, R. G. (2012). *A beginner's guide to structural equation modeling* (3rd ed.). New York, NY: Routledge.
- Silva, E. S., & MacCallum, R. C. (1988). Some factors affecting the success of specification searches in covariance structure modeling. *Multivariate Behavioral Research*, 23, 297-326.
- Smith, C. P., & Graham, J. R. (1981). Behavioral correlates for the MMPI standard F scale and for a modified F scale for Black and White psychiatric patients.

Journal of Consulting and Clinical Psychology, 49(3), 455-459. doi:
10.1037/0022-006X.49.3.455

- Stanton, J. M. (1956). Group personality profiles related to aspects of antisocial behavior. *Journal of Criminal Law, Criminology and Police Science*, 47, 340-349.
- Stevens, M. J., Kwan, K., & Graybill, D. F. (1993). Comparison of MMPI-2 scores of foreign Chinese and Caucasian-American students. *Journal of Clinical Psychology*, 49(1), 23-27. doi: 10.1002/1097-4679(199301)49:1<23::AID-JCLP2270490104>3.0.CO;2-O
- Sue, S., Keefe, K., Enomoto, K., Durvasula, R. S., & Chao R. (1996). Asian American and White college students' performance on the MMPI-2. In J. N. Butcher (Ed.), *International adaptations of the MMPI-2* (pp. 206-218). Minneapolis, MN: University of Minnesota Press.
- Sue, S., & Sue, D. W. (1974). MMPI comparisons between Asian-American and non-Asian students utilizing a student health psychiatric clinic. *Journal of Counseling Psychology*, 21(5), 423-427. doi: 10.1037/h0037074
- Sutker, P. B., Archer, R. P., & Allain, A. N. (1978). Drug abuse patterns, personality characteristics, and relationships with sex, race, and sensation seeking. *Journal of Consulting and Clinical Psychology*, 46, 1374-1378.
- Sutker, P. B., Archer, R. P., & Allain, A. N. (1980). Psychopathology of drug abusers: Sex and ethnic considerations. *The International Journal of the Addictions*, 15(4), 605-613.

- Telander, C. M. (1999). Asian and Hispanic-American performance on the supplementary scales of the MMPI-2. *Dissertation Abstracts International: Section B: The Sciences and Engineering*, 59(11-B), 6108.
- Tellegen, A. (1985). Structures of mood and personality and their relevance to assessing anxiety, with an emphasis on self-report. In A. H. Tuma & J. D. Maser (Eds.), *Anxiety and the anxiety disorders* (pp. 681-706). Hillsdale, NJ: Erlbaum.
- 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). *The MMPI-2 Restructured Clinical Scales: Development, validation, and interpretation*. Minneapolis, MN: University of Minnesota Press.
- Timbrook, R. E., & Graham, J. R. (1994). Ethnic differences on the MMPI-2? *Psychological Assessment*, 6(3), 212-217. doi: 10.1037/1040-3590.6.3.212
- Tsai, D. C., & Pike, P. L. (2000). Effects of acculturation on the MMPI-2 scores of Asian American students. *Journal of Personality Assessment*, 74(2), 216-230. doi: 10.1207/S15327752JPA7402_4
- Tsushima, W. T., & Onorato, V. A. (1982). Comparison of MMPI scores of White and Japanese-American medical patients. *Journal of Consulting and Clinical Psychology*, 50(1), 150-151. doi: 10.1037/0022-006X.50.1.150

- Tsushima, W. T., & Stoddard, V. M. (1990). Ethnic group similarities in the biofeedback treatment of pain. *Medical Psychotherapy: An International Journal*, 3, 1990, 69-75.
- Tsushima, W. T., & Tsushima, V. G. (2009). Comparison of MMPI-2 validity scales among compensation-seeking Caucasian and Asian American medical patients. *Assessment*, 6, 159-164.
- Uecker, A. E., Boutilier, L. R., & Richardson, E. H. (1980). "Indianism" and MMPI scores of men alcoholics. *Journal of Studies on Alcohol*, 41, 357-362.
- United States Department of Health and Human Services. (2001). *Mental health: Culture, race, and ethnicity – a supplement to mental health: A report of the Surgeon General*. Available from <http://www.ncbi.nlm.nih.gov/books/NBK44249/>
- University of Minnesota Press. (2011). *Available translations*. Retrieved from <http://www.upress.umn.edu/test-division/translations-permissions/permissions>
- Velasquez, R. J., Ayala, G. X., & Mendoza, S. A. (1998). *Psychodiagnostic assessment of U.S. Latinos: MMPI, MMPI-2, and MMPI-A results*. East Lansing, MI: Julian Samora Institute.
- Velasquez, R. J., & Callahan, W. J. (1990a). MMPI comparisons of Hispanic- and white- American veterans seeking treatment for alcoholism. *Psychological Reports*, 67, 95-98.
- Velasquez, R. J., & Callahan, W. J. (1990b). MMPIs of Hispanic, Black, and White DSM-III schizophrenics. *Psychological Reports*, 66(3), 819-822. doi: 10.2466/PRO.66.3.819-822

