December 2022

The Impact Of Fake-Bad And Fake-Good Responding On The Millon Clinical Multiaxial Inventory, Fourth Edition (MCMI-IV)

Michael Sasan Jowkar

The Impact of Fake-Bad and Fake-Good Responding on the Millon Clinical Multiaxial Inventory, Fourth Edition (MCMI-IV)

by

Michael S. Jowkar

Bachelor of Science, University of North Dakota, May 2020

A Thesis

Submitted to the Graduate Faculty

of the

University of North Dakota

in partial fulfilment of the requirements

for the degree of

Master of Science

Grand Forks, North Dakota

December

2022

Name: Michael Jowkar

Degree: Master of Science

This document, submitted in partial fulfillment of the requirements for the degree from the University of North Dakota, has been read by the Faculty Advisory Committee under whom the work has been done and is hereby approved.

DocuSigned by:
2C877E791A994CE...
Joseph Miller

DocuSigned by:
John Paul Legerski
CB697C88AA83416...
John-Paul Legerski

DocuSigned by:
Thomas Petros
C4A022268E24476...
Thomas Petros

This document is being submitted by the appointed advisory committee as having met all the requirements of the School of Graduate Studies at the University of North Dakota and is hereby approved.

DocuSigned by:
Chris Nelson
2E0AE088C733403...
Chris Nelson
Dean of the School of Graduate Studies

9/15/2022
Date

PERMISSION

Title **The Impact of Fake-Bad and Fake-Good Responding on the Millon Clinical Multiaxial Inventory, Fourth Edition (MCMI-IV)**

Department Clinical Psychology

Degree Psychology, M.S.

In presenting this document in partial fulfillment of the requirements for a graduate degree from the University of North Dakota, I agree that the library of this University shall make it freely available for inspection. I further agree that permission for extensive copying for scholarly purposes may be granted by the professor who supervised my dissertation work or, in her/his absence, by the Chairperson of the department or the Dean of the Graduate School. It is understood that any copying or publication or other use of this dissertation or part thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of North Dakota in any scholarly use which may be made of any material in my dissertation.

Name Michael Jowkar
Date 10/24/2022

ACKNOWLEDGEMENTS

I wish to express my sincere appreciation to the members of my advisory Committee for their guidance and support during my time in the master's program at the University of North Dakota.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

Abstract

The MCMI-IV is a personality scale that attempts to measure psychopathology as well as normal personality functioning. The MCMI-IV has a number of validity scales designed to measure over-reporting and under-reporting of psychological symptoms. No research has been done on the MCMI-IV's ability to discriminate between respondents who are either over-reporting or under-reporting symptoms. Studies on previous versions of the test suggest elevations on the validity scale Y and elevations on personality scales 4, 5, and 7 in clients who are under-reporting symptoms. Previous research also suggests elevations on the validity scale X and Z in clients who are over-reporting symptoms and multiple clinical scales above 85 BR. This study analyzed differences in validity scales when participants were instructed to over-report and under-report symptoms when compared to honest respondents. It was hypothesized under-reporters would score high on scale Y, while over-reporters would score high on scales X with a low scale Y and Z. The results were consistent with the hypothesis. Additionally, ROC curves were analyzed between the H and FG group to identify specificity and sensitivity at BR 75 and 85 recommended by the test manual. An optimal cutoff score of 73.50 was identified for the best trade-off of specificity and sensitivity.

Introduction

Profile distortion

Self-report measures are often used to supplement diagnostic and clinical decision making. These measures are only valid if the client responds openly and honestly. However, clients occasionally respond in ways, consciously or unconsciously, to present themselves as more or less pathological as they actually are (Bagby & Marshall, 2005), thereby distorting (Morey, 2003) the test score or profile. Negative distortion, associated with a putative overreporting of symptoms, signs of dysfunction, etc., may be variously motivated. If intentional—the often described "fake-bad" set (Greene, 2010; Graham, 2012)— may be diagnosable if the motivation or goal can be identified (e.g., Malingering or factitious Disorder; 5th ed.; DSM–5; American Psychiatric Association, 2013), such as seeking medication, accommodation, attention or avoidance of personal responsibility.

Such responding is expected to occur more frequently in assessment contexts where some form of secondary gain is inherent, such as evaluations for ADHD medication (Alfano & Boon, 2007; Bryant et al., 2017; Sullivan, May, & Galbally, 2007; Suhr et al., 2008; Williamson et al., 2014) disability/SSDI evaluation (Bryant et al., 2017; Chafetz et al., 2007; Fisher & Watkins, 2008; Quinn, 2003) or personal injury litigation (Daubert & Metzler, 2000; Lenny & Dear, 2009). Negative distortion may be unintentional however, as evidenced by correlations between measures of self-unfavorable response and measures of psychopathology characterized by demoralization (Hathaway & McKinely, 1983; Millon et al., 2015; Morey, 1995). Moreover, and especially in the internet age, negative distortion may reflect a confirmation bias toward a self-diagnosed disorder (Ackerman, 2010; Butcher, 2009).

Positive profile distortion occurs when the client responds in a manner reflective of less pathology or dysfunction than is actually present (Craig, 2005). If intentional – the so-called "fake-good" set (Bagby et al., 1991) –the client is attempting to hide symptoms for a variety of reasons. Such distortions are expected in assessment contexts where minimization of symptoms is desirable, such as child custody evaluations (Daubert & Metzler, 2000; Lenny & Dear, 2009; McCann et al., 2001) and employment screening (Detrick et al., 2010; Morey et al., 1998; Rosse et al., 1998). However, as with negative distortion, positive distortion may not be intentional. It is well established non-patient community respondents tend to minimize their dysfunction and exaggerate their socially desirable traits (Craig, 2005), complicating psychometric discrimination between non-pathological respondents and pathological respondents who fake good intentionally (Morey 2003).

