Detection of Over- and Under-reporting with the Computer Adaptive Version of
the Minnesota Multiphasic Personality Inventory-2 (MMPI-2-CA)

Wendy R. Dragon
Kent State University
August 2008
TABLE OF CONTENTS

LIST OF TABLES ........................................................................................................v

ACKNOWLEDGEMENTS ..............................................................................................vi

Chapter I. Introduction ..............................................................................................1

  MMPI-2 Validity Scales that Detect Over-reporting ................................................. 5
  MMPI-2 Validity Scales that Detect Under-reporting .............................................. 6
  Computerized Assessment ...................................................................................... 7
  Effects of over- and under-reporting on substantive scales .................................. 11

Chapter II. Methods .................................................................................................. 13

  Participants .......................................................................................................... 13
  Procedures .......................................................................................................... 13
  Analyses ............................................................................................................ 15

Chapter III. Results .................................................................................................. 17

  Administration Effect on Validity Scales in Detecting Over-/Under-Reporting .... 17
  Over-reporters vs. standard instructions ................................................................. 18
  Under-reporters vs. standard instructions ............................................................... 21
  Classification accuracy of over-reporting indicators ............................................. 25
  Classification accuracy of under-reporting indicators ......................................... 28

  The Effect of Administration Mode and Instructions on the RC Scales .......... 32
  Mean Restructured Clinical Scale comparisons across over-reporting groups ... 33
  Mean Restructured Clinical Scale comparisons across under-reporting groups. 38
List of Tables

1. Validity Scales - Over-reporting vs. Standard Instruction Men ......................... 19
2. Validity Scales - Over-reporting vs. Standard Instruction Women ...................... 20
3. Validity Scales - Under-reporting vs. Standard Instruction Men ...................... 22
4. Validity Scales - Under-reporting vs. Standard Instruction Women..................... 23
5. Classification Accuracy and Predictive Power for Over-reporting Indicators......... 26
6. Classification Accuracy and Predictive Power for Under-reporting Indicators........ 29
7. RC Scales - Over-reporting vs. Standard Instruction Men ............................. 34
8. RC Scales - Over-reporting vs. Standard Instruction Women ........................... 36
10. RC Scales - Under-reporting vs. Standard Instruction Women ....................... 41
Acknowledgements

First of all, I would like to thank my advisor, Yossi Ben-Porath, for his time, guidance, and especially his patience throughout this project.

I’d also like to thank my husband, Scott, and my two children, Samantha and William, for all of your sacrifices and support so that I can pursue my education. I love and appreciate you all more than you will ever know.
CHAPTER 1

Introduction

Psychologists draw upon many forms of information, including interviews and behavioral observations, during the assessment process. They also rely extensively on measures of personality and psychopathology in order to answer assessment questions (Groth-Marnat, 2003). Ideally, respondents give direct and honest answers to questions posed to them during an assessment (Graham, 1990). However, misleading responding on such measures has long been a major concern (Nicholson, et al., 1997), and detection of misleading responding is considered to be an important task in personality assessment (malingering; Meyers, Millis, & Volkert, 2002).

Misleading responding can take two forms: over- and under-reporting. When the target of the assessment attempts to present him or herself as having more psychological problems than he or she actually does, this is known as negative self-presentation, malingering or over-reporting (Weiner, 2003). This can occur in a variety of settings (in forensic settings; Gallagher, 1998; in clinical settings; Weiner, 2003), and for a variety of reasons, ranging from the innocuous to malicious. For example, respondents may exaggerate their symptoms in order to receive services they believe they would not otherwise receive (Weiner, 2003) or as a plea for help (Graham, 2006). Some plaintiffs malinger the extent of their psychological disturbance in order to increase the amount of a settlement in a personal injury case (Baer & Miller, 2002). In the same vein, some
defendants in criminal actions mangle psychological disturbance to minimize the penalties for their crimes (Wiener, 2003).

Occasionally, the target of the assessment wants to present him or herself as being psychologically healthier than he or she actually is. This is known as positive self-presentation, defensiveness or under-reporting (Wiener, 2003). Individuals undergoing psychological assessment often have much to gain by concealing or minimizing psychopathology (Bagby, et al., 2000). Patients who wish to be discharged from a hospital, prisoners who are requesting parole (Bagby, et al., 2000), and parents seeking custody of or access to their children have powerful incentives for under-reporting (Bagby & Marshall, 2004). People who are applying for jobs or training (Baer & Miller, 2002) or who are up for promotion (Weiner, 2003) may also be motivated to present themselves in an overly positive light.

The MMPI-2 (Minnesota Multiphasic Personality Inventory-2), like other self-report measures of personality and psychopathology, is susceptible to both unintentional and deliberate distortion (Bagby, et al. 1994). Test-takers respond to the 567 items on this instrument by indicating if the item is true, or mostly true as applied to them, or false, or mostly false as applied to them. The test instrument contains validity scales designed to assess over- and under-reporting test-taking approaches. Detecting response distortion on the MMPI-2 is particularly important because of its frequent use in psychological assessments across a number of different settings. It is the psychological test most frequently used by clinical psychologists (Camara, Nathan, & Puente, 2000). Only the Wechsler intelligence scales are used more often in forensic evaluations (Lees-Haley,
Smith, Williams, & Dunn, 1996) where it is regularly used (Pelfrey, 2004). It is also by far the most frequently used psychological test in health assessment (Piotrowski & Lubin, 1990).

The MMPI-2 is also the most extensively researched psychological measure of feigned mental disorders (Rogers, Sewell, Martin, & Vitacco, 2003). The development of specific scales to detect respondents’ problematic test taking approaches is an important feature of the test (Timbrook, Graham, Keiller, & Watts, 1993) and has likely contributed to its wide-spread use across settings. The efficacy of the validity scales of the MMPI-2 in detecting both over- and under-reporting response biases has been studied extensively (Baer & Miller, 2002). Three primary research designs have been employed in this research: known groups, groups with differential prevalence rates, or condition simulation (also known as analog design).

Known-groups designs divide participants in a study into groups of ‘known’ over or under-reporters and those who are ‘known’ to be representing themselves honestly (for examples, see: Edens, Poythress, Watkins-Clay, 2007; Kucharski, Johnsen, & Procell, 2004; or McCusker, Moran, Serfass, & Peterson, 2003). The methods of determining who is reporting in a misleading manner may vary widely (from self-report scales to individual interviews; Edens, Poythress, Watkins-Clay, 2007). The validity scale scores of the known groups are compared to see whether they can differentiate between them. Because these studies use naturally occurring groups, their results may have more external validity than those generated by other designs. However, the validity of the
findings is dependent upon the accuracy of the method that researchers use to determine feigning (Baer & Miller, 2002).

Differential-prevalence designs operate on a principle of different motivational circumstances. Scores on validity scales are compared between groups of persons who are believed to have strong incentives for responding in a misleading fashion, and those who do not seem to have such motives. The major limitation of differential-prevalence designs is that they cannot yield precise information about the classification accuracy of validity scales (Baer & Miller, 2002), as it is impossible to determine whether a particular individual is over-reporting (or under-reporting) or not.

The simulation design (also known as a condition-simulation or analog design) is most frequently used. In this design, the scores of a group that has been instructed to present themselves in a misleading manner on the MMPI-2 are compared to the scores of a group who has taken the test under standard instructions. In some studies the two groups are separate, each group taking the test only once (between-groups design); in other studies the same participants take the test twice with different instructions each time (within-groups design). The equivalence of these two methods has not been established consistently. Baer, Wetter and Berry (1992) found no difference in effect sizes in studies using a between vs. those using a within groups design. However, Baer and Miller (2002) found larger effects sizes in studies that used within-groups than those that used between group designs. Both within and between-groups designs have been useful in identifying participants responding in a misleading manner; but there is some question of the generalizability of simulation studies (Alloy, Abramson, Raniere, &
Dyller, 1999, as it is unclear how closely the responses of those who are instructed to respond in a misleading fashion match their real-world counterparts (Baer & Miller, 2002).

MMPI-2 Validity Scales that Detect Over-reporting

Rogers, Sewell, Martin, and Vitacco (2003) conducted a meta-analysis of over-reporting studies that included those who faked psychopathology versus real patients. They found that three MMPI-2 scales, Infrequency (F), Back Infrequency (FB) and Infrequency Psychopathology (FP) could adequately make this distinction (effect sizes, 2.21, 1.62, and 1.90, respectively). These scales are the focus of the over-reporting analyses of the present investigation.

The Infrequency or F scale was created to detect when a test taker responded in a deviant or atypical way (Meehl & Hathaway, 1946). The scale is made up of 60 items endorsed infrequently by the original sample on which the MMPI was normed (Butcher, et al. 2001). Although originally thought to tap into random responding by a test-taker, it is also sensitive to psychological disturbance and over-reporting (Graham, 2006).

