THE UNIVERSITY OF CALGARY

Comparison of "Traditional" and "New" Validity Measures on the MMPI

by

Carole Ann Woychyshyn

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE

DEGREE OF MASTER OF SCIENCE

DEPARTMENT OF EDUCATIONAL PSYCHOLOGY

CALGARY, ALBERTA

NOVEMBER, 1986

ISBN Ø-315-35966-8

THE UNIVERSITY OF CALGARY

FACULTY OF GRADUATE STUDIES

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies for acceptance, a thesis entitled, "Comparison of 'Traditional' and 'New' Validity Measures on the MMPI" submitted by Carole Ann Woychyshyn, in partial fulfillment of the requirements for the degree of Master of Science.

Supervisor, Dr. W. G. McElheran
Holy Cross Hospital

Supervisor, Dr. D. M. Romney
Department of Educational Psychology

Dr. R. C. Conklin
Department of Educational Psychology

Dr. C. R. Bagley
Department of Social Welfare

November 28, 1986

ABSTRACT

The detection of faking on the MMPI remains a problem. The "traditional" validity scales, L (lie), F (frequency) and K (correction), developed to detect faking on the MMPI, contain a number of confounding factors which limit the confidence which can be placed in them. A number of alternatives to these scales have been suggested as possible improvements in the identification of profile distortion. The Positive Malingering (Mp) Scale was developed for the detection of "fake good" response sets, the Obvious minus Subtle (Wo–Ws) Scales for the detection of both "fake good" and "fake bad" response sets, and the Dissimulation (Ds) Scale for the detection of "fake bad" response sets. The Test–Retest (TR) Index and Carelessness (C) Scale were developed to detect random responding.

This study compared the effectiveness of "traditional" validity indicators L and K and "new" validity measures Mp and Wo–Ws for the detection of "fake good" response sets in a sample of psychiatric patients. The F scale and F–K index were compared with the Ds and Wo–Ws scales for the detection of "fake bad" response sets, while the F scale and F–K index were compared with the TR index and the C scale for the

detection of invalid clinical profiles due to random responding in samples of university students.

Results indicated that the Mp and Wo–Ws scales were most effective in detecting "fake good" response sets in a psychiatric population. All validity measures were extremely effective in detecting "general fake bad" response sets in a normal population. The F scale and F–K index found clinical profiles invalid when students were given instructions to respond carelessly (randomly). However, the TR index and C scale were unable to detect random responding.

Given the results of this study, the employment of the Mp and Wo–Ws scales for the detection of "faking good" are recommended for clinical settings.

Acknowledgements

Appreciation is extended to the Holy Cross Hospital Research Committee and the nursing staff on wards M2 and M3 for their cooperation and assistance.

I am particularly grateful to Dr. W. G. McElheran and Dr. D. M. Romney, my supervisors. Dr. McElheran provided guidance, assistance and encouragement throughout the preparation of this thesis. Dr. Romney provided assistance in planning the project, statistical analysis and in the the final stages of this study.

I also wish to express my appreciation to my family for their understanding and support throughout this project, and particularly to Nicole, for her warm encouragement and support (both verbal and practical) during the writing of this thesis.

Finally, I owe much appreciation to Fenna Hersberger for her availability at short notice and excellent typing.

TABLE OF CONTENTS

PAGE

LIST OF TABLES

LIST OF APPENDICES

PAGE

CHAPTER I

INTRODUCTION

Objective personality tests are self–report tests which can be easily distorted according to the desires of the test–takers. Patients may present themselves in a positive light (fake–good), to receive early discharge from a psychiatric hospital. Job applicants may also "fake good" to present themselves in the best possible light. Exaggeration or "fake bad" response tendencies may occur as a plea for help to relieve psychological distress; people found guilty of crimes may "fake bad" so that they are "sentenced" to psychiatric care instead of jail. Random responding may occur due to confusion, carelessness or a lack of cooperation in clinical and correctional settings. The usefulness of these tests as diagnostic instruments appears dependent upon the honesty of the responder or the ability of the interpreter to detect wilful distortion.

Hathaway and McKinley, when constructing their objective personality test, the Minnesota Multiphasic Personality Inventory (MMPI), hoped to overcome this susceptibility to distortion by using an empirical keying procedure and by the development of three validity indicators, the L (lie), F (frequency) and K (correction) scales. An MMPI profile is considered valid when the patient has provided an accurate and internally

consistent self appraisal in response to MMPI items. The L, F and K scales are commonly taken into consideration when interpreting the validity of the clinical scales.

Although there has been some support for the usefulness of these "traditional" validity scales, a number of problems have been identified. The L scale has reportedly been effective in detecting "fake good" response sets of unsophisticated persons with below average intellectual or educational abilities (Good and Brantner, 1974). Conversely, brighter, well educated and psychologically sophisticated persons normally admit to the minor flaws and weaknesses contained in L scale items (Graham, 1977). Responders who utilize the defense mechanisms of repression and denial excessively may validly elevate L scale scores (Graham, 1977), as well as paranoids with grandiose ideation and clergy and others with a strict sense of moral virtue (Good and Brantner, 1974). Therefore, it can be seen that certain populations may "fake good" and avoid detection on the L scale, while other populations may respond legitimately and elevate L scale scores.

The F scale is commonly used for the detection of "fake bad" response sets. It has been found effective for this purpose (Gough, 1947; Meehl and Hathaway, 1946). However, experimental findings have indicated that elevated F scores are also related to other factors such as

age, intelligence, race and diagnostic classification. Elevated F scores have been found to be an indication of significant and severe psychopathology. McKegney (1965) suggests that most juvenile delinquents can honestly and accurately obtain elevated F scale scores, while Rothaizer (1980) found substance abusers produced elevated F scores without indications of psychopathology. Therefore, it appears some minority groups, individuals with deviant lifestyles, and those experiencing severe disorganization or psychosis may validly produce elevated F scores. If the above factors are not applicable, it is likely that the test–taker has produced an invalid profile. However, elevated F scores cannot determine whether the invalidity was due to a "fake bad" response set or random responding as a result of confusion, carelessness or lack of cooperation.

The K scale is used for the detection of "fake good" response sets. Research has shown that this scale is fairly complex. Elevated K scale scores have been found to represent defensiveness in maladjusted populations, but suggest healthy adjustment in normal populations. Socioeconomic class and educational levels also have a bearing in K scale interpretation and must be taken into consideration. Graham (1977) states that college educated individuals normally produce K scale T–scores approximately 10 points higher than those of lower middle class and lower class individuals.

The F and K scales have been used in combination (F–K Index) to detect both "fake bad" and "fake good" response sets. However, this index has only been found effective in detecting "fake bad" response sets (Gough, 1947; Hunt, 1948). Experimental studies have shown that the F–K index is ineffective in differentiating "fake good" and authentic responders, primarily because the F–K index identifies a high percentage of honest profiles as "fake good" ones. The problems inherent in the F and K scales separately also operate in the F–K index.

The above mentioned problems with the "traditional" validity scales provided the impetus for the development of "new" validity indicators to be used as an addition to or instead of the "traditional" validity indices. The Gough Dissimulation (Ds) Scale was developed by differentiating item responses given by neurotic patients and those simulating a psychoneurotic reaction. The Ds scale has been found effective in distinguishing honest and "fake bad" responders in normal populations, psychiatric populations, and prison populations.

Wiener and Harmon's Obvious (Wo) and Subtle (Ws) scales were developed to detect both "fake good" and "fake bad" response sets. The elevation of subtle items suggests a "fake good" response set, while the elevation of obvious items indicates a "fake bad" response set. More recently, the combination of Wo and Ws has been found to be effective in

detecting distorted profiles: low Wo minus Ws scores have been found effective in identifying "fake good" response sets and elevated Wo minus Ws scores indicative of "faking bad" response sets.

The Positive Malingering (Mp) scale was developed as a subtle lie key for the detection of "fake good" profiles. Studies have shown that the Mp scale is very useful in differentiating "fake good" and honest responding in both normal and clinical populations. For example, Wiggins (1959) considers the Mp scale to be one of the best available measures of social desirability response tendencies.

The Test–Retest (TR) index and the Carelessness (C) scale were developed for use as adjuncts to the "traditional" scales for the detection of random responding. The TR index has been found useful in distinguishing random and non–random responders when elevated F scale scores indicate profile invalidity. The TR index has also been found useful in identifying random responding when the "traditional" validity indicators considered the profile valid. The C scale, just recently developed, may be particularly effective for the detection of random responding by more sophisticated individuals.

To the extent that the MMPI correctly categorizes test–takers in the terms of the diagnostic reference groups, its predictive validity and clinical utility will be maximized. To this end, the accuracy and

consistency of item endorsement must be established. This has typically been done via the use of scales designed to detect distorting response patterns, i.e., L, F and K. Research findings suggest inherent flaws in detecting distortion with these scales and, further, that other scales, currently not routinely used for MMPI validity decisions, show promise of either enhancing or supplanting the traditional measures. This study is designed to compare the "traditional" scales with these "new" measures in terms of their effectiveness in distinguishing valid from nonvalid test profiles.

CHAPTER II

REVIEW OF THE LITERATURE

The MMPI is considered the most widely used objective personality inventory in current usage. Well over 5000 references on clinical and research application of the MMPI are cited in Dahlstrom, Welsh and Dahlstrom's (1975) *An MMPI Handbook, Volume II.* Furthermore, a large number of research articles directly concerned with the MMPI have been published since.

The present review of related research was directed by the following objectives:

1. to review briefly the rationale for, development of and uses of the MMPI.
2. to examine in greater detail, the development, interpretation and usefulness of the validity measures built into the MMPI, namely, the L (Lie), F (Frequency), and K (Correction) scales.
3. to examine additional measures of item consistancy and accuracy derived from the MMPI, namely, the Positive malingering (Mp), Gough Dissimulation (Ds), Test–Retest (TR), the Carelessness (C) and the difference between Wiener and Harmon Obvious Subtle (Wo and Ws) scales.

4. To summarize research findings of studies of simulations of "fake good," "fake bad," and carelessness as they apply to MMPI validity.