Millon Inventories

Millon's instruments are based in his biosocial-learning theory (Millon, 2011). This theory is based on Darwinian principles asserting personality is the product of trial and error in human evolution. More adaptive traits tend to persist while less adaptive traits tend to attenuate. These traits make up personality styles are meant to maintain balance between avoiding pain and pursuing pleasure (e.g., negative vs. positive reinforcement), adapting to the environment, and reproduction (Millon, 2011). Millon uses these dimensions to explain so-called "normal" personality function and personality pathology. For example, a functional person seeks to adapt to their environment, but a person who adapts too readily to their environment may not be confident in themselves and change their behavior too readily based on external stimuli.

Prior to Millon, theories of personology had disparate conceptual frameworks. For example, Galton believed bumps on the skull corresponded to different personalities (Millon,

2011), but these concepts were not connected with learning. Some theories completely deemphasize evolution and focus almost exclusively on learning components (e.g. behaviorism). Millon attempts to create a singular conceptual framework for normal and pathological personality development combining both social learning and evolutionary approaches (Millon, 2011). Through trial and error humans learn behavior to get their evolutionary needs met. These different ways of getting needs met are considered personality styles.

These styles are based on polarities in the categories of existence, adaptation, and replication (Millon et al., 2015). Existence describes a person's relationship between pleasure pain. A person high in the pleasure polarity may be motivated to seek pleasure and avoid pain. Some individuals may have reverse polarities (e.g. seeking pain and avoiding pleasure). Adaptation refers to if a person's coping style is active or passive. An active individual will actively seek to change the environment, while the passive individual will accommodate to the environment (Millon, 2011). Replication refers to whether a person emphasize the self or others. An individual's pattern of behavior can be explained by these categories. For example, a person who seeks pleasure for themselves in an active way may be considered to have an Anti-Social style. Millon's first personality inventory was meant to measure these personality traits. They can be classified on a continuum from adaptative to maladaptive.

Millon's original personality assessment, the Millon-Illinois Self-Report Inventory (MISRI), was based on his theory of personality (Choca & Grossman, 2015). Elevations on MISRI scales did not necessarily indicate pathological functioning and could represent adaptive traits (Choca & Grossman, 2015). He later refined the items on his scale leading to the first Millon Clinical Multiaxial Inventory (MCMI). These findings were incorporated into the

personality disorders in the Diagnostic and Statistical Manual of Mental Disorders, third edition (Groth-Marnat & Wright, 2016).

Millon's assessments were designed within a theoretical framework and later tested empirically. Millon used a three-stage process for test construction (Millon, 2011). These stages include the theoretical-substantive stage, internal-structure stage, and external-criterion stage (Loevinger, 1957). During the theoretical-substantive stage, relevant items related to constructs of interest are added to the test. In the internal-structure stage, examinees are given the previous iteration of MCMI and the items for the new installment. Clinicians were then asked to rate the test taker on how well they believed the client would score on different constructs. Items were retained based on their correlation with the representative scale, how well they represented the clinical construct of interest, and endorsement frequency (Loevinger, 1957). Items are then subjected to a confirmatory factor analysis, which further eliminates items. The final step of the internal-structure stage is calculating Cronbach's alpha for the remaining items. The External-Criterion stage compares the MCMI results to other assessments measuring similar constructs (Millon, 2011). The test manual reports MCMI-IV scale internal consistency values between .92 and .67, the former being the melancholic scale and the later being borderline (Millon et al., 2015). Overall the median internal consistency value of the clinical scales is .84. The test manual reports test-retest reliability is around .8 for most groups.

MCMI scales.

The most current iteration of the adult Millon personality instrument is the Millon multiaxial Clinical Inventory, Fourth Edition (MCMI-IV). . This test assesses personality psychopathology, emotional adjustment, and test-taking attitude (Groth-Marnat & Wright, 2016). This measure is made up of 195 true/false items generating 12 clinical personality patterns, 3

severe personality pathologies, 7 clinical syndromes and 3 severe clinical syndromes. The personality scales mirror Millon's personality theory which combines empiricism with prototypic personality styles (Grossman & Amendolace, 2017). The clinical scales allow interpretation of more temporary behavior that might accompany elevations on personality scales (Millon et al., 2015). Clinical scales also allow a more nuanced understanding of personality scale elevations (Grossman & Amendolace, 2017).

The MCMI-IV was constructed to help clinicians identify personality disorders in psychiatric populations (Groth-Marnat & Wright, 2016). The MCMI-IV uses "Base Rate" (BR) scores that are based on the prevalence of the disorders in clinical samples. BR scores account for the difference in base rates between disorders, unlike most tests (e.g. MMPI) which transform skewed data into a normal distribution (Millon et. al, 2015). A BR score of 60 indicates the median raw score of the clinical sample (Millon et. al, 2015). BR 75 represents the raw score of clients who had a DSM-5 diagnosis or had severe dysfunction. BR 85 represents the raw score of clients in the clinical sample that had a DSM-5 diagnosis or severe dysfunction in a more pronounced manner (Grossman & Amendolace, 2017). BR 115 represents the maximum raw score for a scale (Millon et. al, 2015). A BR score of 60-74 suggests a personality style, a score of 75-84 suggests a problematic personality type, and a score of 85 and above suggests a clinical personality disorder (Millon et al., 2015).

The MCMI-IV has two validity measures and three modifying indices. The Invalidity (V) scale is made up of three items designed to detect absurd or random responding by presenting content the client is unlikely to truthfully endorse (Groth-Marnat & Wright, 2016). Endorsing one item on the V scale suggests "questionable validity", while endorsing two or three items invalidates the profile (Groth-Marnat & Wright, 2016).

The Inconsistency scale (W) is made of 25 pairs of items similar statistically and semantically (Groth-Marnat & Wright, 2016). These items are expected to be answered in the same direction. A score of 20 to 25 invalidates the profile and could suggest random or inconsistent responding (Groth-Marnat & Wright, 2016).

The Disclosure (X) Index is calculated from raw scores on scales 1 to 8B (Millon et. al, 2015). This scale may identify respondents who are either over reporting symptoms or under reporting symptoms. A score below 7 or above 114 renders the profile uninterpretable.