- Velasquez, R. J., Callahan, W. J., & Carrillo, R. (1989). MMPI profiles of Hispanic-American inpatient and outpatient sex offenders. *Psychological Reports*, *65*(3), 1055-1058. doi: 10.2466/pr0.1989.65.3.1055
- Velasquez, R. J., Callahan, W. J., & Carrillo, R. (1991). MMPI differences among Mexican-American male and female psychiatric inpatients. *Psychological Reports*, *68*(1), 123-127. doi: 10.2466/PR0.68.1.123-127
- Velasquez, R. J., Callahan, W. J., & Young, R. (1993). Hispanic-White MMPI comparisons: Does psychiatric diagnosis make a difference? *Journal of Clinical Psychology*, *49*(4), 528-534. doi: 10.1002/1097-4679(199307)49:4<528::AID-JCLP2270490410>3.0.CO;2-X
- 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*, 4-69.
- Venn, J. (1988). MMPI profiles of Native, Mexican, and Caucasian-American male alcoholics. *Psychological Reports*, *62*, 427-432.
- Waller, N. G., Thompson, J. S., & Wenk, E. (2000). Using IRT to separate measurement bias from true group differences on homogeneous and heterogeneous scales: An illustration with the MMPI. *Psychological Methods*, *5*(1), 125-146.
- Walker, C. M. (2011). What's the DIF? Why differential item functioning analyses are an important part of instrument development and validation. *Journal of Psychoeducational Assessment*, *29*(4), 364-376. doi: 10.1177/0734282911406666

- Walters, G. D., Greene, R. L., Jeffrey, T. B. (1984). Discriminating between alcoholic and nonalcoholic Blacks and Whites on the MMPI. *Journal of Personality Assessment, 48*, 486-488.
- Watson, D., & Tellegen, A. (1985). Toward a consensual structure of mood. *Psychological Bulletin, 98*, 219-235.
- Weiss, R. W., & Russakoff, S. (1977). Relationship of MMPI scores of drug-abusers to personal variables and type of treatment program. *Journal of Psychology, 96*, 25-29.
- Whitworth, R. H., & McBlaine, D. D. (1993). Comparison of the MMPI and MMPI-2 administered to Anglo- and Hispanic-American university students. *Journal of Personality Assessment, 61*(1), 19-27. doi: 10.1207/s15327752jpa6101_2
- Whitworth, R. H., & Unterbrink, C. (1994). Comparison of MMPI-2 clinical and content scales administered to Hispanic and Anglo-Americans. *Hispanic Journal of Behavioral Sciences, 16*(3), 255-264. doi: 10.1177/07399863940163004
- Woods, C. M., Oltmanns, T. F., & Turkheimer, E. (2009). Illustration of MIMIC-model DIF testing with the Schedule for Nonadaptive and Adaptive Personality. *Journal of Psychopathology and Behavioral Assessment, 31*(4), 320- 330.
- Yu, C. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes*. Dissertation retrieved from <http://www.statmodel.com/download/Yudissertation.pdf>

VITA

Megan Anne Brokenbough
700 Park Avenue/MCAR-410
Norfolk, VA 23504

Education

- | | |
|------|---|
| 2015 | Doctorate of Philosophy in Clinical Psychology,
Virginia Consortium Program in Clinical Psychology |
| 2013 | Masters of Art in Clinical and Community Psychology,
Norfolk State University |
| 2009 | Bachelors of Art in Psychology and Social Behavior,
University of California Irvine |
| 2006 | Associates of Art and Associates of Science in General Studies
Riverside Community College |

Clinical Experience

- | | |
|-------------|--|
| 2015 – 2016 | Post-Doctoral Fellow
Mental and Behavioral Health Capacity Project (MBHCP)
Department of Psychiatry,
Louisiana State University Health Science Center |
| 2014 – 2015 | Predoctoral Internship
Department of Psychiatry,
Louisiana State University Health Science Center |

Research Presentations

Brokenbough, M. A., Archer, R. P., Handel, R. W., & Elkins, D. E. (2013, March). *Internal and external psychometric properties of the Minnesota Multiphasic Personality Inventory-Adolescent-Restructured Form (MMPI-A-RF) Restructured Clinical Scales in a forensic sample*. Paper presented at the 75th Annual Meeting of the Society for Personality Assessment, San Diego, CA

Brokenbough, M. A., Handel, R. W., & Archer, R. P. (2012, April). *Scale-level and item-level factor structure of the MMPI-A content scales in a forensic sample*. Paper presented at the 47th Annual Symposium on Recent Developments in the use of the MMPI-2/MMPI-2-RF/MMPI-A, Las Vegas, NV.