I. THE MMPI CLINICAL SCALES

Prior to 1940, objective personality inventories were constructed solely on a rational basis to measure a particular construct or trait. Items selected were those which the test developer believed, on the basis of clinical experience, represented the construct being measured. However, clinical experience and research findings (Page, Landis & Katz, 1934; Landis & Katz, 1934) demonstrated that these inventories were largely unsuccessful in identifying or distinguishing between various pathologies and, therefore, patients were often misclassified. Also, distortion or falsification of responses, to present oneself as one wished to be seen, appeared a relatively easy task.

Starke Hathaway and J.C. McKinley (1940) endeavored to overcome some of these inadequacies. Their development of a multifaceted inventory, covering a wide range of diagnostic categories, was aimed at eliminating the need to administer a variety of tests, each with its specific purpose. Hathaway and McKinley utilized an innovative approach, the

empirical keying procedure, in the construction of the clinical scales of this new inventory, the MMPI (Hathaway, 1943).

First, a sample of 1000 personality related statements were gathered from previous personality tests, textbooks and clinical practice. From these statements, 504 items were chosen which were deemed to be reasonably independent of each other. The second step was the selection of appropriate criterion groups. The primary normative group consisted of 724 visitors and relatives of patients in the University of Minnesota hospitals. Hathaway and McKinley (1940) found the group to be a fairly representative cross section of the Minnesota population in terms of age (16–55), sex and marital status, although all subjects in this group were white. The normative group was augmented by three other groups of subjects: precollege students, skilled workers and physically ill patients.

The criterion group consisted of eight subgroups of psychiatric patients at the University of Minnesota Hospitals who had received a clinical diagnosis in one, and only one, of the following diagnostic categories: hypochondriasis, depression, hysteria, psychopathic deviate, paranoia, psychasthenia, schizophrenia, hypomania.

The next step was the selection of appropriate items for each of the eight clinical scales. As an example, in the construction of Scale 1 (Hypochondriasis), the 504 sample items were presented to the primary

normative group and the criterion subgroup diagnosed psychiatrically as hypochondriacs. Items selected for Scale 1 (Hs) were those which were answered differently by the 50 hypochondriacs, as found in an item analysis. In this initial process, 55 items were identified for inclusion in the Hs Scale. Administration to the additional normative groups, and cross–validation by administration to new groups of normal and clinical subjects resulted in refinement and further modifications, reducing the Hs scale to 33 items. The construction of the Depression (D), Hysteria (Hy), Psychopathic Deviate (Pd), Psychasthenia (Pt), Hypomania (Ma) and later the Paranoia (Pa) and Schizophrenia (Sc) scales followed, using the same developmental procedure.

The Masculinity–Femininity (Mf) scale and the Social Introversion (Si) scale were added later, and were constructed in a slightly different way (Dahlstrom, Welsh & Dahlstrom, 1972).

Hathaway and McKinley also constructed two validity scales aimed at detecting deviant test–taking attitudes, namely, the L and F scales. Later Meehl and Hathaway developed a third validity scale, the K scale. These scales will be viewed in greater detail in a forthcoming section.

The MMPI can be administered easily to individuals and/or groups who are 16 years of age or older with at least six years of formal education. There are a number of test forms available consisting of 566

items, which can normally be completed in approximately one to one and one half hours. Scoring can be accomplished by hand or by computer.

Scoring of the MMPI consists of giving one point for each statement answered in the direction opposite to that given by the normative group. This raw score is then plotted onto a profile sheet appropriate for the individual's sex. A correction factor is routinely added to 5 of the clinical scales. This is the K factor to be discussed later. A T–score which has a mean of 50 and a standard deviation of 10 is shown on the sides of the profile sheet, for easy visual comparison of the patients score with that of Hathaway and McKinley's normal standardization group. A T–score of 70, two SD's above the mean, is generally considered the cut–off point for identifying psychopathology. A number of other normative groups have been established to aid interpretation of individual profiles if they fall outside the original normative group in terms of age, education, cultural background or race. For example, norms were developed for a sample of normal Minnesota adolescents (Marks, Seeman & Haller, 1974).

Interpretation of an MMPI profile consists of integrating information regarding T–scores on each of the ten clinical scales and three validity scales. Although the T–score on a single scale provides useful information, it is the configuration of the scales in relation to each other which is most useful in providing the clinician with a picture of the

patient's psychopathology. Hathaway (1947), Welsh (1948) and more recently Gilberstadt (1970) and Marks, Seeman, and Haller (1974), have provided coding systems which are commonly utilized by clinicians. *The Eighth Mental Measurements Yearbook* (Buros, 1978, pp. 938–962) also provides examples of the automated interpretive systems available.

The confidence which is invested in the MMPI is partially related to the criterion–keying procedures used in empirically establishing the validity of scored items. Retest reliabilities on the clinical scales, following one week, in both normal groups and psychiatric patients range from the .50's to the mid .90's. Following one year, retest reliabilities range from the .30's to the low .70's. Another factor contributing to the confidence in the test is its utilization of the three validity scales which represent checks on test–taking attitudes.

II. THE MMPI VALIDITY SCALES

The concept of validity traditionally has meant "the degree to which the test actually measures what it purports to measure" (Anastasi, 1968, p. 28). The development of both the MMPI clinical and so–called validity scales was based psychometrically on criterion validity. Criterion validity refers to an empirical–atheoretical approach which, in the development of the MMPI, utilized the method of contrasted groups to distinguish between normal and pathological groups.

In practice, the validity scales are used to appraise the genuineness and consistency of responding. Criterion levels (cut–off scores) are established, violation of which is believed to indicate that scores on the clinical scales are not valid, i.e., do not measure what they were intended to measure. In a strict psychometric sense, "genuineness" and "consistency" have different meaning, the former signifying validity, the latter, reliability. However, in the context of determining if MMPI clinical scales are interpretable, no distinction is made between them, and the term "validity" is routinely used. Hathaway and McKinley constructed three validity scales, L (Lie), F (referring to frequency of items endorsed relatively rarely) and K (referring to a correction factor).

The L (Lie) Scale

The L scale is a rationally derived scale, constructed by Hathaway and McKinley to identify test–takers who were attempting to present an unusually good front by deliberately avoiding honest responding. Hathaway and McKinley produced 15 items similar to those of Hartshorne and May (1928), who originated the idea in an attempt to study deceit among school–age children. Item 225 "I gossip a little at times" and item 45 "I do not always tell the truth" are examples of items which reflect the behaviours which most persons could answer truthfully in the affirmative direction. All deviant, and socially desirable, responses to L

scale items are "false," making the scale susceptible to a general acquiescence set. Harris and Baxter (1965) found only one L scale item, #15, "Once in a while I think of things too bad to talk about," rated as ambiguous by 60% of their sample. Overall, the L scale was rated by 74 university students as one of the least ambiguous scales, with items being relatively obvious. Although the original Minnesota normative group answered over 80% of the items in a non–deviant manner, three items, #15, 135, and 165, were answered in a scored direction by a majority of the normative sample. Gravitz (1970), in a validation study, found these three, plus two more items, #45 and #255, answered in the deviant direction by the majority of his sample of over 10,000 job applicants. It was also found that thirteen of the items were differentially answered by males and females, suggesting the need for separate scales by sex. The T–scores for the L scale were arbitrarily assigned. Clinical experience has shown that they were set too low, and Hathaway and McKinley (1967) suggested a raw L score of 7 equal to a T–score of 70 would be appropriate. The average raw score for the original normative group was four and the distribution of scores was markedly skewed with a cluster at the low end (Dahlstrom, Welsh & Dahlstrom, 1972).

Dahlstrom, Welsh and Dahlstrom (1975) provide reliability coefficients for intervals up to one week, ranging from .70 to .85.

However, for intervals of one year or more, reliability coefficients range from only .35 to .60. These are slightly lower than those for the other validity scales.

The original purpose of the L scale, as an index of falsification, generally appears effective in detecting deception in naive and socially unaware individuals (Meehl & Hathaway, 1946). However, evidence is somewhat contradictory. McKinley, Hathaway and Meehl (1948) reported L was moderately successful with women in detecting a "fake good" set. Hunt (1948) found no significant elevation in L scores for male prisoners "faking good." Grayson and Olinger (1957) found no significant change in L scores when psychiatric patients were instructed to simulate normalcy, although significant improvement in clinical scales were found in 73% of the sample. On the other hand, Gendreau, Irvine and Knight (1973) found the L scale to differentiate honest responding from "faking good" in 83% of their prisoner sample.

Other studies suggest cultural differences have an effect on L scale scores. McDonald and Gynther (1963) found black students scored higher than whites on the L scale when matched on social class. They concluded that differences were culturally determined. Dahlstrom, Welsh and Dahlstrom (1972) indicate that clergy, social activists and reformers, who rigidly control unethical or antisocial impulses are likely to provide

elevated L scale scores. Coyle and Heap (1965) proposed that in some hospitalized paranoids, where the patients have pathologically grandiose ideas of their own perfection, elevated L scale values will result.

Interpretations of the L scale associate elevated scores with social desirability and defensiveness (Butcher, 1969; Crowne & Marlowe, 1960; Dahlstrom, Welsh & Dahlstrom, 1972; Good & Brantner, 1974). Wiggins (1959) provides evidence of the correlation between the L scale and social desirability. To investigate the defensive interpretation of the L scale, Burish and Houston (1976) related it to the denial scale (Dn), an experimental scale of the MMPI developed by Little and Fisher (1958) which is purported to measure defensiveness, and to the Sc and Hs scales of the MMPI which are considered unrelated to defensiveness. Results provided evidence for convergent validity by showing a significant positive relationship between the L and Dn scales, and evidence for discriminant validity by showing no significant correlation between the L scale and the Sc and Hs scales.

Overall, research and clinical experience suggest that unsophisticated individuals with below average intellectual ability or low educational levels who deliberately distort test responses to place themselves in a more favorable light can be detected by the L scale. Brighter and more highly educated individuals appear willing to admit some of the minor social

faults included on the L scale. Research has shown that minority groups are more likely to have elevated L scores; however, it remains unclear whether this reflects deliberate attempts at distortion or merely reflects cultural differences. Also, persons who strongly utilize the defense mechanism of denial and those with little self–insight are likely to demonstrate high scores without deliberate attempts at distortion. It has been suggested that paranoids, with grandiose ideation, as well as clergy and others who may have a strict sense of moral virtue may validly elevate L scores (Good & Brantner, 1974). Therefore, interpreters must take into consideration the above factors before concluding an elevated L score represents a deliberate effort to avoid answering the test honestly.