The Desirability (Y) index measures the proclivity to "appear socially attractive, morally virtuous, or emotionally well composed" (Millon et. al, 2015). BR scores 75 or higher may indicate a person is intentionally denying personality dysfunction or minor flaws (Grossman & Amendolace, 2017).

The Debasement (Z) index is designed to measure feigned psychopathology (Grossman & Amendolace, 2017). Clients who endorse items on this scale BR scores of 75 or higher suggest a person may be presenting as more pathological than they actually are (Millon et. al, 2015). High scores on this scale could also mean the client is experiencing a higher level of emotional distress than the clinical reference group or suggest the client is aware of their need for treatment (Grossman & Amendolace, 2017).

The fourth iteration of the MCMI has many changes from previous versions on both personality scales and validity scales. The turbulent scale is a new addition to the fourth edition. This prototype represents a personality type who has a proclivity towards pleasure and activation (Choca & Grossman, 2015). The MCMI-IV also places more emphasis on dimensional categories of personality, ranging from styles that are generally adaptive to personality traits that can cause distress, impairment or severely limited functioning (Choca & Grossman, 2015).

Detecting Over-reporting.

There are no current studies identifying how well the MCMI-IV modifying scales detect fake-bad profiles, though there is some research relevant to prior versions of the test (Choca & Pignolo, 2022). A study on the MCMI-III indicated Scale X and Scale Z scores are higher in students instructed to fake-bad when compared to students who responded honestly (Schoenberg, Dorr, & Morgan, 2003). Daubert & Metzler (2000) found psychiatric patients who were instructed to fake-bad scored significantly higher on Scale X and Scale Z, while scoring significantly lower on Scale Y on the MCMI-III. Aguerrevere et. al (2011) found MCMI-III scales X and Y were higher in patients with Traumatic Brain Injuries (TBIs) who were malingering symptoms. Aguerrevere et. al (2011) also found lower scale Z scores in patients who were malingering TBIs.

There are no current studies on how MCMI-IV personality scales relate to fake-bad profiles. However, Choca (2004) found fake-bad profiles typically have multiple elevations at 85 BR or higher on the MCMI-III. The most common elevations on malingered profiles were Schizoid, Avoidant, Negativistic, Borderline, Paranoid, and Schizotypal. Elevations on these scales can also indicate authentic pathology, so these scores need to be interpreted with caution.

Detecting Under reporting.

There are no current studies identifying how the MCMI-IV modifying scales detect fake-good profiles. However, Daubert & Metzler (2000) found psychiatric patients who were instructed to under report symptoms scored higher on Scale X and Y and lower on Scale Z on the MCMI-III. Lenny & Dear (2009) found parents who were instructed to look like “good parents” on the MCMI-III had low scores on scale X and Z and elevated Y scores. They also found these participants tended to elevate on scales 4, 5, and 7 more frequent than honest responders.

Students instructed to fake good scored higher on the Y scale when compared to students who were honestly responding or faking bad (Bagby et al., 1990; Retzlaff et al., 1991).

Along with elevations on scale Y, some researchers found elevations on scales 4, 5, and 7 in fake-good response sets (Fals-Steward, 1995; McCann et al., 2001; Lenny & Dear 2009). Fals-Steward (1995) compared the MCMI-II profiles of those trying to hide substance abuse to those openly admitting substance abuse. Those who tried to hide substance abuse had subclinical elevations on scales 4, 5, and 7 when compared to honest responders who clinically elevated scales 4 and 5. These differences in profiles could suggest subclinical elevations on scales 4, 5, and 7 are indicative of positive distortion. Lenny & Dear (2009) found elevated scores on scales Y, 4, 5, and 7 and low scale z scores in a sample of parents instructed to "fake-good" for a hypothetical child custody hearing. McCann (2001) found elevations on the MCMI-II scales 4, 5, and 7 in parents who took personality testing for a child custody hearing. These scores may suggest elevations on scales 4, 5, and 7 may be indicative of under-reporting. However, these scales are positively correlated with healthy personality features and negatively correlated with psychopathology, so they may be indicative of a healthy individual (Craig & Olson, 1992; Craig & Weinberg, 1993; Holliman & Guthrie, 1989). These scores also represent the average profile of a person "faking-good" and may provide little diagnostic utility in finding individuals who are faking good.

There is little research on how well the MCMI-III discriminates between dishonest responding. There is no research on how well the MCMI-IV distinguishes between honest and dishonest responding. Research on clinical scale elevations with fake-good and fake-bad responding is limited to the MCMI-II and MCMI-III, with no such research existing on the MCMI-IV.

History of MCMI validity scales

Millon's first personality inventory, the MISRI, did not have any validity scales (Choca & Grossman, 2015). The first MCMI had one validity scale that contains absurd items. This scale was designed to invalidate profiles of those who were randomly responding or not paying attention to test items. The MCMI-II added Disclosure, Desirability and Debasement scales. Scores on these scales adjusted the BR scores on personality scales. This version of the test also had adjustments for how a number of mental states and environmental factors may affect client scores. These adjustments made profile scoring complicated and were removed in later versions of the test. The third version of the MCMI kept the Disclosure, Desirability, and Debasement scales from the previous version and added an inconsistency scale. This scale was designed to measure erratic responding across the personality measure. The MCMI-IV includes Disclosure, Desirability, Debasement, Inconsistency and Invalidity modifying indices. These scales are discussed more in-depth elsewhere.

Hypotheses

The absence of research on the MCMI-IV validity scales means their ability to identify profile distortion is largely unknown (Craig, 2005; Choca & Pignolo, 2022). All MCMI modifying indices studies are on previous versions of the test. These versions may be similar to those of the fourth edition, but their operarting characteristics must be established directly. This study aims to collect data on the MCMI-IV validity scales with clear cases of profile distortion. It is predicted that respondents attempting to fake-good will elevate Scale Y and produce lower Scale X and Z scores when compared to fake-bad participants and honest responding participants. It is also predicted those who are instructed to fake-bad will elevate scale X with low scores on scale Y and Z relative to the honest responders. Unusually low scores on

personality scales may have utility in identifying fake-good profiles. Unusual elevations may be useful in identifying negatively distorted profiles.