Because all of the F scale items appear in the first 370 items of the test, the F scale is not able to assess whether the test-taker’s approach changes on the later items on the test. To detect the validity of responses to items appearing in the last half of the test booklet, the Back Infrequency (FB) scale was created (Butcher et al, 2001). FB is similar to the F scale, both in design sensitivity to severe psychopathology as well as over-reporting of symptoms (Butcher et al, 2001).
Arbisi and Ben-Porath (1995) developed the Infrequency Psychopathology ($F_P$) scale with the goal that it be less confounded with psychopathology than the $F$ and $F_B$ scales. This goal was accomplished in that the new scale also is less correlated with measures of psychopathology and distress than the other two scales. Arbisi and Ben-Porath (1995) reported that $F_P$ adds incrementally to $F$ and $F_B$ in the discrimination between psychiatric patients and individuals instructed to over report psychopathology. Tolin and colleagues (2004) reported that $F_P$ is less sensitive to genuine psychopathology than are other over-reporting indices and suggested that it may have greater utility in assessing PTSD in veterans.

MMPI-2 Validity Scales that Detect Under-reporting

Early in the construction of the MMPI, Hathaway created a scale, L, which was designed to identify test takers who claimed positive characteristics they either did not have or had to a lesser extent than they claimed (Butcher et al., 2001). In a meta-analysis in 2002, Baer and Miller examined the ability of the MMPI-2 to differentiate between under-reporters vs. persons taking the test under standard instructions. They found effect sizes that were significant but smaller than those found by Rogers et al. (2003) with regard to over-reporting. Three of these scales that could adequately make the discrimination were: the Superlative (S) scale ($d=1.51$), the Lie (L) scale, ($d=1.19$), and the Correction (K) scale ($d=1.13$). These are the under-reporting indicators that will be examined in this study.
Each under-reporting scale was developed to achieve different goals. The L scale was developed to detect when a test taker is engaging in a deliberate attempt to portray him or herself in unrealistically positive light (Meehl & Hathaway, 1946). The items’ content deals with minor personality flaws and weaknesses that most people are willing to admit (Graham, 2006). Meehl and Hathaway (1946) observed that the L scale is a “trap” for naïve subjects, but can easily be avoided by more sophisticated ones.

Hathaway and Meehl (1948) noticed that sometimes a person with significant psychopathology did not register as psychologically disturbed (or appear to be guarded as measured by the L scale) when taking the MMPI. They developed the Correction or K scale as a more subtle measure of defensiveness and to be used as a correction factor for the amount of under-reporting in a protocol (McKinley, Hathaway & Meehl, 1948). Individuals who produce elevated K scores are less likely to report significant psychological problems on the MMPI-2 (Butcher, et al., 2001).

The S scale was developed to detect when a person is trying to appear highly virtuous and responsible. Butcher and Han (1995) designed this scale to identify well-educated individuals who are trying to appear as if they have a sound mind, strong moral values and great capacity to work effectively with others, to be devoid of problems and to be superior in terms of their mental adjustment.

Computerized Assessment

Computerized measures are included in many of the assessment batteries performed in a range of mental health settings today (Butcher, Perry & Hahn, 2004; Hile
There are many advantages of computerization, including efficiency, accuracy, and reliability (Butcher, Keller, & Bacon, 1985). Test takers may find using a computer more interesting than the paper-and-pencil procedure, and they are more motivated to complete the task (Pinsoneault, 1996).

There are two ways in which the MMPI-2 can be administered by computer: conventionally and adaptively. The conventional version of the test parallels how most personality and psychopathology instruments are administered (Forbey & Ben-Porath, 2007). In this version, the test items are given in the traditional order, and all 567 items are administered to the test taker. Although this format significantly reduces administration time when compared with paper-and-pencil administrations, it does not take full advantage of computers’ test-administration technology (Forbey & Ben-Porath, 2007).

In adaptive testing, different items are administered to different individuals depending on the test taker’s standing on the measured trait (Weiss, 1985). Adaptive testing allows a psychologist to obtain an optimal amount of information at a minimal cost in terms of administration time (Roper, Ben-Porath, & Butcher, 1991). There are many forms of adaptive testing, but the most widely used and studied are those based on Item Response Theory (IRT). However, IRT is not appropriate for most MMPI-2 scales. One assumption of IRT is that the trait being measured is one-dimensional (Ben-Porath & Butcher, 1986). This may make IRT incompatible with personality testing, where the constructs are usually heterogeneous (Roper, Ben-Porath, Butcher, 1995). Recent studies attempting to use IRT in personality testing have found substantial savings in the number
of items that need to be administered (Simms & Clark, 2005). However, one study showed this method to decrease the variance of the latent trait scores significantly (Reise & Henson, 2000) whereas another showed significant loss of information (Simms & Clark, 2005).

Because of the limitations of IRT with regard to personality testing, Butcher, Keller, and Bacon (1985) proposed using a countdown method when administering MMPI-2 scales adaptively. The countdown method uses a cutoff criterion, which typically corresponds to the raw score on a scale that is equal to a clinical elevation on that scale (Forbey & Ben-Porath, 2007). In this method, items on a scale are administered until either the number of unendorsed items equals the total number of items on a particular scale minus the raw score required to be elevated + 1, or the individual endorses enough items for their score to be classified as elevated. This method has been found to be most efficient for individuals who are at the extremes of the trait being assessed, and it can lead to substantial items savings (Ben-Porath, Slutske, & Butcher, 1989). To further increase item savings, the items on these scales are administered from those that are least endorsed to those that are most endorsed (Ben-Porath & Butcher, 1986).

Two varieties of the countdown method have been researched: the classification method and the Full Score on Elevated Scales (FSES) method. When using the classification method, administration of items on a scale is stopped once a designated level of elevation on that scale is either ruled in or out. This means that items are only administered until the cutoff for that scale is reached, or it is impossible to reach it (Ben-
Porath, Slutske, & Butcher, 1989). For the FSES method, item administration is terminated only if it is impossible to reach a cut off criterion. If elevation on the scale is reached, all items are administered on that scale so that a full score is obtained (Ben-Porath, Slutske, & Butcher, 1989).

There are many advantages to administering the MMPI-2 via computer, but it is necessary to determine whether MMPI-2-CA and MMPI-2-CC results are comparable to the booklet version. Research has shown both adaptive administration of the MMPI-2 and conventional computerized administration to be comparable to standard booklet administrations of the test (Roper, Ben-Porath, & Butcher, 1991). Meta-analysis indicates there is very little difference in how participants respond on computerized vs. paper-and-pencil versions of the MMPI-2 (Finger & Ones, 1999). The research in this field led Graham (2006) to conclude that the differences between conventional computerized and standard administrations of the MMPI-2 are probably small and not clinically important.

Research in this field has also demonstrated advantages of the countdown method. Ben-Porath, Slutske, and Butcher (1989) conducted a real-data simulation using four archival samples. They found that the countdown method could result in a twenty to thirty-one percent (19.8 -31.3%) reduction in the number of items that were administered. Roper, Ben-Porath, and Butcher (1991) administered two validity scales ten clinical scales and fifteen content scales adaptively. The results of adaptive administration were very similar to the results obtained by booklet administration. Furthermore, Handel and colleagues found that there was a 31.4% savings in the number of items that were
administered with the computer adaptive (CA) version (1999). Forbey and Ben-Porath (2007) reported a large amount of item savings using the classification and FSES method over the conventional computerized administration of the MMPI-2; however, there was only a medium effect for the difference in number of items administered in the FSES administration over the classification administration. An average of only 2.32 minutes was saved by administering the classification method over the FSES method.

There have also been studies showing that the computerized versions of the MMPI-2 display the same degree of criterion validity as the booklet version. Roper, Ben-Porath and Butcher (1995) found that correlations with criterion measures were not significantly different for adaptive, conventional computerized and booklet administrations. Handel, Ben-Porath, and Watt (1999) demonstrated that the computerized adaptive version of the MMPI-2 was correlated with criterion measures in much the same way as conventional administrations with slight attenuations between Scales 1 and 3 and extratest criteria. Forbey and Ben-Porath (2007) reported no consistent differences in correlations between MMPI-2 scales and relevant criteria when comparing the results of conventional and adaptive administration of the test.

Effects of over- and under-reporting on substantive scales

The major concern with over- and under-reporting is that these test-taking approaches will affect scores on substantive scales, leading to interpretations that are not consonant with the test-taker’s functioning. The impact of these approaches on scores on the substantive scales of the MMPI-2 was examined with the Restructured Clinical
(RC) Scales (Tellegen, Ben-Porath, McNulty, Arbisi., Graham, et. al, 2003). These scales were chosen because they are more transparent than the original Clinical Scales, and therefore can better serve as gauges of the impact of over- and under-reporting on the substantive scales of the test.