The F (Frequency) Scale

The F scale was developed to detect unusual or atypical responding on the MMPI. Hathaway and McKinley selected 64 items which had been endorsed by no more than 10% of the Minnesota normative samples. The content of the items taps beliefs, attitudes, experiences and feelings. Comrey (1958) identified 19 content areas indicating such characteristics as hostility, antisocial attitudes and behavior, poor physical health, paranoid thinking, bizarre sensations, strange experiences and feelings of isolation. Harris and Baxter (1965) reported ratings of all but one of the F–scale items as "unambiguous." Examples of items scored in the "True"

direction are "I believe my sins are unpardonable" and "Evil spirits possess me at times." Items such as "I loved my mother" and "I am liked by most people who know me" are scored in the "False" direction. There is considerable overlap between F items and items on four of the clinical scales. The so called psychotic scales (Pa, Pt and Sc) contain 21 items of the F scale, and the Ma scale contains 9 items, leaving 35 items unique to the F scale. T–scores for the F scale were arbitrarily assigned: Hathaway and McKinley (1951) originally suggested a raw score of 12 should equal a T–score of 70. Subsequent research indicated that this was set too high, and Hathaway and McKinley (1967), upon reevaluation, suggested a raw score of 16 equal a T–score of 70. Meehl and Hathaway (1946) stated that the mean score for the normative groups ran between two and four. As in the L scale, the distribution curve is positively skewed. Only 2 to 3% of the population answer 12 or more items in the scored direction (Meehl & Hathaway, 1946).

The original purpose of the scale was to detect those who falsify their responses owing to their inability to read or comprehend the statements, carelessness or deliberate lack of cooperation. It was also considered a check against clerical errors in scoring. Hathaway and

McKinley (1943) acknowledged that two types of individuals may validly obtain high F scores. These are highly individualistic and independent persons and seriously disturbed psychiatric patients.

Dahlstrom, Welsh and Dahlstrom (1975) report test–retest reliability coefficients for the F scale range from .80 to .97 for an interval up to two weeks, and range from .45 to .76 for intervals from 8 months to 3 years.

Subsequent research has confirmed the F scale's ability to detect invalidity resulting from inadequate reading proficiency (Costa, London & Levita, 1963) and inadequate reading comprehension (Dahlstrom, Welsh & Dahlstrom, 1972). Meehl and Hathaway (1946) state that normal individuals can validly obtain high F scores by responding to items in an unusual manner. They also confirmed that some psychotics, because of their delusional mental states, obtain high F scores. This is consistent with Schneck's (1948) study and Kazan and Sheinberg's (1945) study which indicated that generally a high F score was an indication of significant and severe psychopathology. Gynther and Petzel (1967) found no difference in response patterns to F scale items between psychotics and behavior disorders. They hypothesized that nonconformity may be the general dimension related to high F scale responses by both the psychotics and behavior disorders.

Meehl and Hathaway (1946) also indicated high F scores are obtained by those wishing to present themselves in a "bad" light. For example, they found, with a group of 54 armed forces men who were instructed to obtain adverse scores without detection, in an attempt to avoid the draft, 96% of the sample were able to produce abnormal profiles, but obtained raw F scores of 15 or more. This provided evidence that the F scale was capable of detecting "fake bad" attempts. Gough (1947) obtained similar results. Many investigators and clinicians routinely eliminate MMPI profiles as invalid on the basis of high F scores which purportedly reflect a negative or dishonest test–taking attitude.

More recent experimental findings and clinical opinion consider high F scores related to other factors, such as age, intelligence, diagnostic classification and race. Gynther (1961) found that younger male court referrals more frequently obtained elevated F scores than older ones. When the group was equated for age and intelligence, it was found that 67% of invalid (F>16) scores were produced by persons with behavior disorders, 33% with psychosis, while no neurotic's F scores exceeded 16. Gynther also found a tendency for aggressive criminals to obtain elevated scores (F>16) more frequently than passive criminals, a finding that was compatible with Leary's (1957) view that elevated F scores are related to aggression, hostility and sadism. Overall, Gynther (1961) found a positive

relationship between elevated F scores and younger aged males, behavior disorders and aggressive behavior. He concluded that routinely discarding MMPI profiles because of elevated F scores was highly questionable. Gynther and Shimkunas (1965) showed that age and intelligence affect F scores, i.e. as age and intelligence increases, the F scores are likely to decrease. But it was found that, for low intelligence individuals, as age increased F scores decreased, whereas individuals with average intelligence did not produce lower F scores as age increased. It was suggested that the age/intelligence interaction is the critical factor affecting changes in F scores. An exception to the age and F score relationship was found with persons 60 years and older with IQs below 90. These individuals all produced relatively high F scores. However, the majority were foreign born and educated, and had been diagnosed as psychotic, confounding the possible causes of F scale elevations. Gynther and Shimkunas (1965) also found that intelligence was a better predictor of F scores than educational level. Schenkenberg, Gottfredson and Christensen (1984) found F scores for military veterans applying for mental health care, aged 20 to 40 years, were significantly higher than those who were aged 55 to 60 years. Archer, White and Orvin (1979) found the mean F score for male and female adolescent psychiatric inpatients was markedly elevated (F>16). Clinical elevations were on the Pd and Sc scales. No significant difference

in scores was found between sexes. They also found a large majority of elevated profiles reflected the individuals' extensive degree of psychopathology, rather than illiteracy, carelessness, random responding or malingering. McKegney (1965) suggests that a raw F score over 16 is not indicative of invalidity for most juvenile delinquents: most delinquents can honestly and accurately achieve elevated F scores. In a correctional institution setting, 21 F scale items were answered in a deviant direction significantly more frequently than the norm, demonstrating that certain F items have special meaning for delinquents as a group. Rothaizer (1980) indicated that substance abusers produced high F scores, and most MMPI profiles would have been rejected if traditional validity scale cut–offs had been used. Individuals in this sample did not exhibit any signs of severe psychopathology.

Studies (Harrison & Kass, 1968; McDonald & Gynther, 1963) indicate race or cultural differences may have an effect on F scale scores. As a result of general findings that blacks score higher on the F scale, Gynther, Lachar and Dahlstrom (1978) constructed an alternate F scale for blacks. There is some evidence (Montgomery & Orozco, 1985; Plemons, 1977) suggesting Mexican–Americans are likely to defend against any indication of psychological distress. Sue and Sue (1974) found Chinese and Japanese students elevated F scores and some of the clinical scores.

However, research is too sparse to draw any conclusions as to the validity of MMPI profiles with elevated F scores in these minority groups.

Overall, clinical experience and research evidence indicates that there are several factors to consider when interpreting F scale scores and the validity of the MMPI profile. Persons who are experiencing severe disorganization or psychosis, adolescents with delinquent tendencies, drug users and those with cultural differences may validly produce extreme F scores. If the above factors do not apply to the test–taker, it is likely the elevated F score resulted from exaggeration of symptoms as a "plea for help," resistance to testing or malingering (Dahlstrom, Welsh & Dahlstrom, 1972; Good & Brantner, 1974; Graham, 1977). In these cases, the elevated F score is thought to indicate an invalid profile. Numerous simulation studies (which will be examined presently) provide evidence that "faking bad" can be detected by elevated F scores.

Finally, scores on the F scale are often viewed in conjunction with K scale scores, in determining attempts at both "faking good" and "faking bad." This will be discussed later in the chapter.

The K (Correction) Scale

Early clinical experience with the original validity indicators, scales L and F, indicated that although they appeared effective in identifying gross instances of test–taking distortions, there seemed to be a more

subtle type of distorting attitude which remained untapped by these validity scales. Meehl and Hathaway (1946) stated "it is presumably a significant fact about a person that, in answering a personality inventory, he tends to behave as a 'liar' or a 'plus–getter'" (p. 544). It was felt that this tendency toward defensiveness or plus–getting was fairly common and often unconscious. Meehl and Hathaway developed the K scale as a means of detecting these subtle score–enhancing or score–diminishing attitudes and as a way of providing statistical corrections to the clinical scales to offset defensive test–taking tendencies.

Meehl and Hathaway (1946) utilized an empirical procedure in developing the K scale. The criterion group consisted of 25 male and 25 female patients, who were diagnosed as psychopathic, alcoholic or having other behavior or characterological disorders, but who obtained normal MMPI profiles with an L scale score of T=60 or greater. Responses of the criterion group were compared with those of the original normative group. Item analysis identified 22 items which showed a "percent difference of 30 or more between the criterion cases and the control group" (p. 541). Items on the L scale had been excluded and males and females were considered separately. The content of these items covers such areas as family, interpersonal relationships and self control. Scored responses generally suggest a denial of inferiority, worry, or health

problems while presenting a disposition to see only good in oneself and others. An example of an item scored in the affirmative direction is "I have very few quarrels with my family," while "I worry over money and business" is scored in the negative direction. Following careful study, it was found that severe depressive and schizophrenic patients obtained normal MMPI clinical profiles and L scale T–scores of 60 or greater, leading to an under–interpretation of their abnormality. Meehl and Hathaway then studied responses of male subjects who had been instructed to adopt "fake good" and "fake bad" response sets. Items which showed no tendency to change with altered testing–attitude instructions were noted. From these items, eight items were selected which demonstrated differences between the depressive and schizophrenic criterion groups and the normative group. The combination of these eight items with the previous 22 items completed the 30 items of the K scale. All the eight additional items are scored in the negative direction. An example is "What others think of me does not bother me." One of these items "At times I feel like swearing" is also scored on the L scale. Twenty–five items are also scored on one or more of the clinical scales, leaving only five items unique to the K scale. All but seven items are scored in the same direction on the K and clinical scales. Six items shared with the Si scale are scored in the opposite direction.

Dahlstrom, Welsh and Dahlstrom (1975) report test–retest reliability coefficients for the K scale, range from .78 to .92 for periods up to two weeks. They range from .52 to .67 for intervals from 8 months to 3 years. Rosen (1952) found a significant difference in K scale scores over a four day period, for 40 male psychiatric patients. He interpreted this change as reflecting changes in the defensive structure of these hospitalized patients.