Methods

Participants

One-hundred fifteen participants were recruited from undergraduate psychology courses. The participants were between ages of 18 and 27 (M = 16, F = 79). 29 participants were assigned to the Honest (H) group, 39 to the Fake-Good (FG) group, and 47 to the Fake-Bad (FB) group.

Materials

Participants in the H, FB, and FG groups completed a consent form, a demographics questionnaire, the Patient Health Questionnaire-9 (PHQ-9), the Generalized Anxiety Disorder-7, and the MCMI-IV. All information was gathered in-person in a laboratory setting.

The PHQ-9 is a 10-item Likert scale assessment designed to measure depressive symptoms over a 2-week interval (Kroenke et al, 2001). The First 9 items are designed to assess depressive symptoms over the last two weeks (Hunsley & Mash, 2018). The 10th item assesses the level of functional impairment from these symptoms (Hunsley & Mash, 2018). Scores from 0-4 are considered normal, 5 to 9 are considered mild, 10 to 14 are considered moderate, 15 to 19 as moderately severe, and 20+ are considered severe depressive symptoms (Hunsley & Mash, 2018). Menea et al., (2012) found the optimal PHQ-9 cutoff score for detection of Major Depressive Disorder to be between 8- and 11.

The GAD-7 is a 9-item Likert scale assessment designed to capture the core features of the DSM-IV diagnosis Generalized Anxiety Disorder (Spitzer et al., 2006). Plummer et al., (2015) suggests using cutoff scores of 7-10 to identify generalized anxiety disorder.

The PHQ-9 and the GAD-7 were used to identify any pre-existing groupwise differences in psychological symptoms that might influence results on the MCMI-IV protocol. For the purpose of this study we examined the mean differences in anxiety and depression symptoms between experimental groups.

Procedure

Participants were tested in small groups of 1-10 participants in a laboratory setting. They were first required to read the study information sheet which described the study including what was expected, the risks and benefits, and the approaches to maintaining confidentiality. After the participants read through the online consent form and had any questions answered, the researcher logged participants into an online survey according to the instructions of the experimenter using the confidential study number.

During the online survey, the participant completed a demographic sheet including questions about age, gender, race and education level. In addition, participants were asked to respond to whether they are currently being treated or have been treated by a Psychologist or a Psychiatrist, and if so, what they have been treated for. They then completed the PHQ-9 and GAD-7. When these questionnaires were completed, the participant logged into a second online survey using their subject number and asked to complete the MCMI-IV under one of three instructional sets (See Appendix A). In the honestly-responding (H) group, they were asked to complete the inventory under the standard instructions. In the fake-good (FG) condition, the participants were instructed to portray themselves in the best possible light. In the fake-bad (FB) condition, the participants were instructed to respond to the items to portray themselves in a negative light.

Results

PHQ-9 and GAD-7 analysis

The mean of total scores on the GAD-7 was computed for each group. These mean scores suggest mild anxiety is presented in across conditions. It is unlikely MCMI-IV scores were significantly influenced by reported anxious symptoms. A one-way ANOVA determined there were no significant group differences ($p > .05$) in mean GAD-7 scores. This data suggests there were no significant differences in reported anxiety symptoms between groups.

The mean total scores of the PHQ-9 was computed for each group. These mean scores suggest mild depression is present across conditions. It is unlikely MCMI-IV scores were significantly influenced by reported depressive symptoms. A one-way ANOVA determined there were no significant group differences ($p > .05$) in mean PHQ-9 scores. These data suggest there were no significant differences in psychiatric symptoms between groups.

Table 1

Means of the GAD-7 and PHQ-9 as a function of group

	Group		
	Honest Response (H) *N*=29	Fake-Good (FG) *N*=31	Fake-Bad (FB) *N*=35
GAD-7	6.68 (4.64)	6.69 (4.29)	6.72 (5.34)
PHQ-9	6.31 (6.12)	5.93 (4.72)	6.71 (5.68)

Note: Standard Deviations are in parentheses.

MCMI-IV Analyses

Profiles that elevated the V ($V \geq 2$) or W ($W \geq 20$) scale were considered invalid and excluded from data analysis, eliminating twenty total profiles and leaving the participant count as

twenty nine for H, thirty-one for FG, and Thirty-five for FB. A series of one-way ANOVAs were conducted to compare differences on scales X, Y, and Z between the Honest (H), Fake-Good (FG), and Fake-Bad (FB) groups. Due to unequal group sizes and violations of homogeneity of variance, Welch's F tests were calculated to compensate for increased Type I error risk (Howell, 2013). The results of Welch's test was the same as the ANOVAs, so only uncorrected F values will be reported. Subsequent analyses are done using Tukey pairwise comparison procedure with alpha set at .05.

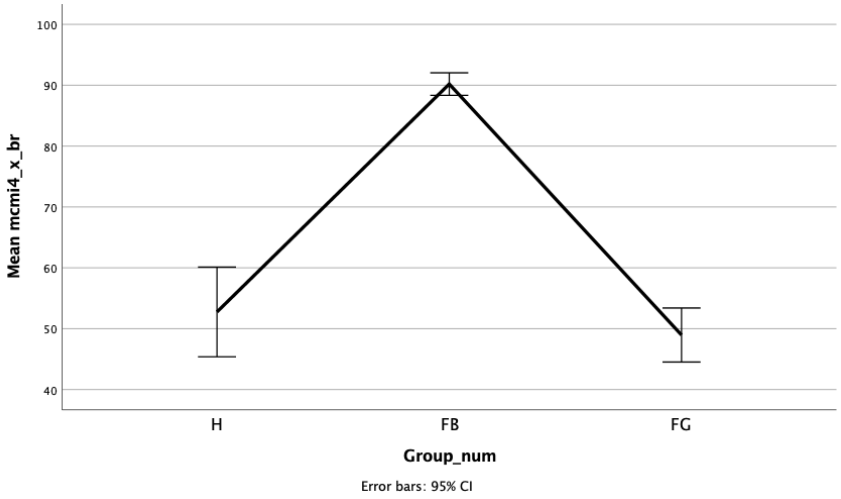
Validity and Modifying Scales.