As described above, a computerized adaptive administration of the MMPI-2 changes the order of item administration. Previous studies have established that this method results in scores on substantive scales that are not statistically different than those obtained using conventional administration (for example, Roper, Ben-Porath & Butcher, 1991). However, to date no studies have examined computerized adaptive MMPI-2 administration in the context of detecting the effects of misleading responding. The goal of the current investigation was to compare the ability of the MMPI-2 Validity Scales to identify over- and under-reporting under conventional and adaptive administration.

The following research questions are addressed in this study:

1) Do adaptively-administered validity scales function as well as conventionally-administered ones in detecting over- and under-reporting?
2) When test-takers are instructed to over or under-report their symptomatology, does administration method differentially affect their scores on the MMPI-2 RC Scales?
CHAPTER 2

Methods

Participants

The original sample consisted of 968 (498 male and 470 female) undergraduate college students. The students volunteered to participate in the study in order to fulfill requirements in a general psychology course. The volunteers had an average of 12.99 (SD=1.10) years of education. The student ranged in age from 18 to 65 and the mean was 19.77 (4.78) years old. Caucasians comprised 86.3% of the sample, whereas 8.8% was African-American, 2.6% was Hispanic, 1.2% was Native American, 2.1% was Asian, .7% was Native Hawaiian or Pacific Islander, and 2.3% self-identified as other. Participants were permitted to identify with more than one ethnic group.

Procedures

Students were assigned randomly to one of six groups, Conventional Computer Administration with Over-reporting Instruction, Conventional Computer Administration with Under-reporting Instruction, Conventional Computer Administration with Standard Instruction, Adaptive Computer Administration with Over-reporting Instruction, Adaptive Computer Administration with Under-reporting Instruction, and Adaptive Computer Administration with Standard Instruction. Under adaptive administration conditions, all scales were administered as Full Scores on Elevated Scales (FSES) to
minimize range restriction. Although all six groups were collected simultaneously as part of this study, at no time are all six groups included in a single analysis. Instead, the analyses were broken down into over-reporters versus those taking the test under standard instruction and under-reporters versus those taking the test under standard instruction. For further clarification, see below. The standard instructions given were the ones that are printed in the front of the test booklet. The over- and under-reporting instructions were the same as those used by Graham, Watts, & Timbrook (1991). The students in the over-reporting conditions were given the following instructions:

This is the Minnesota Multiphasic Personality Inventory -2 (MMPI-2). In responding to the items in the inventory, please try to answer in the way that you think one would answer if he or she wanted to give the impression of being a person who has very serious psychological or emotional problems.

The students in the under-reporting condition were given these instructions:

This is the Minnesota Multiphasic Personality Inventory – 2 (MMPI-2). In responding to the items in the inventory, please try to imagine that you are graduating from college, are being assessed for a highly desirable job, and for that reason are trying to appear very well adjusted.

A post-test questionnaire was administered to every participant immediately after taking the MMPI-2-CA or MMPI-2-CC for two reasons. First, because results can be affected when participants engage in random or non-content based responding during validity scale studies (Rogers, 1997) it is important to assess the degree of random responding in which a participant engages. Second, when participants do not follow the instructions (Gallagher, 1998), results may be distorted with respect to the ability of the validity scales in detecting over- and under-reporting. The post-test questionnaire that was used was adapted from the one employed by Gallagher (1998) and designed to assess
for random responding and compliance with instructions. A copy of this questionnaire is included in Appendix A.

Analyses

There were no missing values on any variable in these analyses. This is because data for each participant in this study were collected at one time point, and participants were encouraged to complete both measures fully before leaving. In determining whether administration format for the MMPI2-CA affects how the validity scales detect over-reporting, three scales (F, Fb, and Fp) were investigated to see whether they were affected by administration (conventional versus adaptive). For analyses looking at these scales, only four of the six groups were investigated: conventional and adaptively administered tests given under both over-reporting and standard instruction conditions. For the under-reporting analyses, three scales (L, K, and S) were investigated to assess whether administration type affects scores on the validity scales. These analyses only included four of the six groups: both under-reporting and standard instruction groups who either took the test adaptively or conventionally. All scales were within skewness and kurtosis standards (as set forth in Tabachnick & Fidell, 2001; p. 74). Due to the purpose of this study, no participant was excluded based on their scores on F, Fb, Fp, L, K or S scales. However, conventionally-administered protocols with a Cannot Say Score at or above 30 were excluded as large numbers of unanswered responses can lead to artificially deflated scores on the scales of interest (Graham, 2006). Protocols having a VRIN (Variable Response Inconsistency) or TRIN (True Response Inconsistency) score at or above 80T (indicating random or fixed responding as opposed to responding to the
content of the items on the test) were also excluded. Seventy-eight individuals (66 male and 12 females) were excluded on this basis. Participants who indicated that they did not follow the instructions given to them about how to approach the test (as assessed by the first item of the post-test questionnaire) were excluded as well. Two hundred and six individuals were excluded because they answered the question “What instructions did you receive for completing the MMPI-2?” in such a way as to indicate they did not follow the instructions given to them at the beginning of the testing session. Analyses were run to determine whether these individuals did actually respond differently than those who reported following the instructions. Results indicated that non-compliant participants responded differently than those who followed instructions. The non-compliant participants were excluded from further analyses.

The final sample consisted of 684 (333 male and 351 female) undergraduate college students. The participants had an average of 13.01 (3.87) years of education. The students ranged in age from 18 to 65 and the mean was 19.56 (1.12) years old. The sample was 87.3% Caucasian, 8% African-American, 2.6% Hispanic, 1.2% Native American, 1.9% Asian, .9% Native Hawaiian or Pacific Islander, and 2% of the sample self-identified as other. No significant differences were found on demographic variables between those clients included and those excluded from the study.
CHAPTER 3

Results

The Effect of Administration Mode on Validity Scales in Detecting Over- and Under-Reporting

The first research question addressed in this study examined whether adaptively-administered validity scales function as well as conventionally-administered ones in detecting over- and under-reporting. To address this question, two by two ANOVAs were conducted to determine whether the type of administration affected scale scores differently based on whether the participants took the test under standard or modified (over- or under-reporting) instruction. Next, $t$ tests were conducted to examine the effects of administration and instruction. The three subject groups consisted of students taking the test under instructions to over-report psychological and/or emotional problems, students taking the test under instructions to under-report; that is to try to appear very well-adjusted, and students taking the test under standard instructions. Analyses compared the over-reporters with the standard instruction group and the under-reporters with the standard instruction group by administration type (conventional or adaptive). Analyses were conducted separately for men and women to ascertain whether results could be replicated across gender. Bonferroni corrections were applied within groups of scales to control for family-wise error (three scales F, Fb, and Fp, for over-reporting
analyses; and three L, K and S, for under-reporting analyses) with alpha set at .05 - \( \frac{.05}{3} = .017. \)

**Over-reporters vs. standard instructions.**

T score means and standard deviations on the over-reporting indicators (F, Fb and Fp) as well as effect sizes for the mean score comparisons are reported by administration and instruction type for men in *Table 1* and for women in *Table 2*. Two (administration: conventional vs. adaptive) by two (instruction: over-reporting vs. standard) between-participants ANOVAs were conducted to test interactions between administration and instruction types on the over-reporting indicators. No statistically significant interaction effects (reported as \( \eta^2 \)) were found on any of the scales for either men or women. For each of the over-reporting indicators, \( t \) tests were conducted to compare the effects of administration (adaptive vs. conventional) on over-reporting and standard instruction groups. Next, \( t \) tests were conducted to compare the effects of instruction (over-reporting vs. standard instruction) on adaptively and conventionally administered tests. All \( t \) tests were conducted separately by gender.

As can be seen by looking across the first row in *Table 1*, men in both over-reporting groups (adaptive = 112.30[21.32] and conventional = 110.19[21.20]) produced much higher scores on the F scale than the standard administration groups (adaptive = 57.23[15.62] and conventional = 60.96[15.07]). The effect sizes in that table indicate that mean scores were not significantly different for administration type (conventional vs. adaptive) for either over-reporting or standard instruction men on the F scale, indicating
Table 1. Over-reporting vs. Standard Instruction Men: Means, Standard Deviations and Effect Sizes

<table>
<thead>
<tr>
<th>Scale</th>
<th>Over-reporting</th>
<th>Standard</th>
<th>Con. vs. Adp.</th>
<th>Std vs. Over-</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Adp. (N=54)</td>
<td>Con. (N=54)</td>
<td>Adp. (N=82)</td>
<td>Con. (N=69)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>112.30(21.32)</td>
<td>110.19(21.20)</td>
<td>57.23(15.62)</td>
<td>60.96(15.07)</td>
</tr>
<tr>
<td></td>
<td>110.19(21.20)</td>
<td>112.30(21.32)</td>
<td>60.96(15.07)</td>
<td>57.23(15.62)</td>
</tr>
<tr>
<td>Fb</td>
<td>112.59(20.99)</td>
<td>112.41(18.34)</td>
<td>55.50(16.01)</td>
<td>61.17(19.12)</td>
</tr>
<tr>
<td></td>
<td>112.41(18.34)</td>
<td>112.59(20.99)</td>
<td>61.17(19.12)</td>
<td>55.50(16.01)</td>
</tr>
<tr>
<td>Fp</td>
<td>111.46(21.91)</td>
<td>108.50(22.86)</td>
<td>58.91(16.81)</td>
<td>61.77(14.63)</td>
</tr>
<tr>
<td></td>
<td>108.50(22.86)</td>
<td>111.46(21.91)</td>
<td>61.77(14.63)</td>
<td>58.91(16.81)</td>
</tr>
</tbody>
</table>