A high (K>65) K scale score was intended to detect defensiveness, and a low score (K≤45) was thought to reflect candidness and/or self–criticism. Meehl and Hathaway recommended subjectively correcting profiles on the basis of K scores. They found with borderline profiles (borderline profiles are those with clinical scale T–scores between 65 and 80) and low K, interpretations of pathology were over–estimated, and conversely, with borderline profiles and high K scores, profiles were under–interpreted. McKinley, Hathaway and Meehl (1948) therefore, developed a statistical procedure, using the K scale, to correct clinical scores as needed. They reasoned that since K scores represented defensiveness, the most reasonable solution to the problem would be to add K or some portion of K to the raw score on each personality variable. The difficulty was in determining the optimal weight or value of the K factor for each scale. This optimal weight would be the one which

best differentiated between the criterion group in question and normals. A trial–and–error method was used in determining the optimal weight for each personality variable. It was determined that five clinical scales benefited by the use of K–corrections. The scales and optimal values assigned to each are as follows: Hs + .5K, Pd + .4K, Pt + 1.0K, Sc + 1.0K and Ma + .2K. These K–corrections provided the best differentiation of psychoneurotic and psychotic groups from the Minnesota normative group. The K–corrections are built into the MMPI and routinely used, and the standard profile sheet provides all needed fractions of K for all raw scores on the K scale to assist in easy computation of K–corrections.

Research regarding the appropriateness of the K–corrections has been fairly sparse. Tyler and Michaelis (1953) and Yonge (1966) found little if any difference between uncorrected and K–corrected profiles in samples of college women and students, and advised against the use of K–corrections. Hunt, Carp, Cass, Winder and Kantor (1948) and Schmidt (1948) also found K–corrections of little use in contributing to effective diagnosis in a military setting. Silver and Sines (1962) obtained similar results with a sample of state hospital patients, while Ruch and Ruch (1967) found K–corrected scales reduced the discriminating power between good and poor salesmen. McKinley, Hathaway and Meehl (1948)

emphasized that the K–weights were intended to differentiate psychoneurotics and psychotics from normals, and cautioned that these particular K–weights may be inappropriate for other populations. It appears that their concerns were not groundless, at least in populations other than psychiatric settings. Heilbrun (1963) developed his own system of K–weighting with a sample of adjusted and maladjusted college students. Results showed significant differences in the systems. Heilbrun applied a negative weight to the Hy scale and deleted weights from the Hs, Pd and Ma scales. He also found that different K weights were needed for males and females on the Hy and Pt scales. Cross–validation was shown with a sample of college educated subjects and with diagnosed psychopaths. No further research has been done using this revised system, leaving open the question of generalizability to other populations. Overall, research findings suggest that standard K–corrections may not be appropriate for normal persons taking the MMPI, but they appear useful when psychopathology is suspected.

The appropriateness of the K scale as a measure of test–taking defensiveness has been investigated by a number of researchers. Results have indicated that the K scale is considerably more complex than was originally thought. Wheeler, Little and Lehner (1951) found abnormal groups scored lower on the K scale than normal groups. Sweetland and

Quay (1953) in summarizing several studies, suggest that besides measuring "defensiveness" and "plus–getting" the K scale may also be a measure of healthy emotional adjustment or personality integration. Berger (1955) and Block and Thomas (1955) found a positive correlation between K scores and degree of self–acceptance in their samples of college students. Smith (1959) found a significant positive relationship between individual defensiveness as measured by the K scale, and insightfulness. He suggested that high K scores signify psychological health in normal populations but defensiveness in abnormal ones. Harris and Baxter (1965) found students who scored high on the K scale, indicating healthy defensiveness or ego strength, see less ambiguity in MMPI items than do low K scorers. Heilbrun (1961) found some support that the K scale is a measure of psychological health in a normal female college population. Correlational data supported the hypothesis that the K scale is a better measure of defensiveness with maladjusted college subjects than with adjusted subjects.

A number of studies found that K scores increased following treatment (Gallagher, 1953; Schofield, 1953) These findings are compatible with the above studies which suggest high K scores measure adjustment. Ries (1966) found that when patient's K scale scores fell within the middle range, i.e., 9 to 15, prognosis was better. Sines, Baucom and Gruba

(1979) advise extreme caution when interpreting defensive (L or K score above 70, or both L and K T–scores above 60) profiles with nonpsychotic–appearing clinical scale configurations, as approximately half of his sample of psychiatric inpatients with these profiles were actually psychotic. In their study, when validity scales were absent, judges were inaccurate in correctly classifying profiles from defensive psychotics and nondefensive nonpsychotics. Accuracy improved when validity scales were present. Overall, there seems to be support for Smith's (1959) contention that high K scale scores suggest defensiveness in a maladjusted population but may suggest healthy adjustment in a normal population.

Dahlstrom, Welsh and Dahlstrom (1972), Graham (1977), Good and Brantner (1974), and Greene (1980) emphasize the importance of socioeconomic class and educational level of individuals when interpreting the K scale score. Graham (1977) states that average college educated persons produce K scale T–scores in the 55 to 70 range, lower middle class and lower class average persons produce in the 40 to 60 range. T–scores for the K scale were derived in a standardized manner. However, as seen from the above findings, no definite K score can indicate the profile is invalid. Greene (1980) proposes a K raw score below 9 (T–score below 45) should be considered low and the individual considered to be "faking bad" or experiencing acute psychotic distress. A K raw score

above 15 (T–score above 65) should be viewed as high and an interpretation of defensiveness or "faking good" should be considered, particularly in lower class individuals (Greene, 1980). Overall, research has shown that the traditional scales L, F and K may effectively detect simulation. However, a number of confounding factors have been found in each of these validity indices which limits the confidence that can be placed on these interpretations.

Validity Scale Configurations

In clinical settings, the most common validity scale configuration consists of an F scale above a T–score of 60, with L and K T–scores below 50. This configuration is thought to indicate an acknowledgement of personal or emotional difficulties. However, as the F T–score increases and K T–score decreases an interpretation of a plea for help or simulation of psychopathology may be warranted. The second configuration, L and K T–scores of 60 or above and F T–scores below 50, are most frequently found in "defensive" normals, in unsophisticated job applicants, or in psychiatric patients diagnosed as hysterics and hypochondriacs (Greene, 1980). The third configuration, showing the L scale less than F, and F less than K, is found in college–educated individuals who demonstrate sophisticated defensiveness, and in normal individuals with marital conflict. It is also found in normal individuals who have appropriate resources to

deal with stress and conflict. However, prison inmates or job applicants who are "faking good" may show this configuration. The last configuration, with K less than F, and F less than L, is less common than the previously mentioned configurations. It is normally found in lower socioeconomic class members who are attempting to look good but are experiencing neurotic symptoms (Greene, 1980).

The validity scales have also been used in other combinations. Cofer, Chance and Judson (1949) suggested that an additive combination of L and K would be useful in the detection of positive malingering. However, Exner, McDowell, Pabst, Stackman and Kirk (1963), Grow, McVaugh and Eno (1980), and Lanyon and Lutz (1984) have found L + K index less effective in detecting positive malingering than other strategies. One combination which has received considerable attention and usage is the F–K index, a technique used for the detection of both "fake good" and "fake bad" attempts.

F–K Index

Gough (1947, 1950) in his studies examining the problem of simulation on the MMPI, developed a new means of detecting invalidity due to falsification. In Gough's (1947) study, eleven clinical workers attempted to feign an acute, severe anxiety neurosis and a nondeteriorated, acute, paranoid schizophrenic psychosis. Judges were asked to identify

these simulated profiles which had been intermixed with 57 authentic psychoneurotic and 13 psychotic profiles. The judges were able to identify 8 of the 11 simulated psychoneurotic profiles and all of the psychotic simulations. The F and K scales, used singly were fairly successful in identifying the feigned profiles. However, the F raw score minus the K raw score combination, with a cutting score of plus four for neurotic profiles, and a cutting score of plus 16 for psychotic profiles was most effective. Ten out of 11 simulated records in each situation were detected with the use of the F–K index. Hunt (1948) extended the use of the F–K index in an attempt to detect "faking good." He arbitrarily selected cutting scores of F–K = +11 or above, to suggest "faking bad," and F–K = –11 or below to suggest "faking good." Results of his study showed that 88% of "fake bad" profiles produced by male students were identified using these indices, while only 2% of honest profiles (F–K = +10 to –10) were misclassified as "fake bad." In the two prisoner samples, 85% and 88% of "fake bad" profiles were correctly identified while 3% and 20% of honest profiles were misclassified as "fake bad." Results on the use of the F–K index with the students to detect "fake good," on the other hand, were discouraging. Although all of the profiles which successfully "faked good" were identified by the F–K index, 93% of honest profiles were also identified by the F–K index as "fake good" profiles. Results with the

prison samples were somewhat more encouraging. Of the fake good profiles, 62% and 55% were correctly classified, while 18% and 12% of the honest profiles were misclassified. Hunt (1948) noted that research indicates that K is likely to be elevated by higher socio–economic status groups. When F is not elevated for this group, the F–K index is likely to indicate "fake good" responding.

Gough (1950) found that all clinical and normal groups had F–K means of less than zero, while all dissembling groups produced F–K means above zero. All authentic profiles, normal and clinical (1,773 cases) were consolidated and compared with a sample of dissemblers (319 cases) to determine optimal cutting scores. It was found that a cutting score of plus 9 correctly classified 97% of authentic profiles and 75% of simulated records. Therefore, if the F–K index exceeded plus nine it was deemed that the profile should be considered invalid due to "faking bad." F–K index scores from 0 to 9 indicated a valid profile while scores below 0 suggested "faking good," although detection of positive dissimulation with the F–K index was less efficient. Gough (1950) also stated that a single cutoff score cannot be established for all settings and suggested that there may be appropriate cutting scores for special situations.

Since Gough's pioneering work, research results regarding the effectiveness of the F–K index for detection of simulation have been less

conclusive. Exner, McDowell, Pabst, Stackman and Kirk (1963) found that with a cutoff of +10, 17 of 25 "fake bad" profiles could be detected, while a cutoff of +12 detected 24 or 25 malingered profiles. The F–K index was thus successful in detecting "fake bad" attempts in this college population. However, when students responded either honestly or "faked good," scores on the F–K index suggested only a "fake good" profile. Anthony (1971) in his study of male military personnel with nonpsychotic disturbances who were asked to exaggerate their problems on the MMPI, found the F–K index successfully differentiated 62% of the exaggerated profiles from valid profiles of other patients with similar profile configurations. The F–K cutoff score used was +10. Vincent, Linsz and Greene (1966) found the F–K index, using Gough's 1950 cutoff score of +9, was only able to detect five cases of deception when 100 university students were instructed to "fake good." It was concluded that the F–K index could only identify simple, direct and naive sorts of faking good deception.