The means and standard deviations for the response styles are presented in Table 2. A series of one-way ANOVAs was conducted with each of the three response style scale scores as DV, producing significant differences for Scale X $F(2, 94) = 99.90$, Scale Y $F(2, 94) = 116.25$, and Scale Z $F(2, 94) = 91.48$, with all $p < .001$.

Follow up Tukey tests for scale X indicate the FB (90.20) group scored significantly higher than H (52.76) and FG (48.97), which did not differ significantly from each other.

Figure 1

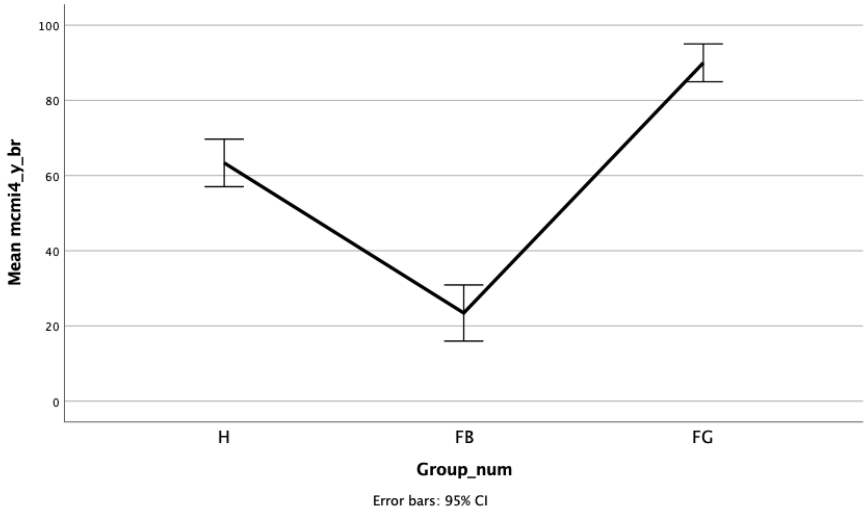
Mean scale X scores across conditions



Tukey tests for Scale Y indicate FG (90) group scored significantly higher than H (63.38) and FB (23.46) groups. Additionally, the H (63.38) group scored significantly higher on scale Y compared to the FB (23.46) group.

Figure 2

Mean scale Y scores across conditions



Tukey tests for Scale Z indicate FB (89) scored significantly higher than H (46.45) and FG (17.48). Additionally, the H (46.45) group scored significantly higher on scale Y compared to the FG (17.48) group.

Figure 3

Mean scale Z scores across conditions

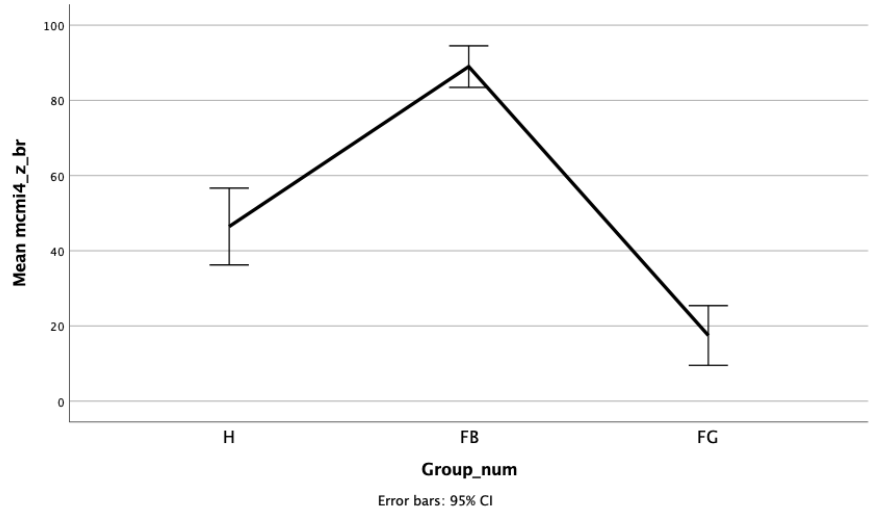


Table 2

Mean MCMI-IV Modifying Scale Scores as a Function of Group

	Group		
	Honest Response (H) *N*=29	Fake-Good (FG) *N*=31	Fake-Bad (FB) *N*=35
Scale X	52.76 (19.36)	48.97 (12.11)	90.20 (5.39)
Scale Y	63.38 (16.61)	90 (13.71)	23.46 (21.72)
Scale Z	46.45 (26.86)	17.48 (21.66)	89 (16.10)

Note: Standard Deviations are in Parentheses.

A Receiver Operator Characteristic (ROC) analysis (McFall & Treat, 1999, Zhou et al, 2011) was conducted for FG versus H on Scale Y to quantify sensitivity and specificity across all possible cut scores, and, in particular, at BR score cut scores recommended in the MCMI-IV

manual. At BR 75, the sensitivity = .903 and specificity = .828. At BR 85, the sensitivity = .742 and specificity = .931. Additionally, the AUC statistic is .919, suggesting excellent discrimination overall. An ROC curve can be used to generate an optimal cut-off score by looking at the "bend" in the curve. Our analysis suggests an optimal cutoff score of BR 73.50 with sensitivity = .935 and specificity = .759. This cutoff score has higher sensitivity and lower specificity when compared to the recommended cutoff score at BR 75 and BR 85.