*<p< .017

*a Interaction effect sizes are reported as $\eta^2$ all others are Cohen’s $d$

Table 2. **Over-reporting vs. Standard Instruction Women: Means, Standard Deviations and Effect Sizes**

<table>
<thead>
<tr>
<th>Scale</th>
<th>Over-reporting</th>
<th>Standard</th>
<th>Con. vs. Adp.</th>
<th>Std vs. Over-</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>116.03(14.56)</td>
<td>113.40(16.77)</td>
<td>62.11(18.00)</td>
<td>58.85(17.30)</td>
</tr>
<tr>
<td>Fb</td>
<td>115.61(16.41)</td>
<td>114.82(15.01)</td>
<td>61.40(18.48)</td>
<td>58.23(17.45)</td>
</tr>
<tr>
<td>Fp</td>
<td>115.43(15.89)</td>
<td>111.34(20.34)</td>
<td>63.22(18.83)</td>
<td>59.38(15.85)</td>
</tr>
</tbody>
</table>

*p< .017

$^a$ Interaction effect sizes are reported as $\eta^2$ all others are Cohen’s $d$

that the scale performed similarly regardless of how the test was administered. However, the mean scores were significantly different across instruction type, with men who were instructed to over-report scoring statistically significantly higher than men who were given standard instructions, for both those who took the test adaptively ($t = 17.38$, $df = 134$, $p < .017$, $d = 3.03$) and conventionally ($t = 15.04$, $df = 121$, $p < .017$, $d = 2.72$). Similar results were obtained on the other over-reporting indicators (see the effect sizes in Table 1). The findings were replicated in the analyses that examine the effect of administration type and instruction for women (Table 2).

*Under-reporters vs. standard instructions.*

Effect sizes for the mean score comparisons, as well as means and standard deviations on the under-reporting indicators (L, K and S) are reported by administration and instruction type in Table 3 for men and Table 4 for women. Two (administration: conventional vs. adaptive) by two (instruction: under-reporting vs. standard) between-participants ANOVAs were conducted to test interactions between administration and instruction types on the under-reporting indicators. No statistically significant interaction effects were found on any of the scales for either men or women. For each of the under-reporting indicators, $t$ tests were conducted to compare the effects of administration (adaptive vs. conventional) on under-reporting and standard instruction groups. Next, $t$ tests were conducted to compare the effects of instruction (under-reporting vs. standard instruction) on adaptively and conventionally administered tests. All $t$ tests were conducted separately by gender.
<table>
<thead>
<tr>
<th>Scale</th>
<th>Under-reporting</th>
<th>Standard</th>
<th>Effect Sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Adp. (N=46)</td>
<td>Con. (N=28)</td>
<td>Adp. (N=82)</td>
</tr>
<tr>
<td>L</td>
<td>65.57(19.09)</td>
<td>62.75(13.21)</td>
<td>49.28(10.65)</td>
</tr>
<tr>
<td></td>
<td>&lt;.01</td>
<td>.16</td>
<td>.04</td>
</tr>
<tr>
<td>K</td>
<td>51.35(12.11)</td>
<td>54.39(9.01)</td>
<td>42.56(6.83)</td>
</tr>
<tr>
<td></td>
<td>&lt;.01</td>
<td>.27</td>
<td>.22</td>
</tr>
<tr>
<td>S</td>
<td>53.20(14.32)</td>
<td>57.54(9.97)</td>
<td>41.27(5.97)</td>
</tr>
<tr>
<td></td>
<td>&lt;.01</td>
<td>.33</td>
<td>.56*</td>
</tr>
</tbody>
</table>

*p< .017

* Interaction effect sizes are reported as $\eta^2$ all others are Cohen’s $d$

### Table 4. Under-reporting vs. Standard Instruction Women: Means, Standard Deviations and Effect Sizes

<table>
<thead>
<tr>
<th>Scale</th>
<th>Under-reporting</th>
<th>Standard</th>
<th>Con. vs. Adp.</th>
<th>Std vs. Over-</th>
</tr>
</thead>
<tbody>
<tr>
<td>L</td>
<td>72.28(16.69)</td>
<td>75.59(16.88)</td>
<td>48.65(10.13)</td>
<td>48.63(9.93)</td>
</tr>
<tr>
<td>K</td>
<td>58.70(9.21)</td>
<td>59.03(8.35)</td>
<td>44.00(8.98)</td>
<td>47.59(10.54)</td>
</tr>
<tr>
<td>S</td>
<td>60.83(12.25)</td>
<td>63.54(9.46)</td>
<td>41.96(8.15)</td>
<td>48.28(10.81)</td>
</tr>
</tbody>
</table>

\(a\) Interaction effect sizes are reported as \(\eta^2\) all others are Cohen’s \(d\)


\(^*\)p < .017
As can be seen in Table 3, men in both under-reporting groups (adaptive = 65.57[19.09] and conventional = 62.75[13.21]) produced much higher scores on the L scale than did men in the standard administration groups (adaptive = 49.28[10.60] and conventional = 49.65[9.55]). Mean scores were not significantly different for administration type (conventional vs. adaptive) for either under-reporting or standard instruction men on the L scale, indicating that the scale performed similarly regardless of how the test was administered. On the other hand, the mean scores were significantly different across instruction type, with men who were instructed to under-report scoring statistically significantly higher than men who were given standard instructions for both those who took the test adaptively \((t = 6.21, df = 126, p<.017, d=1.14)\) and conventionally \((t = 5.46, df = 95, p<.017, d=1.21)\). Similar results were obtained for both men and women on the other under-reporting indicators. For both the L and K scales, there were no significant differences on any of the under-reporting indicators when adaptive and conventionally administered tests were compared to each other in either the under-reporting or standard instruction conditions. For the S scale, there was no significant difference between adaptive and conventional administration for men and women who were asked to under-report or try to appear very well-adjusted. However, under standard instruction condition, the men who took the test conventionally \((45.59[9.35])\) scored significantly higher than those who took the test adaptively \((41.57[5.97])\). The difference between these two types of administration reached a medium effect size (Cohen’s \(d = .56\)). A similar effect was found among women as well (Cohen’s \(d = .66\)).
Classification accuracy of over-reporting indicators.

Classification accuracies were calculated for the participants who were instructed to over-report their symptomatology versus those who were given standard instructions (see Table 5) in order to further examine whether adaptively-administered validity scales function similarly to conventionally-administered ones in detecting over-reporting. Because similar results were found for men and women on the above analyses, separate classification analyses by gender were not deemed to be necessary. For each of the indicators (F, Fb, and Fp), the ability of the over-reporting indicators to accurately classify over-reporters and members of the standard instruction group was examined at three different levels (T scores $\geq 80$, T scores $\geq 90$, and T scores $\geq 100$). These three levels were chosen as they indicate clinical cutoff points where protocols may be considered to be invalid for nonclinical test-takers, for outpatients, and for inpatients, respectively (Butcher et. al, 2001). For each group, sensitivities, specificities, positive and negative predictive powers (PPP and NPP) and hit rates were calculated.