Gendreau, Irvine and Knight (1973) instructed 24 prisoners to fake good adjustment and maladjustment. It was found the F–K index successfully classified 100% of both maladjusted and honest instructional set profiles, but a high cutoff score of F–K = 24 was used. It was reasoned that since previous research found valid profiles with F scores 20

to 30 points above T = 70 in cases of acting–out aggressive behaviour disorders, and since this particular inmate sample fit this description, cutoff scores should be adjusted accordingly. It was found that the prisoners could simulate normalcy quite easily. The mean score on the F–K index was –10.5 with a SD of 8.2, showing a hit rate in the low 80 percentage range. The exact cutoff score used for the "fake good" condition was not mentioned. Results were considerably more favorable than most comparable research findings. It was felt that simulation instructions put into concrete terms which specifically applied to the samples' present circumstances, contributed to the positive results. Wilcox and Dawson (1977) found that ten female university students given instructions to role–play a paranoid role were detected by the F–K index with all scores exceeding +7. However, a group of ten students given the paranoid syndrome suggestion under hypnosis, avoided detection with the F–K index, showing that while role–playing simulation was easily detected, in this sample, hypnosis allowed the avoidance of detection of simulation.

Post and Gasparikova–Krasnec (1979) define a "plea for help" profile as one in which raw F scores are at least 11 points greater than raw K scores. Twenty psychiatric inpatients with "plea for help" validity profiles were compared with 20 patients who obtained average validity

profiles (F moderately elevated and greater than K) and with 20 patients with hyper–defensive profiles (raw K scores were elevated at least 13 points higher than raw F scores). Results showed that the ward staff rated the "plea for help" patients as exhibiting greater feelings of frustration and acting–out more frequently. Patients hospital records indicated significantly higher frequencies of sexual "acting–out," aggression and self–destructive actions than patients in the other groups. Post and Gasparikova–Krasnec suggest a large F–K discrepancy may be clinically useful as predictive of behavioural disorganization and in alerting staff to potential management problems due to a patient's poor impulse control.

Grow, McVaugh and Eno (1980) evaluated the effectiveness of a number of different strategies and techniques for detecting faking on the MMPI. It was generally found that students who were "faking bad" were more easily identified than those who were "faking good." The use of F–K $\geq$ 7 accounted for 79% of the variance associated with students "faking bad." F–K $\geq$ 7 correctly identified 98% of those "faking bad" while F–K $\geq$ 9 correctly identified 92%. Identification of "faking good" was best accomplished by F–K $\leq$ –11, which accounted for 36% of the variance associated with students "faking good." Results were cross–validated with a clinical population which had been judged as responding to the MMPI in either a "fake good," "fake bad" or straightforward manner. Results

indicated F−K ≥ 7 was the best strategy to identify "faking bad," while F−K ≤ −11 was best for identifying "faking good."

Gallucci (1984) evaluated the effectiveness of the F−K index to detect dissimulation in psychiatric patients who were applying for disability benefits and were assumed to be intrinsically motivated to exaggerate psychopathology. Claimants were divided into four groups, based on their inferred motivation to exaggerate. Results showed mean F−K scores of −.796, 2.06, 4.72 and 11.97, as predicted with scores increasing as motivation to exaggerate increased.

Osborne, Colligan and Offord (1986) report normative tables for the F−K index based on a contemporary sample of 335 normal women and 304 normal men who matched the sex and age distribution of the U.S. population in the 1980 census. Results showed males had a mean F−K index of −9.77, women a mean score of −11.22, while the combined score was −10.53. Gough (1950) had reported an overall mean score of −8.96 which was significantly different. Osborne et al. (1986) suggest that the changes which have occurred in MMPI response patterns over 40 years warrants further study regarding the appropriateness of cutting scores used to identify dissimulation.

Overall, results of the above studies suggest the F−K index is more effective in detecting exaggeration of problems and deliberate "fake bad"

attempts than it is at detection of "fake good" attempts. Results also suggest that a single cutoff score is not necessarily appropriate for all populations. High F–K index scores appear associated with individuals who exhibit low impulse control and behave generally in an aggressive manner. Results regarding the effectiveness of the F–K index in situations where the subjects are intrinsically motivated to dissimulate as compared to research studies where subjects are deliberately instructed to dissimulate remains inconclusive.

III. ADDITIONAL AND ALTERNATIVE ("NEW") VALIDITY MEASURES

In addition to the traditional validity measures, L, F, K and their combinations, a number of other measures of test–taking attitudes have been developed. Greene (1980) suggests the clinician should first measure the consistency with which an individual has responded to items on the MMPI. The Test–Retest (TR) Index (Buechley & Ball, 1952) and the Carelessness Scale (Greene, 1978) are two measures of consistency which can be useful in identifying invalid profiles. Greene (1980) suggests that once consistency of response has been verified, the accuracy of item endorsement should be examined. Cofer, Chance and Judson's (1949) Positive Malingering (Mp) Scale, Gough's (1954) Dissimulation (Ds) Scale, and Wiener's (1948) Subtle and Obvious Items (Ws and Wo) are three

measures of the accuracy of item endorsement developed to determine the validity of profiles. Greene (1980) contends the addition of the above validity measures should increase confidence in interpreting MMPI clinical profiles.

The Test–Retest (TR) Index

Buechley and Ball (1952) developed a new scale which consists of 16 repeated items on the MMPI. These items are taken from the Pa, Pt, Sc, and Si scales. An example, #16 and #315, is "I am sure I get a raw deal from life." It was felt the TR index would be useful as a supplement to the F scale, by helping to separate profiles with high F scores, due to psychopathology, from high F scores, due to random responding. The two responses to identical items are compared. The score represents the number of contradictory responses. Fourteen of the paired items are separated by at least 260 items, while the remaining two paired items are separated by 44 and 60 items. There appears little chance that the average responder will be aware of the duplicate items.

Buechley and Ball (1952) examined the scores on the F scale and TR index of 137 male delinquent adolescents. A correlation of +.63 between F and TR scale scores was found. This provides empirical evidence that the TR index serves an independent function. A cutting score of 3 or more was used in this study to designate profile invalidity.

It was found that approximately 15% of cases considered invalid on the F scale ($F \geq 12$), were found valid on the TR index, separating responders who were consistent from those who were unable or unwilling to respond consistently.

Coche' and Steer (1974) compared the mean TR scores of female nursing–school applicants with those of neurotic and psychotic women inpatients who had taken the MMPI under standard instructions. Results showed that the patient sample averaged response inconsistency on 2.5 items. No significant difference was found between neurotics and psychotics. However, a significant difference was found between inpatients and nursing school applicants who averaged less than one inconsistent response on the TR index. It was suggested in this study that a TR cutting score of 4 inconsistent responses would be appropriate.

Jones, Neuringer and Patterson (1976) found that brain–damaged male patients produced more inconsistent responses than nonbrain–damaged psychiatric patients. Normal individuals obtained an average of 1.54 on the TR index, and all psychiatric groups, brain–damaged or not, with the exception of brain–damaged schizophrenics, received mean scores under the cutoff criterion suggested by Buechley and Ball (1952) for invalidating the test results. Therefore, Jones et al. felt that the criterion for invalidation was questionable in differentiating valid and invalid

profiles. Their normative data suggest that for this clinical group response inconsistency is not a typical means of dissimulation on the MMPI.

Greene (1979) provided normative data on the TR index for four populations: 50 all male psychiatric patients, 50 clients at a university clinic, one third of whom were male, 50 predominantly male adolescents at a Juvenile Probation Office and 50 university students who were approximately equally divided by sex. The frequency of inconsistent responses varied considerably both within and between samples. The TR index mean score was 1.86 (SD 1.97) for psychiatric patients, 1.90 (SD = 1.78) for clinic clients, 1.86 (SD = 1.78) for university students and 4.14 (SD = 2.84) for adolescent probationers. It was felt the significantly higher scores for adolescents reflected their general poor motivation and uncooperativeness. Using the cutting score of 4, it was found that 52% of adolescent profiles were classified as invalid, while the cutting score of 3 classified 68% of profiles invalid. With the exception of the adolescent probationers, the TR index appeared relatively independent of the degree or type of psychopathology. This is congruent with the findings of Jones et al. (1976) and Coche' and Steer (1974). Correlations between the F scale raw score and the TR index were positive and statistically significant (range .33 to .59). It was found 4% to 14% of invalid profiles (F >80) were valid according to the TR index (TR >4). In addition, 8% to 16%

of valid profiles (F <71) were classified as invalid on the TR index. This finding was particularly useful with the adolescent sample where a number of individuals obtained raw F scores less than 5 but scored 6 or more on the TR index. Greene (1983) suggested that the optimal cutting score should be at least 6.

Rogers, Dolmetsch and Cavanaugh (1983) compared a sample of 40 computer–generated random profiles with 40 profiles of forensic outpatients who were intrinsically motivated to respond in a specific manner. Groups were compared on validity and clinical scales and on scales designed to detect randomness. It was found that the most effective means of classifying random and non–random responders was the combined clinical decision rules of F >80 and TR index > 4. Using this rule, 95% of non–random responders and 97.5% of random responders were correctly classified.

Evans and Dinning (1983) contend that it is important to make the distinction between very high F scores from those whose responses are content–dependent (malingering and "plea for help") and those whose responses are content–independent (random responding). The MMPI profiles of 51 male psychiatric inpatients, who obtained F T–scores >90 and were classified as either consistent or inconsistent responders by the TR index (TR >4 = inconsistency) were compared. It was found that

the consistent group obtained lower L and K scores, higher Ds–r (see p. 50) and Subtle–Obvious scores, as well as higher degrees of psychopathology on most of the clinical scales. It was felt that this indicated consistent high F scores represented a plea for help. The inconsistent group obtained higher L and K scores. The Ds–r and Obvious–Subtle scores were equivalent to those expected by chance and a lower degree of psychopathology was indicated by the clinical scales. It was felt this result suggested random responding which could be due to reading problems, confusion or a blatant refusal to cooperate.