Figure 4

ROC Curve for FG versus H on Scale Y

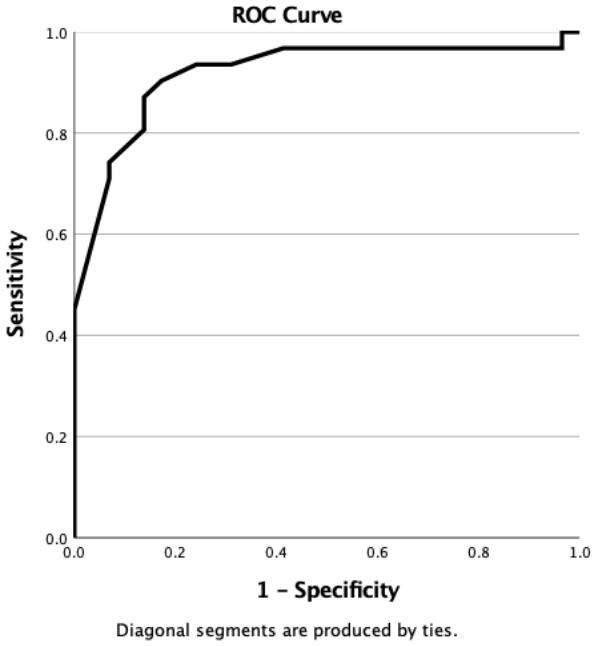


Table 3

ROC cutoff scores for FG versus H on Scale Y

Coordinates of the Curve

Test Result Variable(s): mcmi4_y_br

Positive if Greater Than or Equal To[a]	Sensitivity	1 – Specificity
24.00	1.000	1.000
27.50	1.000	.966
32.50	.968	.966
37.50	.968	.931
42.50	.968	.897
47.50	.968	.793
52.50	.968	.724
57.50	.968	.690
61.50	.968	.655
64.50	.968	.552
67.50	.968	.414
70.50	.935	.310
73.50	.935	.241
76.50	.903	.172
79.50	.871	.138
83.00	.806	.138
87.00	.742	.069
91.00	.710	.069
95.00	.452	.000
98.50	.161	.000
101.00	.000	.000

The test result variable(s): mcmi4_y_br has at least one tie between the positive actual state group and the negative actual state group.

a. The smallest cutoff value is the minimum observed test value minus 1, and the largest cutoff value is the maximum observed test value plus 1. All the other cutoff values are the averages of two consecutive ordered observed test values.

Discussion

Conclusions

This study evaluated the responsiveness of the three MCMI-IV modifying indices to respondents in three different groups, instructed to respond honestly or to either “fake bad” or “fake good” Per the MCMI-IV manual, Scale X elevates in direct proportion to the elevations on personality scales 1 through 8B, while Scale Z (Debasement) was designed specifically to elevate when responses indicate an unusually negative self-appraisal, as is frequently seen in fake-bad response sets. Therefore, BR scores for these scales were expected to be higher in “fake-bad” respondents, relative to “fake-good” or honest responders. This was verified by the

relevant ANOVAs and post hoc tests. On the other hand, Scale Y (Desirability), was expected, owing to its construction, and data from prior versions of the test and in the MCMI-IV manual, was expected to elevate in the "fake-good" condition, relative to the "fake-bad" and honest conditions. This was also born out by analysis of variance. Together, these data suggest that, at least within the somewhat artificial context of the experiment, MCMI-IV Scales X, Y, and Z behave as expected across the imposed instructional groups.

These group mean comparison data, however, do not offer estimates of their discrimination only that some degree of discriminate power is likely in all cases. This is an important point, as one purpose of the modifying indices is to identify positively or negatively distorted response sets, which, in the case of the current study, was dictated by instructions to intentionally "fake-good" or "fake-bad". Discriminative power in clinical tests is typically represented by the test's sensitivity (the probability that the test will detect an actual case of distortion) and specificity (the probability that the test will accurately identify an honest profile). The MCMI-IV manual recommends that caution is warranted when modifying indices are elevated above 75BR, though it is possible that experimental or clinical data could reveal more "optimal" thresholds for identifying positive (Scale Y) or negative (Scales X & Z) distortion.

ROC analysis, frequently used to evaluate the sensitivity and specificity of a scale across all possible thresholds (cutting scores), requires that scores from two groups be compared. However, the choice of groups is important. Discriminative data should reflect decisions that are typical in actual clinical practice to protect ecological validity and diagnostic utility. For example, if a clinician is concerned that a patient might be minimizing pathology (i.e., "faking good") in order to appear as if they are an honestly-reporting, non-pathological, respondent, then they are interested in a comparison between "fake-good" and "honest" respondents (assuming

the honest responders are also non-clinical). In this case, data on the test's ability to discriminate between "fake-good" and "fake bad" response would be *irrelevant* to the clinical question at hand, Worse, because we would expect that the score differences would be much larger between 'fake good" and fake bad" respondents than between "fake good" and "honest" non-pathological respondents, data based on the prior comparison would, in this hypothetical clinician's case, generally over-estimate the test's discrimination accuracy, and increase the odds of a false diagnostic conclusion.

Of the three conditions in this study (H, FB and FG), only the FG versus H comparisons reflect the sorts of questions typically asked by clinicians of their test data, where response distortions are suspected. That is, we can think of no referral issue where the diagnostic question could include "is the evaluee 'faking bad' or are they honestly non-pathological?" (FB vs H) or "is the evaluee 'faking bad' or are they 'faking good'?" (FB vs FG). This is why only Scale Y's ability to discriminate between FG and H, a possible and reasonable referral question, was established via ROC analysis, and shown to be quite good. At the recommended BR threshold of 75, sensitivity is in excess of .90, while specificity is only slightly lower. The optimal cut-off score at BR 73.50 has a sensitivity of .935 and a specificity of .759. Suggesting the optimal cut-off score better classifies clinical cases but is less effective at screening out non clinical cases.

These data partially supports the hypothesis that FG participants would elevate on Scale Y and have significantly lower Scales X and Z when compared to FB and H participants. The participants in the FG group did score significantly higher on Scale Y than FB and H participants. The FG group also had had a significantly lower X score when compared to FB respondents, but not when compared to respondents in group H. The data supports the hypothesis that FB participants would elevate on Scale X and have significantly lower scores on Scale Y

and Z when compared to group H. The ROC curve comparing the H and FG groups on scale Y, suggest the MCMI-IV is capable of detecting under-reporting of psychological data. It is unclear how well the MCMI-IV detects over-reporting of psychological data without a clinical group to compare the FB group to because scales X and Z are designed to detect feigned pathology when compared to authentic pathology.