As can be seen in Table 5, T scores on the F scale $\geq 80$ correctly identified over-reporters (sensitivity) 93% of the time in the adaptive group, and 89% of the time in the conventional group. They identified those who took the test under standard instruction (specificity) 87% of the time in the adaptive group and 91% of the time in the conventional group. Of the adaptive group, 84% of those who were identified by the F
**Table 5. Classification Accuracy and Predictive Power for Over-reporting Indicators**

<table>
<thead>
<tr>
<th>Scale</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPP</th>
<th>NPP</th>
<th>Hit rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>F ≥ 80</td>
<td>.93</td>
<td>.89</td>
<td>.87</td>
<td>.91</td>
<td>.84</td>
</tr>
<tr>
<td>F ≥ 90</td>
<td>.92</td>
<td>.85</td>
<td>.93</td>
<td>.94</td>
<td>.91</td>
</tr>
<tr>
<td>F ≥ 100</td>
<td>.90</td>
<td>.84</td>
<td>.95</td>
<td>.96</td>
<td>.94</td>
</tr>
<tr>
<td>Fb ≥ 80</td>
<td>.93</td>
<td>.91</td>
<td>.88</td>
<td>.88</td>
<td>.85</td>
</tr>
<tr>
<td>Fb ≥ 90</td>
<td>.91</td>
<td>.90</td>
<td>.92</td>
<td>.92</td>
<td>.88</td>
</tr>
<tr>
<td>Fb ≥ 100</td>
<td>.91</td>
<td>.89</td>
<td>.97</td>
<td>.95</td>
<td>.95</td>
</tr>
<tr>
<td>Fp ≥ 80</td>
<td>.92</td>
<td>.86</td>
<td>.85</td>
<td>.89</td>
<td>.82</td>
</tr>
<tr>
<td>Fp ≥ 90</td>
<td>.90</td>
<td>.84</td>
<td>.93</td>
<td>.94</td>
<td>.90</td>
</tr>
<tr>
<td>Fp ≥ 100</td>
<td>.89</td>
<td>.82</td>
<td>.96</td>
<td>.99</td>
<td>.94</td>
</tr>
<tr>
<td>----------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
</tr>
</tbody>
</table>

PPP = Positive Predictive Power; NPP = Negative Predictive Power; Adp. = Adaptive; Con. = Conventional

Adaptive Over-reporting (N = 115), Conventional Over-reporting (N = 122), Adaptive Standard (N = 154), Conventional Standard (N = 140)
scale as over-reporting symptomatology were in the over-reporting group, whereas in the conventional group 90% were identified as over-reporters (PPP). The proportions of participants who were identified as being part of the standard instruction group, and actually were part of that group (NPP) were 94% (adaptive) and 91% (conventional). At $T \geq 80$, F accurately classified 90% of the adaptive group and 90% of the conventional group (base rates .43 and .47, respectively). The results for Fb and Fp are similar to those found for F, in that sensitivity increased and specificity decreased with higher cutoff scores, but overall classification accuracies are comparable across administration types for both of these scales. As expected, sensitivity rates declined and specificity rates rose with higher cutoff scores, but the overall pattern remained the same: adaptively generated over-reporting scales produced comparable classification accuracies when compared with those found for conventionally generated ones.

*Classification accuracy of under-reporting indicators.*

Classification accuracies were calculated for the participants who were instructed to appear more well-adjusted than they actually are versus those who were given standard instructions (see Table 6) in order to further examine whether adaptively-administered validity scales function as well as conventionally-administered ones in detecting under-reporting. Because similar results were found for men and women on the ANOVAs and $t$ tests above, separate classification analyses by gender were not deemed to be necessary. For each of the under-reporting indicators (L, K and S), the ability of the under-reporting indicators to accurately classify under-reporters and members of the standard instruction...
Table 6. Classification Accuracy and Predictive Power for Under-reporting Indicators

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>L ≥55</td>
<td>.73</td>
<td>.78</td>
<td>.71</td>
<td>.74</td>
<td>.59</td>
<td>.59</td>
<td>.83</td>
<td>.87</td>
<td>.72</td>
<td>.75</td>
</tr>
<tr>
<td>L ≥60</td>
<td>.63</td>
<td>.76</td>
<td>.84</td>
<td>.84</td>
<td>.68</td>
<td>.69</td>
<td>.80</td>
<td>.88</td>
<td>.76</td>
<td>.81</td>
</tr>
<tr>
<td>L ≥65</td>
<td>.58</td>
<td>.64</td>
<td>.90</td>
<td>.92</td>
<td>.76</td>
<td>.80</td>
<td>.79</td>
<td>.84</td>
<td>.78</td>
<td>.83</td>
</tr>
<tr>
<td>L ≥70</td>
<td>.52</td>
<td>.54</td>
<td>.95</td>
<td>.95</td>
<td>.87</td>
<td>.84</td>
<td>.78</td>
<td>.81</td>
<td>.80</td>
<td>.82</td>
</tr>
<tr>
<td>L ≥75</td>
<td>.45</td>
<td>.43</td>
<td>.99</td>
<td>.98</td>
<td>.98</td>
<td>.91</td>
<td>.77</td>
<td>.78</td>
<td>.80</td>
<td>.80</td>
</tr>
<tr>
<td>L ≥80</td>
<td>.37</td>
<td>.36</td>
<td>1.0</td>
<td>.99</td>
<td>1.0</td>
<td>.96</td>
<td>.74</td>
<td>.76</td>
<td>.78</td>
<td>.79</td>
</tr>
<tr>
<td>K ≥55</td>
<td>.56</td>
<td>.66</td>
<td>.92</td>
<td>.80</td>
<td>.79</td>
<td>.61</td>
<td>.79</td>
<td>.83</td>
<td>.79</td>
<td>.75</td>
</tr>
<tr>
<td>K ≥60</td>
<td>.44</td>
<td>.42</td>
<td>.97</td>
<td>.90</td>
<td>.88</td>
<td>.67</td>
<td>.76</td>
<td>.76</td>
<td>.78</td>
<td>.74</td>
</tr>
<tr>
<td>K ≥65</td>
<td>.22</td>
<td>.21</td>
<td>.99</td>
<td>.96</td>
<td>.90</td>
<td>.74</td>
<td>.69</td>
<td>.72</td>
<td>.71</td>
<td>.72</td>
</tr>
<tr>
<td></td>
<td>PPP</td>
<td>NPP</td>
<td>Adp.</td>
<td>Con.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----</td>
<td>-----</td>
<td>-----</td>
<td>------</td>
<td>------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K ≥ 70</td>
<td>.05</td>
<td>.06</td>
<td>1.0</td>
<td>.99</td>
<td>1.0</td>
<td>.80</td>
<td>.65</td>
<td>.69</td>
<td>.66</td>
<td>.69</td>
</tr>
<tr>
<td>K ≥ 75</td>
<td>.00 (^a)</td>
<td>.00 (^a)</td>
<td>1.0</td>
<td>1.0</td>
<td>-(^a)</td>
<td>-(^a)</td>
<td>.64</td>
<td>.68</td>
<td>.64</td>
<td>.68</td>
</tr>
<tr>
<td>K ≥ 80</td>
<td>.00 (^a)</td>
<td>.00 (^a)</td>
<td>1.0</td>
<td>1.0</td>
<td>-(^a)</td>
<td>-(^a)</td>
<td>.64</td>
<td>.68</td>
<td>.64</td>
<td>.68</td>
</tr>
<tr>
<td>S ≥ 55</td>
<td>.52</td>
<td>.72</td>
<td>.98</td>
<td>.78</td>
<td>.94</td>
<td>.61</td>
<td>.79</td>
<td>.85</td>
<td>.82</td>
<td>.76</td>
</tr>
<tr>
<td>S ≥ 60</td>
<td>.48</td>
<td>.57</td>
<td>.98</td>
<td>.86</td>
<td>.93</td>
<td>.66</td>
<td>.77</td>
<td>.81</td>
<td>.80</td>
<td>.76</td>
</tr>
<tr>
<td>S ≥ 65</td>
<td>.44</td>
<td>.46</td>
<td>.99</td>
<td>.93</td>
<td>.95</td>
<td>.76</td>
<td>.76</td>
<td>.78</td>
<td>.79</td>
<td>.78</td>
</tr>
<tr>
<td>S ≥ 70</td>
<td>.21</td>
<td>.22</td>
<td>.99</td>
<td>1.0</td>
<td>.95</td>
<td>1.0</td>
<td>.69</td>
<td>.73</td>
<td>.71</td>
<td>.75</td>
</tr>
<tr>
<td>S ≥ 75</td>
<td>.06</td>
<td>.07</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>.66</td>
<td>.70</td>
<td>.67</td>
<td>.70</td>
</tr>
<tr>
<td>S ≥ 80</td>
<td>.00 (^a)</td>
<td>.00 (^a)</td>
<td>1.0</td>
<td>1.0</td>
<td>-(^a)</td>
<td>-(^a)</td>
<td>.64</td>
<td>.68</td>
<td>.64</td>
<td>.68</td>
</tr>
</tbody>
</table>

PPP = Positive Predictive Power; NPP = Negative Predictive Power; Adp. = Adaptive; Con. = Conventional

\(^a\) No participants scored above the cutoff.

Adaptive Under-reporting (N = 86), Conventional Under-reporting (N = 67) Adaptive Standard (N = 154),

Conventional Standard (N = 140)
group was examined at six different levels (T scores $\geq 55$, T scores $\geq 60$, T scores $\geq 65$, T scores $\geq 70$, T scores $\geq 75$, and T scores $\geq 80$). Levels of 65, 70, 75 and 80T were examined, as they indicate clinical cutoff points which affect interpretation of the validity of the test protocol (Butcher et. al, 2001). Lower than customary cutoffs (55T and 60 T) were examined because of the relatively low classification accuracy rates found with the higher cutoffs. For each group, sensitivities, specificities, positive and negative predictive powers (PPP and NPP) and hit rates were calculated.