Overall, the above research indicated that the TR index is a useful measure of response consistency and an adjunct to the standard validity measures. The utility of the TR index appears to include both the detection of random responding (and invalidity), particularly when traditional validity indicators consider the profile valid, and the distinguishing of "plea for help" responders from random responders in high F scale scorers.

The Carelessness Scale

The Carelessness scale was developed by Greene (1978) as an additional validity indicator and adjunct to the TR index. Haertzen and Hill (1963) had developed a carelessness scale for the Addiction Research Center Inventory which included both repeated items (as in the TR index)

and items which were psychological opposites. They found psychologically opposite items to be more sensitive in the detection of inability or unwillingness to respond validly. Dahlstrom, Welsh and Dahlstrom (1972) suggested that psychologically opposite items would be a valuable adjunct to the TR index.

Greene (1978) had 50 patients at a Veterans Administration (VA) hospital, 50 University Psychology Clinic clients and 50 college students complete the MMPI. Items were selected empirically on the basis of being answered in a consistent direction more than 90% of the time. The next step was the selection of pairs of items which were psychological opposites. Greene and two clinical psychology graduate students independently picked pairs of items, from the empirically selected items. Only pairs which all three judges agreed were psychological opposites were retained. This procedure produced 12 pairs of items which comprise the Carelessness Scale. Deviant responses are those which are answered in a psychologically opposite direction. In seven cases, answering item pairs in the same direction (both true or both false) is considered a deviant response. For example, if the test–taker answered both #10 "There seems to be a lump in my throat much of the time" in the true (or false) direction and #405 "I have no trouble swallowing" in the true (or false) direction, one point would be scored for a deviant response. Five item

pairs answered in different directions are scored as deviant responses. An example, is #17 "My father was a good man" and #65 "I loved my father." The mean score on the Carelessness scale was 1.70 for the VA group, 2.20 for the psychology group and 1.48 for the university students. From the mean and standard deviation of total scores, it was determined that a cutting score of 4 or more deviant responses would be optimal in identifying invalid profiles. Correlations of the Carelessness scale with the TR index was .57 for the VA group, .23 for the psychology clinic and .28 for university students. Correlations of the Carelessness scale with the F scale was .52 for VA, .53 for psychology clinic and .43 for university students, showing that the Carelessness scale is measuring related but not identical test–taking behaviour as the F scale and TR index. It was found that the TR index and Carelessness scale were in agreement in 82% of the cases on whether a profile was valid or invalid. An additional 10% of profiles were identified as invalid which had been considered valid by the TR index. Greene suggests that the Carelessness scale, due to the subtle nature of the items, may be able to detect deviant test–taking behavior in sophisticated persons who recognized the fact that a number of items were repeated, and for whom, therefore, the TR index would be invalid.

Rogers, Dolmetsch and Cavanaugh (1983) found a significant difference between random and patient responses on the Carelessness scale. Discriminant analysis demonstrated differentiating patterns of responses between the random and patient samples. The Carelessness scale (Carelessness > 4) correctly classified 80.1% of the random responders and 95.0% of the non–random responders. Using the clinical decision rules F >80 and Carelessness > 4, 80.0% of random responders and 97.5% of non–random responders were identified. Although this combined decision rule was not the most effective at correctly classifying random and non–random responders, Rogers et al. suggested other clinical rules be applied when examining particular profiles. For example, when the Carelessness scale is the only one elevated, it should definitely be taken into account as an indicator of random test–taking behaviour.

Overall, the above research indicates that the Carelessness scale can distinguish between random and non–random responders, and may be particularly useful for detecting random responses made by more sophisticated individuals. Further research is needed to validate the Carelessness scale externally.

The Gough Dissimulation (Ds) Scale

Gough (1954) conducted an extensive survey of the specific stereotypes about neuroticism held by both professionals and laymen. Fifty male and 50 female university students and 11 professional staff members of an army hospital (psychiatrists, psychologists, and social workers) were given the MMPI under instructions to answer items in the way an individual would who was "experiencing a psychoneurotic reaction." Responses were contrasted with profiles of male and female psychiatric patients diagnosed as psychoneurotic. Item analysis identified 74 items that differentiated role–players from patients. For example, from 64% to 90% of role–players gave a "true" response to item #16 "I am sure I get a raw deal from life" while only 7% to 28% of neurotics gave a "true" response. Conversely, only 10 to 36% of role–players gave a "true" response to item #257 "I usually expect to succeed in things I do" while 67 to 88% of neurotics gave a "true" response to this item. For each of the 74 items, neurotic patients responded in the opposite direction from what was expected, suggesting that these items pertained more to the stereotypic view of neuroticism rather than to neuroticism per se. Gough felt differences produced by subjects feigning neuroticism, and neurotic patients, on the 74 items were sufficiently important to be considered as a scale for dissimulation, the Ds scale.

New samples were obtained for the purpose of cross–validation. These consisted of three groups; those simulating neuroticism, neurotic patients and high school students given standard instructions on the MMPI (control). Results showed a mean raw score on the Ds scale of 15.94 (SD = 9.99) for neurotics (N=915). The high school students (N=507) obtained a mean score of 15.88 (SD=7.90). The group simulating neuroticism (N=354) obtained a mean score of 54.13 (SD=11.69). As is apparent, the simulators scored significantly higher than either the clinical or control groups. Of great importance is the fact that mean Ds scores of the control group and clinical group are indistinguishable. A raw cutting score of 35 was established as 93% of simulators scored at or above that point, and only 6% of clinical patients and 2% of the control group were mislabeled.

The Ds scale was later revised to 40 items (Gough, 1957), and it is the revised scale that is now employed in research studies. Thirty–four of these items are scored in the "true" direction making the Ds–r scale fairly susceptible to unsophisticated deviant response sets to answer all items "true." The mean raw score for males on the Ds–r was 6.22 (SD=4.33) and for females it was 7.11 (SD=4.72).

Exner, McDowell, Pabst, Stackman and Kirk (1963) investigated methods to detect malingering with a sample of 25 male and female

college students. The MMPI was administered twice, under standard instructions and instructions to be "sufficiently deviant to be exempt from some social responsibility such as military service but not so deviant that institutionalization would be required." Results showed a significant difference ($p<.01$) between Ds scores under honest and "fake bad" conditions. The mean raw Ds score for the "fake bad" group was 40.27 (SD=11.76). Results support the usefulness of the Ds scale as a method of detecting malingering.

Anthony (1971) examined a variety of detection strategies for malingering. Forty U.S. Air Force nonpsychotic male patients were given the MMPI under standard instructions and instructions to "fake bad." These exaggerated profiles were compared with valid profiles of other patients with configurations similar to the exaggerated ones and the exaggerated profiles were matched with the similar valid ones. Results showed a mean raw Ds score for honest responders of 9.82 (SD=5.98). The "faking bad" mean raw score was 32.25 (SD=16.12) while the valid matching mean score was 26.56 (SD=13.25). The exaggerated matching group obtained a mean score of 32.43 (SD=18.19). The Ds scale, with a cutoff score of 30, correctly identified 20 of the "fake bad" profiles and 21 of the matching profiles as valid. The Ds scale produced a 64% hit rate in correctly identifying the 32 valid matching and exaggerated matching

profiles. The Ds scale significantly distinguished ($p<.05$) the "fake bad" from the matching profiles.

Gendreau, Irvine and Knight (1973) reported the Ds scale significantly differentiated honest and fake maladjustment sets in a sample of prisoners, however no cutoff score was provided. Mean raw Ds scores were 8.6 (SD=4.9) for the honest group, and 23.5 (SD=6.5) for the "fake bad" group. The successful classification percentage for the honest and maladjustment instructional set using the Ds scale was 96%.

Grow, McVaugh and Eno (1980) evaluated the effectiveness of the Ds scale to detect "fake bad" responses on the MMPI in a sample of university students. Results showed that the Ds scale (cutoff $\geq$ 35) correctly identified 86% of "faking bad" profiles, and only 2% of honest profiles were incorrectly identified. This shows a total of 95% of profiles were correctly identified by the Ds scale in a population of university students. The mean raw score for those "faking bad" was 51.22, while honest responders obtained a mean Ds score of 15.5. A cross–validation experiment with profiles of a clinical population showed the Ds scale (cutoff $\geq$35) correctly identified 56% of "faking bad" responders, while 15% of honest responders were incorrectly identified for a total correctly identified of 87%. The mean raw Ds score for those "faking bad" was 35.13 while the mean score for honest responders was 23.69.

Overall, the Ds scale appears very effective in distinguishing between honest and fake bad responders in a normal population of university students. Although the Ds scale (cutoff ≥35) is also effective in a clinical population, it appears less so. The "faking bad" raw mean Ds score was considerably higher for the normal group compared with the clinical group (51.22 vs 35.13) and conversely, the honest mean response for the clinical group (23.69) was considerably higher than for the normal group (15.5). In speculation of these results, the normal group was instructed to "fake bad," whereas the clinical group were 'judged' to have "faked bad." Instructions may have led to a more blatant attempt at distortion. The fact that the clinical group produced higher mean scores than the normal group when responding legitimately may have reflected a variety of psychopathologies other than neuroticism as the clinical group was not restricted to neurotics. The Ds scale correctly identified 95% of profiles in a student population but only 87% in a clinical population. It appears that the Ds scale is more effective in detecting "fake bad" response sets in a normal population. Refinement of cutting scores appear indicated for clinical populations.

The Positive Malingering (Mp) Scale

Cofer, Chance and Judson (1949) in the process of investigating both positive and negative malingering, developed an additional validity scale to detect positive malingering. Three groups of male and female college students completed the MMPI twice. The first group was instructed to answer honestly during one administration and to make the best possible impression ("fake good") during another administration. The second group was first instructed to answer honestly, and then as they thought an emotionally disturbed person would respond ("fake bad"), while the third group was a control group. The order of instruction was divided between the first two groups as a check on order effects. Results of an item analysis of the data showed 462 items changed significantly under instructions to "fake bad," while 108 items changed under instructions to "fake good." The scales which showed the greatest percentage of change under "fake good" instructions were the L, K, Mf and Ma scales, in that order. Further analysis found 39 items which were not affected by "fake bad" instructions but changed under "fake good" instructions. Each of these items had been responded to in a deviant direction by at least one half of the subjects when answering in an honest manner. However, the control group had significantly changed 5 of the 39 items under test–retest conditions. Cofer et al. proposed that the remaining 34 items could be

useful as a subtle lie key for the detection of fake good profiles. An example of an Mp scale item which is scored with a "true" response is "I am not afraid of fire." Item #102, "My hardest battles are with myself" is scored with a "false" answer. An equal number of true and false responses are scored making the Mp scale insusceptible to "all true" or "all false" test–taking sets.