Limitations

This study was conducted with a relatively young (18 – 27) college aged sample. It is unclear how well these results would generalized to other populations. This study also explicitly instructed participants to distort their profiles, there may be differences in overt profile distortion versus more subtle or unconscious attempts to distort profiles. Most problematically, the study lacked an honestly-responding clinical group, which, as described earlier , would be essential to establish the discriminative power of the X and Z scales. That is, clinicians do frequently ask "is the evaluee faking pathology, or is pathology actually present (FB versus Clinical). The absence of an actual clinical group makes meaningful analysis of sensitivity and specificity of the X and Z scales impossible.

Future Directions

This study provided preliminary evidence that the MCMI-IV modifying indices are behaving as expected. More research needs to be done with community and clinical samples to improve the generalizability of these results. Future studies may also look to see if external reward (i.e. providing a reward for malingering) generating different profiles on the modifying indices when compared to verbally instructing someone to malinger without a reward.

References

Ackerman, M. J. (2010). *Essentials of forensic psychological assessment*. John Wiley & Sons.

Aguerrevere, L. E., Greve, K. W., Bianchini, K. J., & Ord, J. S. (2011). Classification accuracy of the Millon Clinical Multiaxial Inventory-III modifier indices in the detection of malingering in traumatic brain injury. *Journal of Clinical and Experimental Neuropsychology*, *33*(5), 497–504. https://doi.org/10.1080/13803395.2010.535503

Alfano, K., & Boone, B. K. (2007). The use of effort tests in the context of actual versus feigned attention-deficit/hyperactivity disorder and learning disability. In K. B. Boone (Ed.), *Assessment of feigned cognitive impairment: A neuropsychological perspective*. New York: The Guilford Press.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). https://doi.org/10.1176/appi.books.9780890425596

Bagby, R. M., & Marshall, M. B. (2005). Assessing Response Bias with the MCMI Modifying Indices. In R. J. Craig (Ed.), *New directions in interpreting the Millon™ Clinical Multiaxial Inventory-III (MCMI-III™)* (pp. 227–247). John Wiley & Sons Inc.

Bagby, R. M., Gillis, J. R., & Dickens, S. (1990). Detection of dissimulation with the new generation of objective personality measures. *Behavioral Sciences & the Law*, *8*(1), 93–102. https://doi.org/10.1002/bsl.2370080111

Bagby, R. M., Gillis, J. R., Toner, B. B., & Goldberg, J. (1991). Detecting fake-good and fake-bad responding on the Millon Clinical Multiaxial Inventory-II. *Psychological*

Assessment: A Journal of Consulting and Clinical Psychology, *3*(3), 496–498. https://doi.org/10.1037/1040-3590.3.3.496

Bryant, A. M., Lee, E., Howell, A., Morgan, B., Cook, C. M., Patel, K., Menatti, A., Clark, R., Buelow, M. T., & Suhr, J. A. (2017). The vulnerability of self-reported disability measures to malingering: A simulated ADHD study. *The Clinical Neuropsychologist*, *32*(1), 109–118. https://doi.org/10.1080/13854046.2017.1346145

Butcher, J. N. (2009). *Oxford Handbook of Personality Assessment*. Oxford University Press.

Chafetz, M., Abrahams, J., & Kohlmaier, J. (2007). Malingering on the Social Security Disability Consultative Exam: A new rating scale. *Archives of Clinical Neuropsychology*, *22*(1), 1–14. https://doi.org/10.1016/j.acn.2006.10.003

Choca, J.P., & Grossman, S.D., (2015). Evolution of the Millon Clinical Multiaxial Inventory, *Journal of Personality Assessment*, 97:6, 541-549, https://doi.org/10.1080/00223891.2015.1055753

Craig, R. J. (2005). *New Directions in interpreting The Millon Clinical Multiaxial Inventory-III (MCMI-III)*. John Wiley.

Craig, R., & Weinberg, D. (1993) MCMI: Review of the literature. In R. Craig (Ed.), *The Millon Clinical Multiaxial Inventor: A clinical research information synthesis (pp.23 – 70).* Hillsdale, NJ: Laurence Erlbaum Associates

Craig, R., Olson, R. (1992). Relationship between MCMI-II scales and normal personality traits. *Psychological Reports*, *71*(7), 699–705. https://doi.org/10.2466/pr0.71.7.699-705

Daubert, S. D., & Metzler, A. E. (2000). The detection of fake-bad and fake-good responding on the Millon Clinical Multiaxial Inventory III. *Psychological Assessment*, *12*(4), 418–424. https://doi.org/10.1037/1040-3590.12.4.418

Detrick, P., Chibnall, J. T., & Call, C. (2010). Demand effects on positive response distortion by police officer applicants on the revised Neo personality inventory. *Journal of Personality Assessment*, *92*(5), 410–415. https://doi.org/10.1080/00223891.2010.497401

Fals-Stewart, W. (1995). The effect of defensive responding by substance-abusing patients on the Millon Clinical Multiaxial Inventory. *Journal of Personality Assessment*, *64*(3), 540–551. https://doi.org/10.1207/s15327752jpa6403_11

Fisher, A. B., & Watkins, M. W. (2008). ADHD rating scales' susceptibility to faking in a college student sample. *Journal of Postsecondary Education and Disability, 20*, 81–92.

Greene, R. L. (2010). *MMPI-2/MMPI: An interpretive manual*. Pearson Education.

Grossman, S., & Amendolace, B. (2017). *Essentials of MCMI-IV Assessment*. John Wiley & Sons.

Hathaway, S. R., & McKinley, J. C. (1983). *Minnesota multiphasic personality inventory: Manual for Administration and scoring*. University of Minnesota Press.

Holliman, N. B., & Guthrie, P. C. (1989). A comparison of the Millon Clinical Multiaxial Inventory and the California Psychological Inventory in assessment of a nonclinical population. *Journal of Clinical Psychology*, *45*(3), 373–382. https://doi.org/10.1002/1097-4679(198905)45:3<373::aid-jclp2270450305>3.0.co;2-r

Hunsley, J., & Mash, E. J. (2018). *A guide to assessments that work*. Oxford University Press.