As can be seen in Table 6, scores on the L scale $\geq 55$ correctly identified under-reporters (sensitivity) 73% of the time in the adaptive group, and 78% of the time in the conventional group. This score identified those who took the test under standard instruction (specificity) 71% of the time in the adaptive group and 74% of the time in the conventional group. For both the adaptive and conventional groups, 59% of those who were identified by the L scale as under-reporting symptomotology were in the under-reporting group (PPP). The proportion of participants who were identified being part of the standard instruction group, and actually were part of that group (NPP) was 83% (adaptive) and 87% (conventional). At T $\geq 55$, L accurately classified 72% of the adaptive group and 75% of the conventional group (base rates .36 and .32, respectively). For K, as with L, lower cutoff scores were more effective (had a higher hit rate) than higher cutoff scores. However, at T $\geq 55$, K performed slightly better in the adaptive sample than in the conventional sample. No one scored at or above 75T in either sample on this scale. The hit rate of the S scale showed that the adaptive sample performed slightly better at lower cutoffs (T $\geq 55$ and 60), but there were insufficient high scorers on
S to make comparisons at higher cutoffs meaningful. For under-reporting scales, at higher cutoffs, sensitivity rates declined and specificity rates rose. However, the pattern overall remained the same: adaptively generated scales had comparable classification accuracy when contrasted with conventionally generated ones, although at certain cutoffs the classification accuracy rates were slightly higher for the conventionally administered scales. Overall, the analyses show that mode of administration had little impact on the indicators’ abilities to correctly classify the participants in both over- and under-reporting conditions.

The Effect of Administration Mode and Instructions on the RC Scales

The second research question examined whether administration mode differentially affects scores on the MMPI-2 RC Scales when test-takers are instructed to over or under-report their symptomatology. To address this question, two by two ANOVAs were conducted to determine whether the type of administration affected RC scale scores differently based on whether the participants took the test under standard or modified (over- or under-reporting) instruction. Next, t tests were conducted to examine the effects of type of administration on these scales. Analyses compared members of each instruction group who took the test adaptively with those who took the test conventionally. The over-reporters were also compared with the standard instruction group and the under-reporters with the standard instruction group by administration type (conventional or adaptive). Analyses were also conducted separately for men and women to ascertain whether results could be replicated across gender. Bonferroni
corrections were applied within the RC scales to control for family-wise error with alpha set at .006 (.05/9).

Mean Restructured Clinical Scale comparisons across over-reporting groups.

T score means and standard deviations on the RC scales for the over-reporting groups as well as effect sizes for the mean score comparisons are reported for men in Table 7 and for women in Table 8. For each RC scale, two (administration: conventional vs. adaptive) by two (instruction: over-reporting vs. standard) between-participants ANOVAs were conducted to test interactions between administration and instruction types on the RC Scales. No statistically significant interaction effects were found on any of the scales for either men or women. Next, t tests were conducted to compare the effects of administration (adaptive vs. conventional) on the RC Scale scores of over-reporting and standard instruction men and women. All t tests were conducted separately by gender. In addition, t tests were conducted to compare how instruction (over-reporting vs. standard instruction) affected adaptively and conventionally generated scales.

As can be seen in Table 7, men in both over-reporting groups (adaptive = 80.33[10.15]) and conventional = 79.70 [10.80]) produced elevated scores on the RCd scale. Mean scores were not significantly different for administration type (conventional vs. adaptive) on RCd, indicating that the participants answered similarly independently of
Table 7. **RC Scales - Over-reporting vs. Standard Instruction Men: Means, Standard Deviations and Effect Sizes**

<table>
<thead>
<tr>
<th>Scale</th>
<th>Over-reporting</th>
<th>Standard</th>
<th>Effect Sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td>RCd</td>
<td></td>
<td>80.33 (10.15)</td>
<td>79.70 (10.80)</td>
</tr>
<tr>
<td>RC1</td>
<td></td>
<td>88.76 (16.09)</td>
<td>87.87 (15.91)</td>
</tr>
<tr>
<td>RC2</td>
<td></td>
<td>75.74 (16.19)</td>
<td>74.96 (17.44)</td>
</tr>
<tr>
<td>RC3</td>
<td></td>
<td>71.98 (10.64)</td>
<td>70.96 (10.74)</td>
</tr>
<tr>
<td>RC4</td>
<td></td>
<td>84.19 (15.42)</td>
<td>85.00 (13.92)</td>
</tr>
<tr>
<td>RC6</td>
<td></td>
<td>91.85 (14.56)</td>
<td>90.98 (14.39)</td>
</tr>
<tr>
<td>RC7</td>
<td></td>
<td>82.98 (10.89)</td>
<td>81.54 (12.16)</td>
</tr>
<tr>
<td>RC8</td>
<td></td>
<td>92.80 (12.18)</td>
<td>90.59 (15.55)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>------</td>
<td>------</td>
<td>------</td>
</tr>
<tr>
<td>RC9</td>
<td>67.85 (10.24)</td>
<td>68.13 (11.77)</td>
<td>56.46 (11.94)</td>
</tr>
</tbody>
</table>

*p< .006  

*a Interaction effect sizes are reported as $\eta^2$ all others are Cohen’s $d$

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>RCd</td>
<td>78.90 (8.89)</td>
<td>78.99 (8.34)</td>
<td>54.94 (11.32)</td>
<td>52.51 (11.52)</td>
<td>&lt;.01</td>
<td>.01</td>
<td>.21</td>
<td>2.30*</td>
<td>2.59*</td>
</tr>
<tr>
<td>RC1</td>
<td>87.74 (15.43)</td>
<td>87.47 (13.27)</td>
<td>57.86 (13.22)</td>
<td>52.51 (11.30)</td>
<td>&lt;.01</td>
<td>.02</td>
<td>.43</td>
<td>2.07*</td>
<td>2.81*</td>
</tr>
<tr>
<td>RC2</td>
<td>88.25 (14.86)</td>
<td>84.04 (15.25)</td>
<td>48.81 (11.13)</td>
<td>46.80 (9.86)</td>
<td>&lt;.01</td>
<td>.28</td>
<td>.19</td>
<td>3.00*</td>
<td>2.88*</td>
</tr>
<tr>
<td>RC3</td>
<td>73.87 (9.38)</td>
<td>72.54 (9.05)</td>
<td>54.42 (9.34)</td>
<td>54.69 (9.63)</td>
<td>&lt;.01</td>
<td>.14</td>
<td>.03</td>
<td>2.05*</td>
<td>1.88*</td>
</tr>
<tr>
<td>RC4</td>
<td>91.95 (13.21)</td>
<td>88.93 (15.65)</td>
<td>56.11 (12.01)</td>
<td>53.85 (12.60)</td>
<td>&lt;.01</td>
<td>.20</td>
<td>.18</td>
<td>281*</td>
<td>2.44*</td>
</tr>
<tr>
<td>RC6</td>
<td>94.97 (11.71)</td>
<td>94.91 (10.60)</td>
<td>61.04 (10.23)</td>
<td>54.96 (12.31)</td>
<td>.01</td>
<td>.01</td>
<td>.53*</td>
<td>3.06*</td>
<td>3.43*</td>
</tr>
<tr>
<td>RC7</td>
<td>81.36 (10.94)</td>
<td>80.91 (11.56)</td>
<td>55.33 (12.21)</td>
<td>52.20 (12.95)</td>
<td>&lt;.01</td>
<td>.04</td>
<td>.25</td>
<td>2.21*</td>
<td>2.31*</td>
</tr>
<tr>
<td>RC8</td>
<td>93.82 (11.96)</td>
<td>92.87 (12.32)</td>
<td>60.82 (11.64)</td>
<td>54.99 (13.64)</td>
<td>&lt;.01</td>
<td>.08</td>
<td>.45</td>
<td>2.76*</td>
<td>2.87*</td>
</tr>
<tr>
<td>RC9</td>
<td>69.66 (11.45)</td>
<td>72.26 (11.08)</td>
<td>58.33 (12.73)</td>
<td>58.44 (11.83)</td>
<td>&lt;.01</td>
<td>.23</td>
<td>.01</td>
<td>.92*</td>
<td>1.19*</td>
</tr>
</tbody>
</table>

*Table 8. RC Scales - Over-reporting vs. Standard Instruction Women: Means, Standard Deviations and Effect Sizes*
*p < .006  \( ^a \) Interaction effect sizes are reported as \( \eta^2 \) all others are Cohen’s \( d \)

how the scales were administered. However, the mean scores were significantly different across instruction type, with men who were instructed to over-report scoring statistically significantly higher than men who were given standard instructions for both those who took the test adaptively ($t = 12.99$, $df = 121$, $p<.006$, $d=2.39$) and conventionally ($t = 13.78$, $df = 134$, $p<.006$, $d=2.33$). Similar results were obtained for both men (Table 7) and women (Table 8) on the other RC scales with one exception. For RC6, there was no significant difference between conventional administrations for men or women who were asked to over-report their symptomatology. Also, no difference was found between men who took the test under standard instructions, independently of whether they took the test adaptively or conventionally. However, for women who took the test under standard instruction, those who took the test conventionally (54.96 [12.31]) scored significantly lower on RC6 than those who took the test adaptively (61.04[10.23]; $t = 3.22$, $df = 141$, $p<.006$, $d=.53$).