Profiles of the three groups were rescored using the 34 items of the Mp scale. Scale scores of 20 and more were considered "fake good," while scores of 19 or less were considered honest. Using this decision rule, 96% of honest records were correctly classified and 86% of "fake good" profiles were correctly identified. Using a cutoff score of 20 did not misclassify any "fake bad" profiles. Twenty–three of the Mp scale items were not included in Wiener and Harmon's Obvious (Wo) and Subtle (Ws) scales (see p. 58). Of the 11 items which were included, 9 were classified as subtle and 2 as obvious. The 2 obvious items were scored in the reverse direction from the scoring on the Wo scale.

Wiggins (1959) conducted a study in which 190 male and female university students completed the MMPI in the standard way and 250 other students were instructed to answer the items in the way people in general would consider to be more desirable. A highly significant difference ($p<.001$) was found between both men's and women's Mp scores

under standard and role–playing instructions. The mean raw score for males in the experimental group was 20.25 (SD=7.29) and the mean raw score for females was 21.43 (SD=6.21). The Mp scale correctly classified 65% of the social desirability group and 96% of the honest group. Wiggins explained the shrinkage in percentage of correct identification in terms of cross–validation. No previous cross–validation studies had been reported although the development of the Social Desirability (Sd) scale (Wiggins & Rumrill, 1959) was considered a partial cross–validation as it contains 14 of the items included in the Mp scale. Wiggins considers the Mp and Sd scales as the best available measures of social desirability response tendencies in the MMPI (Wiggins, 1959).

Wales and Seeman's (1968) study consisted of 38 university students who completed the MMPI under honest conditions and under instructions to create a very good impression ("fake good"), and 25 students who formed the control group. Results showed that the Mp scale (cutoff score = 20) was successful in classifying 68% of the "fake good" profiles and 94% of the honest profiles. These results are almost identical to those of Wiggins (1959), and identification of honest profiles is also similar to findings of Cofer et al.'s original study. The Mp scale appears particularly effective and useful in correctly classifying honest responders.

The above mentioned investigations of the Mp scale have all utilized university students attempting to create an extremely good impression. Gendreau, Irvine and Knight (1973) examined how effectively prisoners could feign good adjustment. Results showed that the mean Mp score under an honest instructional set was 8.7 (SD=2.6), while under "fake good" adjustment instructions the mean Mp score was 18.5 (SD=4.5). The Mp scale significantly differentiated "faking good" from honest response sets. The reported percentage hit rate for the Mp scale was 92%. The cutoff raw score used was 16. This is 4 points below the originally recommended cutoff score of 20. Gendreau et al. suggested that the applied clinical use of the Mp scale is worthwhile and they recommended the routine use of this scale by correctional psychologists.

Grow, McVaugh and Eno (1980) evaluated the effectiveness of a number of techniques to detect faking. Responses on the MMPI of university students "faking good," and answering honestly were examined. The Mp scale, cutoff score $\geq$ 20, correctly identified 46% of students "faking good," and 4% of honest responders. The mean score in the faking good group was 18.98 and in the honest group it was 11.38. A second experiment examined profiles of a clinical population which had been judged as "faking good," or responding in a

straightforward manner. Results showed the Mp scale, cutoff score ≥ 20, correctly identified 38% of those "faking good," and did **not** misidentify any honest profiles. The mean score of those "faking good" was 16.17 while the mean score of honest responders was 8.54.

Generally, research on the Mp scale suggests that it can be very useful in distinguishing between "faking good" response sets and honest responding. Positive results have been shown using a cutoff score of 20 with a normal population of university students. Research on the effectiveness of the Mp scale with a clinical population is extremely sparse. Gendreau et al. (1973) with a cutting score of 16 found the Mp scale effective; however, Grow et al. (1980) with a cutting score of 20 found the Mp scale less effective than other strategies in identifying "fake good" profiles. The mean score in Grow et al.'s (1980) clinical population of "faking good" responders was 16.17 suggesting that the cutoff score of the Mp scale may need refinement for a clinical population. Further research is needed in this area.

Wiener and Harmon's Subtle (Ws) and Obvious (Wo) Scales

Wiener and Harmon (Wiener, 1948), in keeping with Meehl and Hathaway's (1946) earlier opinion that subtle items minimized test distortion, developed a method of measuring MMPI item subtlety and

obviousness. They defined an "obvious" item as one whose psychological meaning is very clear. For example, most persons could easily identify item #41, "I have had periods of days, weeks, or months when I couldn't take care of things because I couldn't 'get going'" as a measure of depression. A "subtle" item was seen as one which was difficult to detect as being relevant to psychopathology. For example, item #12, "I enjoy reading detective or mystery stories" is unlikely to be viewed as relevant to the Hy scale even by a trained psychologist. Therefore, the subtle–obvious dimension can be viewed as the degree of perceived relevance between content of item and the psychological concept being measured.

Wiener and Harmon (Wiener, 1948) divided all MMPI items into two groups, those in which deviant responses were easy to detect as indicating emotional problems (obvious) and those which were not easy to detect (subtle). Items for each scale were divided into these two groups. F scale items by definition seldom occur in the normal population so they automatically were placed in the obvious group. A rational inspection of all items identified 110 subtle and 140 obvious items. A representative sampling of 100 profiles of the original normative group were rescored using the obvious and subtle keys. It was found that the empirically determined deviant response was in the opposite direction than expected for 59% of the subtle items while only 5% of the obvious item were in the

opposite direction. Wo raw scores indicated a positive skew in the distribution of the normative group while Ws raw scores formed a relatively normal distribution. T–scores were assigned to the Wo and Ws scales in the same manner as with the clinical scales. Originally, Wiener and Harmon intended to develop obvious and subtle keys for each clinical scale, but results for Sc, Pt and Hs were mainly negative, as these scales are comprised primarily of obvious items.

It was found that the Wo and Ws scales were uncorrelated. A high positive correlation, approximately +.60, was found between the various Wo scales, and a lower positive correlation, approximately +.21, was found between Ws scales. Correlations between Wo minus Ws scores and the K scale were found to be highly negative, ranging from –.47 to –.78. This shows that low K scale score and high Wo–Ws scale scores indicate similar ("fake bad") test taking attitudes. Conversely, high K scores and low Wo – Ws scale scores indicate similar ("fake good") test taking attitudes.

Wiener and Harmon also found individuals with higher intelligence and education obtained approximately equal scores on Wo and Ws, while individuals with lower intelligence and less education obtain higher Wo scores than Ws scores. Psychologically sophisticated individuals in a psychology training program successfully avoided Wo items while Ws T–

scores were average or above average. It was also found, in a group of veterans in school and on–the–job training, that elevation of Wo scores predicted failure, while elevated Ws scores were not significantly related to either success or failure.

Anthony (1971) found that male military clients with nonpsychotic disturbances were successful in manipulating obvious items when attempting to "fake bad." However, subtle items reflected less pathology when "faking bad" than when clients took the test honestly. With a raw cutoff score of 36, the subtle scale had a hit rate of 56% in correctly identifying exaggerated and valid matching profiles, while the obvious scale, with a raw cutoff score of 170 indicated a hit rate of 59%. Although the obvious–subtle scales differentiated honest and exaggerated profiles, it was found the Wo and Ws scales were not successful in distinguishing exaggerated profiles from bona fide profiles with the same configurations.

Harvey and Sipprelle (1976) found that male and female university students simulating job applicants for a highly desired job ("faking good"), produced significantly higher mean Ws scores than Wo scores. Conversely, students simulating evaluation for desired psychotherapy ("fake bad"), produced significantly higher mean Wo scores than Ws scores. Also, the mean of the combined Wo and Ws scores was found to be higher for the

psychotherapy group than for the job group. Harvey and Sipprelle suggest the Wo and Ws scales are a function of the demand characteristics of the situation in a normal population. Individuals are likely to respond to MMPI items according to perceived social desirability for that particular situation. Therefore, the Wo and Ws scales appear useful in determining which situations elicit defensive strategies for an individual.

Gendreau, Irvine and Knight (1973) found that the Wo scale significantly discriminated between honest and "fake bad" instructional sets in a sample of prisoners. No explicit criteria for defining either "fake bad" or "fake good" was provided in this study. Mean Ws and Wo raw scores in the honest condition were 35.1 and 42.5 respectively, for "fake bad" condition, 34.8 and 87.1 and for "fake good" 44.0 and 19.3. The Ws scale was not influenced by "fake bad" instructions, while the Wo scale effectively discriminated between honest and "fake bad" sets. Gendreau et al. report that the hit rate for the Wo scale was 88%, while in the "fake good" sets, the hit rates were Wo=81% and Ws=78%. Both scales were effective in differentiating "fake good" from "honest" responding.

Grow, McVaugh and Eno (1980) found with a sample of university students that the use of Wo $\geq$ 100, a technique to detect "faking bad" correctly identified 48% of those faking bad while not incorrectly identifying any in the the "fake good" or honest anonymous groups,

showing a total correctly identified of 83%. Subtle items ≤ 45, another technique to detect "faking bad" was significantly less effective. Three "faking good" techniques were examined. Obvious items ≤ 65 correctly identified 100% of "faking good," incorrectly identified 16% of "faking bad," incorrectly identified 94% of "honest anonymous" profiles for a total correctly identified profiles of 63%. Subtle items ≥ 61 were significantly less effective. The most successful technique was obvious − subtle items ≤ −4. This technique correctly identified 100% of "faking good," incorrectly identified 6% of "faking bad" and 86% of anonymous for a total correctly identified of 69%. A second experiment was conducted with a clinical population which had been judged as being indicative of either "faking good," "faking bad" or straightforward. It was found that subtle items ≤ 45 were superior to obvious items ≥ 100 in detecting "faking bad." This is opposite from findings with a university sample where obvious items were more effective in detecting "fake bad" profiles. Results of the techniques to detect faking good showed the same order of effectiveness for the three techniques. However, each technique was significantly more effective with the clinical sample. Obvious − subtle items ≤ −4. correctly identified a total of 91% of the profiles, obvious items ≤ 65 identified 85% and subtle items ≥ 61 identified 70%. Grow et

al. suggest Wiener's Wo and Ws items show promise for the detection of faking good with both student and clinical groups.