Lenny, P., & Dear, G. E. (2009). Faking good on the MCMI–III: Implications for child custody evaluations. *Journal of Personality Assessment*, *91*(6), 553–559. https://doi.org/10.1080/00223890903228505

Loevinger, J. (1957). Objective tests as instruments of Psychological Theory: Monograph Supplement 9. *Psychological Reports*, *3*(7), 635. https://doi.org/10.2466/pr0.3.7.635-694

Manea, L., Gilbody, S., & McMillan, D. (2011). Optimal Cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): A meta-analysis. *Canadian Medical Association Journal*, *184*(3). https://doi.org/10.1503/cmaj.110829

McCann, J. T., Flens, J. R., Campagna, V., Collman, P., Lazzaro, T., & Connor, E. (2001). The MCMI-III in child custody evaluations. *Journal of Forensic Psychology Practice*, *1*(2), 27–44. https://doi.org/10.1300/j158v01n02_02

McFall & Treat (1999). Quantifying the information value of clinical assessments with signal detection theory. *Annual Review of Psychology, 50*, 215-241.

Millon, T. (2011). *Disorders of personality: Introducing a DSM/ICD spectrum from normal styles to abnormal*. Wiley.

Millon, T., Grossman, S., & Millon, C. (2015). Millon Clinical Multiaxial Inventory-IV (*MCMI-IV)*. Pearson.

Morey, L. C. (1996). *An interpretative guide to the personality assessment inventory (PAI)*. PAR.

Morey, L. C., & Lanier, V. W. (1998). Operating characteristics of six response distortion indicators for the Personality Assessment Inventory. *Assessment*, *5*(3), 203–214. https://doi.org/10.1177/107319119800500301

Plummer, F., Manea, L., Trepel, D., & McMillan, D. (2016). Screening for anxiety disorders with the GAD-7 and Gad-2: A systematic review and diagnostic metanalysis. *General Hospital Psychiatry*, *39*, 24–31. https://doi.org/10.1016/j.genhosppsych.2015.11.005

Quinn, C. A. (2003). Detection of malingering in assessment of adult ADHD. *Archives of Clinical Neuropsychology*, *18*(4), 379–395. https://doi.org/10.1093/arclin/18.4.379

Retzlaff, P., Sheehan, E., & Fiel, A. (1991). MCMI-II report style and bias: Profile and validity scales analyses. *Journal of Personality Assessment*, *56*(3), 466–477. https://doi.org/10.1207/s15327752jpa5603_8

Rosse, J. G., Stecher, M. D., Miller, J. L., & Levin, R. A. (1998). The impact of response distortion on preemployment personality testing and hiring decisions. *Journal of Applied Psychology*, *83*(4), 634–644. https://doi.org/10.1037/0021-9010.83.4.634

Schoenberg, M. R., Dorr, D., & Morgan, C. D. (2003). The ability of the Millon Clinical Multiaxial Inventory-third edition to detect malingering. *Psychological Assessment*, *15*(2), 198–204. https://doi.org/10.1037/1040-3590.15.2.198

Spitzer, R. L., Kroenke, K., Williams, J. B., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder. *Archives of Internal Medicine*, *166*(10), 1092. https://doi.org/10.1001/archinte.166.10.1092

Suhr, J., Hammers, D., Dobbins-Buckland, K., Zimack, E., & Hughes, C. (2008) The relationship of malingering test failure to self-reported symptoms and neuropsychological findings in adults referred for ADHD evaluation. *Archives of Clinical Neuropsychology*, *23*(5) 521-530. https://doi.org/10.1016/j.acn.2008.05.003

Sullivan, B. K., May, K., & Galbally, L. (2007). Symptom exaggeration by college adults in attention-deficit hyperactivity disorder and learning disorder assessments. *Applied Neuropsychology*, *14*, 189–207.

Williamson, K.D., Combs, H.L., Berry, D.T., Harp, J.P., Mason, L.H., & Edmundson, M. (2014). Discriminating among ADHD alone, ADHD with a comorbid psychological disorder, and feigned ADHD in a college sample. *The Clinical Neuropsychologist, 28*, 1182–1196.

Zhou, X.H., Obuchowski, N.A. & McClish, D.K. (2011). *Statistical Methods in Diagnostic Medicine: 2nd Edition*. Hoboken, New Jersey : John Wiley & Sons.

Appendix A

Instructions given to respondents

The standard instructions read: “The following pages contain a list of statements that people use to describe themselves. They are printed here to help you in describing your feelings and attitudes. Try to be as honest and serious as you can in marking the statements.”

The fake-bad instructions read: “In this experiment we are examining the ability of computer-scored personality tests to detect individuals who are trying to “fake-bad” in their responses. Situations in which might motivate a person to “fake-bad” may include trying to be acquitted of a legal charge on the grounds of insanity, applying for rehabilitative services, or trying to qualify for disability benefits. Please take this test as if you were trying to “fake-bad” as if you were in one of the circumstances I just mentioned. Normally, persons must be honest with their doctor or treatment team in order to be helped the most. However, this is an experimental situation. Just during this experiment, we would like for you to take this test in such a way as to create a negative impression of yourself, more psychologically troubled than would actually be the case, but in a believable and convincing manner.”

The fake-good instructions read: “In this experiment we are examining the ability of computer-scored personality tests to detect individuals who are trying to “fake-good” in their responses. Situations in which might motivate a person to “fake-good” may include trying to get custody of one’s children in a divorce case, attempting to secure release from a mental hospital,

and applying for a good job such as an airline pilot or a police officer. Please take this test as if you were trying to “fake-good” as if you were in one of the circumstances I just mentioned. Normally, persons must be honest with their doctor or treatment team in order to be helped the most. However, this is an experimental situation. Just during this experiment, we would like for you to take this test in such a way as to create a positive impression of yourself, more psychologically healthy than would actually be the case, but in a believable and convincing manner.”