**Mean Restructured Clinical Scale comparisons across under-reporting groups.**

Effect sizes for the mean score comparisons, as well as means and standard deviations for the RC Scales are reported by administration and instruction type in Table 9 for men and Table 10 for women. Two (administration: conventional vs. adaptive) by two (instruction: under-reporting vs. standard) between-participants ANOVAs were conducted to test interactions between administration and instruction types on the RC Scales. No statistically significant interaction effects were found on any of the scales for either men or women. For each of the RC Scale, $t$ tests were conducted to compare the
Table 9. RC Scales - Under-reporting vs. Standard Instruction Men: Means, Standard Deviations and Effect Sizes

<table>
<thead>
<tr>
<th>Scale</th>
<th>Under-reporting</th>
<th>Standard</th>
<th>Effect Sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td>RCd</td>
<td>49.30 (12.64)</td>
<td>53.99 (11.37)</td>
<td>56.10 (9.32)</td>
</tr>
<tr>
<td>RC1</td>
<td>52.63 (11.72)</td>
<td>55.82 (10.63)</td>
<td>55.14 (10.47)</td>
</tr>
<tr>
<td>RC2</td>
<td>43.74 (9.78)</td>
<td>47.57 (11.22)</td>
<td>47.55 (8.33)</td>
</tr>
<tr>
<td>RC3</td>
<td>49.67 (11.21)</td>
<td>54.34 (10.10)</td>
<td>57.17 (10.42)</td>
</tr>
<tr>
<td>RC4</td>
<td>44.59 (9.83)</td>
<td>51.87 (10.09)</td>
<td>55.32 (9.90)</td>
</tr>
<tr>
<td>RC6</td>
<td>55.41 (12.77)</td>
<td>56.26 (10.95)</td>
<td>58.93 (11.45)</td>
</tr>
<tr>
<td>RC7</td>
<td>48.43 (14.93)</td>
<td>54.20 (11.85)</td>
<td>55.33 (10.04)</td>
</tr>
<tr>
<td>RC8</td>
<td>56.91 (14.29)</td>
<td>62.56 (10.26)</td>
<td>59.72 (11.65)</td>
</tr>
<tr>
<td></td>
<td>RC9</td>
<td>50.43 (12.62)</td>
<td>50.54 (8.97)</td>
</tr>
<tr>
<td>---</td>
<td>------</td>
<td>---------------</td>
<td>---------------</td>
</tr>
</tbody>
</table>

*p< .006  

\(^a\) Interaction effect sizes are reported as \(\eta^2\) all others are Cohen’s \(d\)

Table 10. RC Scales - Under-reporting vs. Standard Instruction Women: Means, Standard Deviations and Effect Sizes

<table>
<thead>
<tr>
<th>Scale</th>
<th>Under-reporting</th>
<th>Standard</th>
<th>Effect Sizes</th>
</tr>
</thead>
<tbody>
<tr>
<td>RCd</td>
<td>41.93 (8.59)</td>
<td>41.28 (7.22)</td>
<td>54.94 (11.32)</td>
</tr>
<tr>
<td>RC1</td>
<td>47.85 (9.37)</td>
<td>45.10 (7.84)</td>
<td>57.86 (13.22)</td>
</tr>
<tr>
<td>RC2</td>
<td>38.55 (6.39)</td>
<td>40.46 (7.59)</td>
<td>48.81 (11.13)</td>
</tr>
<tr>
<td>RC3</td>
<td>46.88 (7.99)</td>
<td>47.97 (9.59)</td>
<td>54.42 (9.34)</td>
</tr>
<tr>
<td>RC4</td>
<td>41.03 (8.45)</td>
<td>41.10 (6.71)</td>
<td>56.11 (12.01)</td>
</tr>
<tr>
<td>RC6</td>
<td>51.85 (9.44)</td>
<td>46.15 (7.57)</td>
<td>61.04 (10.23)</td>
</tr>
</tbody>
</table>

<sup>a</sup> Interactions were not significant.
<p>| | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>RC7</td>
<td>40.50</td>
<td>9.69</td>
<td>38.03</td>
<td>7.62</td>
<td>&lt;.01</td>
<td>.28</td>
<td>.25</td>
<td>1.29*</td>
</tr>
<tr>
<td>RC8</td>
<td>50.78</td>
<td>9.18</td>
<td>46.13</td>
<td>7.10</td>
<td>&lt;.01</td>
<td>.56</td>
<td>.45</td>
<td>.92*</td>
</tr>
<tr>
<td>RC9</td>
<td>44.33</td>
<td>9.04</td>
<td>47.67</td>
<td>9.11</td>
<td>&lt;.01</td>
<td>.36</td>
<td>.01</td>
<td>1.20*</td>
</tr>
</tbody>
</table>

*p < .006  a Interaction effect sizes are reported as η² all others are Cohen’s d

effects of administration (adaptive vs. conventional) on under-reporting and standard instruction groups. In addition, t tests were conducted to compare the effects of instruction (under-reporting vs. standard instruction) on adaptively and conventionally administered tests. All t tests were conducted separately by gender to see if the results were replicated across the sexes.

As can be seen in Table 9, men in both under-reporting groups (adaptive = 49.30 (12.64) and conventional = 47.71 [11.99]) produced lower scores on RCd than the standard administration groups (adaptive = 53.99 [11.37]) and conventional = 56.10 [9.32]) did. Mean scores were not significantly different for administration type (conventional vs. adaptive) for either under-reporting or standard instruction men on RCd, indicating that the scale performed similarly regardless of how the test was administered. The mean scores were different across instruction type.

Similar results were obtained for men on many of the other RC Scales. Small effect sizes were obtained for all of the RC Scales except for RC6, with standard instruction men scoring higher than those who were instructed to under-report their symptomatology. For RC6, almost no difference was obtained for those who took the test under standard instruction versus those who were instructed to under-report. Larger effect sizes were obtained for men in the conventional condition, but generally the same pattern emerged, with larger effect sizes for instruction than administration. Again, the only difference was RC6, which showed a larger effect size for administration than it did for instruction in the adaptive (but not the conventional) condition. Although the effect sizes are larger among the women (Table 10), the results were similar to those found with
the men. Only RC6 showed significant differences between the adaptive and conventional conditions. These differences were apparent for both the under-reporting ($t = 2.95, df = 77, p<.006, d=.66$) and standard ($t = 3.22, df = 141, p<.006, d=.53$) instruction conditions.
CHAPTER 4

Discussion

The analyses conducted in this study had two primary goals. The first was to determine whether adaptively-administered validity scales function as well as conventionally-administered ones in detecting over- and under-reporting. The second was to examine whether administration mode differentially affects scores on the substantive scales of the MMPI-2 (represented by the MMPI-2 RC Scales) when a test-taker is instructed to over- or under-report their symptomatology. Overall, the findings indicate that adaptive administration of the MMPI-2 does not affect the ability of the validity scales to detect over- and under-reporting. In both the over- and under-reporting conditions, instruction type influenced the scores on the RC scales in the expected ways, raising scores in the over-reporting condition and lowering them in the under-reporting condition, regardless of mode of administration.

An interesting finding emerged in examining the effects of administration on the S scale. Under standard instructions, both men and women who took the test conventionally scored significantly higher than those who took the test adaptively. However, this difference is not likely to affect the clinical utility of the scale, as it did not affect classification statistics at or above 70 T, the level at which the S scale is likely to be interpreted. Indeed, classification accuracies for the S scale under the two types of administration were similar across all of the levels at which they were compared,
indicating that although there were difference between the two types of instruction on this scale, those differences did not affect the ability of the scale to detect under-reporting.

One issue that should be addressed is the need to use lower than typically recommended cutoffs on the under-reporting scales L, K, and S in analyses examining classification accuracies on these scales. For this study an undergraduate sample was used, made up of individuals who tend to be rather high-functioning. Therefore, there is not likely to be much psychopathology among the participants in the sample. Consequentially, the participants in this sample were not likely to have to distort their responses as much as others would to present themselves in a favorable light. Since they engaged in less distortion than others might in a similar setting, it became necessary to lower the cutoffs in order to be sensitive to these more subtle changes in presentation. These lower elevations in under-reporting college samples have been reported by other authors as well. For example, Bagby, Rogers, Buis, and Kalemba (1994) reported mean elevations of 64.3T (16.9) on L and 58.1T (8.1) on K on their sample of under-reporting college students. Similar mean elevations on K [56.6T (9.4) for males, 60.7T (6.8) for females] were reported by Graham, Watts and Timbrook (1991) although elevations on L were higher [72.3T (13.6) for males, 76.5T (15.5) for females]. However, even under standard instruction, college samples often score somewhat lower. Graham, Watts and Timbrook (1991) reported that K was lower than average in their standard instruction sample [46.3T (9.2) for men and 45.1T (8.1) for women]. Bagby et. al (1994) reported that their college sample controls scored 47.4T (10.1) on L and 45.0T (9.8) on K. Neither of these samples reported scores on the S Scale. Lim and Butcher (1996) reported
average scores for men and women on S [51.0T (11.9) and 50.9T (9.0), respectively]. However, Bagby et al. (1997) reported raw scores on S for college students [22.9 (6.1)] that fell below a score comparable to a T score of 50 (raw score of 25). These scores suggest that college students may engage in less defensiveness in their responding under standard conditions than the normative sample did. Although they had lower elevations when under-reporting, the differences between the two administrations allow us to classify the groups at a rate similar to other populations.

For men, none of the RC Scales were sensitive to administration type. For women, only one scale, RC6, was sensitive to administration format. In both the standard and under-reporting instruction conditions RC6 was higher in the adaptive version than in the conventional version. Three possible explanations can be noted here. First, the difference between the two administration formats may be exaggerated due to the brevity of the scale. Of all of the RC scales, this is the second shortest and requires the fewest items (only three) to obtain an elevation. Endorsement on one item causes a 17 T score point jump versus endorsing no items. This means that with a relatively small discrepancy between respondents of the two groups, one could see a large change in mean T scores. Second, the mean difference on RC6 between the two types of administration was only 6T score points in both the under-reporting and standard conditions. Convention indicates that a 5 T score point difference is needed for a clinically significant difference. It is unknown at this point whether this slightly larger difference would translate to a change in the predictive ability of the scale. Third, the
The mean T score for both groups did not rise above 65T, the level at which clinical interpretations are made.

This study did not look at the Symptom Validity Scale (FBS) in evaluating the equivalence of adaptively and conventionally administered scales in detecting over-reporting. This is because FBS has been shown to be sensitive mainly to cognitive and somatic over-reporting and less sensitive to emotional and psychological forms of over-reporting (Graham, Benitez, & Chamberlain, 2007). The instructions to the participants were instructed to act as if they had serious emotional and psychological problems, and so FBS would not be a good measure of detecting over-reporting in this sample. Future research could examine FBS under more appropriate over-reporting conditions.

Forbey and Ben-Porath (2007) noted that a number of studies have been conducted comparing the computer adaptive versions of the MMPI–2 with the conventional versions of the instrument. As they indicated, these studies generally indicate the equivalence of the two methods, with the added advantage that the adaptive version of the MMPI-2 takes less time to administer. This study adds to that literature as a first step in demonstrating the ability of the adaptive version to detect threats to protocol validity in the forms of over- and under reporting.

This study was the first to examine the utility of the MMPI-2-CA in detecting over- and under-reporting using an experimental design. Studies such as this are needed to expand the use and utility of computerized adaptive tests such as the MMPI-2-CA. To date, adaptive personality testing is not routinely employed, even though there are clear time savings in using adaptive methods. As research mounts confirming the
comparability of the adaptive version of the MMPI-2-CA, the confidence that clinicians can place in the results obtained from adaptive testing will increase. This study expands the comparability research into experimental domains as well as testing the MMPI-2-CA’s ability to detect threats to validity.

In an era of managed mental health care where there are likely to be fewer outpatient sessions per client (Goldman, McCulloch, & Sturm, 1998), it is important to get the maximum amount of benefit from a psychological assessment with a minimum cost in time. As computerized testing is faster, more flexible and can provide immediate results, a psychologist can achieve many assessment goals in far less time than with traditional methods. As equivalence of the MMPI-2-CA is demonstrated, its potential benefits over more traditional administration may make it a vital tool in assessment settings.

Whereas future research will be needed to further establish the equality of the MMPI-2-CA and the MMPI-2-CC, the current study adds to the evidence supporting the equivalence of the two measures. Although Simms and Clark (2005) suggested that the MMPI-2-CA may lead to a loss of information, this did not seem to be true in the current study, as both validity and RC scales performed as expected in both administration conditions.

There are some limitations to this study that should be noted. Using a simulation college sample may give us information only about how well the MMPI-2-CA can discriminate between college students who are over- and under-reporting and those taking the test under standard instruction. Using other samples in future research may
help to clarify to what degree the MMPI-2-CA can identify over- and under-reporters across settings and samples. Also the level of motivation of college students is not likely to match those of “real world” over- and under-reporters, as the student sample was only offered points toward fulfillment of a college research requirement for participating in this study. Some authors have proposed that participants be provided with incentives for successful feigning (Rogers, 1997). Since persons who over- or under-report in real situations often do so because of the incentives or rewards that they will get for successful (undetected) over- and under-reporting, it is believed that offering incentives will cause participants to act more realistically. Although some studies try to offset these motivational issues by offering monetary incentives (e.g. Rogers, Sewell, & Ustad, 1995), it is unlikely that researchers would be able to provide the same level of motivation as one would in the real world (for example, over-reporting so that one can gain psychiatric help or under-reporting so that one can retain custody of one’s child). However, this type of research should be repeated with differing levels of motivation to begin answering whether the MMPI-2-CA validity scales can detect over-and under-reporters who have a higher level of motivation to conceal or enhance their pathology.

The use of a simulation sample limits the utility of this study in another way. Although the degree to which the MMPI-2-CA can detect persons who are over-reporting versus those who are following the standard instruction is of interest, it is more important and challenging to be able to distinguish between those who over-report psychopathology and others who truly have psychopathology. The use of a psychiatric patient sample in the future as a comparison group would help to increase the utility and generalizability of
these results. Comparison of over-reporters with psychiatric inpatient and outpatient groups will advance our understanding of the ability of the MMPI-2-CA to detect differences between over-reporters and these groups.

There is also a limitation of breadth. The current investigation analyzed only the RC scales to determine the effects of over- and under-reporting on these substantive scales. Following this research with an examination of other MMPI-2 scales such as the Clinical and Content Scales would help to determine whether these computer administered scales respond to over- and under-reporting in the same fashion.

Also issues such as coaching will need to be addressed by future studies, to determine whether the computerized version can detect those persons who have been given information about either particular disorders or how the validity scales of this test work or both. There is a growing awareness that research dealing with threats to validity should take into account the possibility that actual over- and under-reporters may be coached to be more effective (less detectable). This concern stems in part from a study that showed that almost half of a sample of practicing attorneys felt obligated to give their clients information and about how the validity scales of the MMPI-2 work (Wetter & Corrigan, 1995). Different types of coaching have different effects and so should be evaluated separately (Storm & Graham, 2000).

Finally, the MMPI–2–CA software makes it possible to administer subsets of MMPI–2 scales rather than the entire set of measures. This process may have therapeutic utility (Forbey & Ben-Porath, 2007). Whether administration of select scales affects the validity scales ability to detect over- and under-reporting should be investigated. The
outcome of this study, and others designed to replicate and extend it, will increase the utility of the MMPI-2-CA in standard assessment settings.
References


MMPI-2 Post Test Inquiry

Please circle the correct response or answer in the space provided. Remember to respond honestly regardless of the way you answered the questions on the MMPI-2.

1) What instructions did you receive for completing the MMPI-2?
   a. To answer each statement as it applied to me
   b. To ‘fake bad’ or appear as if I have psychological problems
   c. To ‘fake good’ or appear more psychologically healthy than I really am

2) I followed the instructions:
   a. All of the time
   b. Most of the time
   c. Half of the time
   d. Less than half of the time
   e. Not at all

3) I tried to appear as crazy as I could when I took the test:
   a. All of the time
b. Most of the time
c. Half of the time
d. Less than half of the time
e. Not at all

4) I tried to answer in the way they would look best to the researchers:
   a. All of the time
   b. Most of the time
c. Half of the time
d. Less than half of the time
e. Not at all

5) I tried to answer the questions so I would look normal:
   a. All of the time
   b. Most of the time
c. Half of the time
d. Less than half of the time
e. Not at all

6) I tried to answer the questions as truthfully as possible:
   a. All of the time
   b. Most of the time
7) I got tired or bored and answered some of the questions without reading them carefully:
   a. All of the time
   b. Most of the time
   c. Half of the time
   d. Less than half of the time
   e. Not at all

8) About how many questions did you answer without reading carefully?
   a. 0-5
   b. 6-15
   c. 16-25
   d. 26-35
   e. 36-45
   f. 46-55
   g. More than 55

9) Where were the questions you answered without reading?
a. Mostly in the beginning of the test
b. Mostly in the middle of the test
c. Mostly in the end of the test
d. All over the test

10) Do you think the university will see the result of this MMPI-2?
   a. Yes
   b. Probably
   c. Not sure
   d. Probably not
   e. No

11) I know that there are ways a psychologist can determine if you are answering truthfully to the MMPI-2.
   a. True
   b. False