Wales and Seeman (1968) examined subtle and obvious items in a slightly different manner. They found subtracting the deviant number of obvious items from the deviant number of subtle items, i.e., Ws–Wo , was useful for detecting "fake good" response sets, they determined 315 obvious items which were endorsed by a minority of the original MMPI normative group, and 84 subtle items which were scored for abnormality, even though they were scored in the deviant direction by the majority of the original normative group. Christian, Burkhart and Gynther (1978) developed another method of defining subtle and obvious items on the MMPI. College students were asked to rate each item on a 5–point scale from very obvious to very subtle, measuring the degree to which the item indicated a psychological problem. These scales show promise and more investigation has recently occurred to determine the generalizability to other populations and the relative predictive validity of obvious, neutral and subtle items on the various MMPI scales (Burkhart, Gynther, & Fromuth, 1980; Hovanitz & Gynther, 1980; Hovanitz, Gynther, & Green, 1985; Hovanitz, Gynther, & Marks, 1983; Hovanitz, Jordan, & Brown, 1986).

Overall, research on the Wo and Ws scales suggests that they can be useful in the detection of "faking bad" and particularly in the detection of "faking good." However, explicit criteria for defining "fake bad" and "fake good" profiles are questionable. Wiener (1948) assigned T–scores to the Wo and Ws scales, but did not suggest specific cutoff scores. Consequently, a variety of cutoff scores, some raw scores and some T–scores have been tried. In a number of articles it is unclear whether T–scores or raw scores have been used. The most effective means of detecting a "fake good" response set appears to be the decision rule, obvious – subtle items ≤ -4. Greene (1980) suggests it is best to assume a "fake good" response set when all five subtle scales are elevated to a T–score of 70 or greater while all obvious scale T–scores remain around 50. Conversely, the opposite relationship between Wo and Ws scales should suggest a "fake bad" response set. Greene (1983) has also suggested that Wo – Ws ≥ 149, and Wo – Ws ≤ -59 may be useful in detecting "fake bad" and "fake good" profiles.

Summary

Review of the research literature has indicated some support for the usefulness of the traditional validity scales. However, elevated scores on these scales have been shown to be related to factors other than validity. For example, the L scale scores appear related to cultural factors (as blacks, Mexican–Americans, Chinese and Japanese tend to score higher than whites), socioeconomic class, and level of education and sophistication. Good and Brantner (1974) indicates that high L T–scores (>70) likely reflected only the most naive and blatant attempts at faking by those who were intellectually, educationally, culturally, or economically deprived. The usefulness of the L scale therefore appears fairly restricted.

F scale raw scores greater than 16 (T–scores >70) are purported to detect invalidity due to exaggeration of problems or deliberate "fake bad" response sets (Hathaway and McKinley, 1967). Simulation studies have provided evidence that F T–scores > 70 can detect "faking bad" (Gough, 1947; Grow, McVaugh & Eno, 1980). Therefore, many clinicians routinely discard MMPI profiles as invalid when F T–scores exceed 70. However, experimental findings indicate F scale scores are related to such factors as age, intelligence, race and diagnostic classification. Of particular importance is the finding that generally a high F score is an indication of

significant and severe psychopathology (Kazan & Sheinberg, 1945; Schneck, 1948). It appears as if elevated F scores in psychiatric settings reflect particular psychopathology, i.e., psychosis and behavior disorders, rather than an indication of "plea for help" response sets or blatant "faking bad" response sets. A number of researchers have found a positive relationship between elevated F scores, younger aged males, behavior disorders and aggressive behaviors. It is felt that item content is such that most delinquents can honestly and accurately achieve F T–scores greater than 70 (McKegney, 1965). Elevated F scores may also reflect random responding due to carelessness or lack of cooperation.

The K scale was developed as a means of detecting subtle defensiveness and as a means of correcting a defensive test–taking tendency. Subsequent studies have confirmed that elevated K scale scores may suggest defensiveness in a maladjusted population (Smith, 1959). However, elevated scores also appear related to socioeconomic class and educational levels. Graham (1977) states a K T–score > 70 for persons with higher status or education may be considered defensive while persons with lower status or education may be considered defensive when K T–scores are greater than 60. Conversely, low K scale scores are suggestive of a lack of defensiveness or a tendency to be self–critical and

exaggerate problems. Again, research findings indicate that socioeconomic status and educational levels must be considered in the interpretations.

The F–K index is routinely used to detect "faking bad" and "faking good" attempts. Studies suggest the F–K index is more successful at detecting "fake bad" attempts (Hunt, 1948; Gough, 1950). It was suggested by Gough (1950) that an F–K index > 9 should be considered invalid due to "faking bad." However, Gendreau et al. (1973) suggest a cutoff score of F–K ≥ 24 is more appropriate with a prisoner population, as this population may validly elevate the F–K index. Post and Gasparkiova–Krasnec (1979) suggest that F–K ≥ 11 profiles can be interpreted in ways other than "faking bad." The difficulty in detecting "faking good" lies in the degree of overlap between "fake good" and honest responding. The problem appears related to the K scale, which is elevated when persons of higher economic or educational levels respond honestly (Hunt, 1948). Overall, the traditional validity measures may effectively detect simulation, particularly where "faking bad" is concerned. However, there are confounding factors which restrict the confidence one can place on interpretations.

The problems with traditional validity measurement interpretations, as cited in this review, provided the impetus for the development of new validity measures to be used additionally or alternatively to the traditional

measures (Greene, 1980). The TR index and the Carelessness scale were developed to measure response consistency, and they have been used as supplements to the standard validity measures. Research has shown that the TR index is useful in detecting random responding particularly with adolescent populations where F scale scores are elevated and the possibility exists that F scores are elevated validly. Of particular significance was the finding that 8% to 16% of valid profiles (F<71) were classified as invalid on the TR index (Greene, 1979). The distinction between invalid profiles due to random responding (reading problems, confusion or a blatant refusal to cooperate) and a "plea for help" response set may be of importance to clinicians. The TR index has been found to distinguish between these content–dependent and content–independent response sets. The Carelessness Scale has also been found to correctly classify a large majority of random responders and non–random responders (Greene, 1978). Although little research has been conducted on this newest validity indicator, it holds promise of detecting randomness effectively in more sophisticated individuals.

The Ds scale was designed to detect individuals attempting to emulate particular psychopathological roles. Research has shown that it is successful in differentiating honest and "fake bad" responders.

The Mp scale was constructed as a subtle lie key for the detection of "fake good" profiles. Results show that the Mp scale is very effective in distinguishing "fake good" from "honest" responders, a task not very successfully accomplished by the traditional validity measures. Although research with the Mp scale is extremely sparse, results to date appear consistently positive and encouraging.

The Wo and Ws scales, developed to detect both "fake good" and "fake bad" simulation, were based on the premise that individuals were capable of manipulating obvious items, but were unable to manipulate subtle items. Particular cutting scores to determine simulation were not provided in the original research. Research studies have typically utilized various cutoff scores for both the Wo and Ws scales with some success. The decision rule to subtract subtle items from obvious items has received considerable attention and appears promising as an effective means of detecting both "fake good" and "fake bad" response sets.

Overall, the new validity measures appear useful in detecting simulation. Their use as an adjunct or alternative to traditional MMPI validity scales may reduce the interpretation uncertainty caused by the confounding factors contained within the traditional scales. The L, F, and K scales are routinely used, as templates are available for scoring, as well as space provided on the answer and profile sheets for recording these

scores. The new validity indicators, on the other hand, do not appear to be routinely used. This may be partially due to lack of convenience, uncertainty regarding appropriate cutting scores, and general uncertainty regarding the usefulness of including these additional validity measures in routine scoring procedures.

The purpose of the present study is to compare the traditional approaches to the measurement of MMPI validity with the new set of scales that show a priori potential for being better validity indicators.

The following experimental hypotheses are proposed:

1. The new validity measures are significantly better at detecting "fake bad" MMPI profiles than are the traditional scales and indices.
2. The new validity measures are significantly better at detecting randomly answered MMPI's than are the traditional validity scales and indices.
3. The new validity measures are significantly better at detecting "fake good" MMPI profiles than are the traditional scales and indices.

CHAPTER III

METHODOLOGY

Subjects

Subjects were 16 male and 44 female students at the University of Calgary, and 16 male and 24 female psychiatric inpatients at Holy Cross Hospital. Two samples were used to maximize the difference within conditions. It was felt that differences might not be significant between university students responding honestly and faking good, or psychiatric inpatients responding honestly and faking mental illness. University students were recruited in the classrooms. All students volunteered to be tested on two occasions, one week apart. University students received $10.00 each for their participation in the study. Ages of the university students ranged from 18 to 47 years, with a mean age of 30.7. Approximately 60% were undergraduate education students, while 40% were teachers obtaining upgrading. All students were recruited from educational psychology classes.

Psychiatric patients were recruited in their wards and volunteered to be tested on two occasions, one week apart. The large majority of patients were diagnosed as depressive (27) with varing degrees of severity, while 7 were diagnosed as schizophrenic. The six remaining patients

received one of the following diagnoses: impulse control disorder, hypomania, hypochondriasis, anxiety neurosis, toxic induced psychosis and situational crisis. Approximately 85% of patients were receiving drug treatment. Most patients were receiving individual therapy, group therapy or combinations of both. Twelve patients were on social assistance, 11 patients held blue collar or clerical jobs prior to hospitalization, 5 were housewives, 1 was a registered nurse, 3 held supervisory positions, and the remaining 8 patients did not indicate any occupation. In all, at least 50%, and possibly as high as 75% of patients were assumed to be of lower–middle to middle–class socioeconomic status. Ages of psychiatric inpatients ranged from 18 to 58 years, with a mean age of 33.17.

Because the performance of the two groups (university students and psychiatric patients) was not going to be compared, it was not necessary to make sure they were matched for sex and age. Information regarding sex, age and socioeconomic status of patients was obtained for descriptive purposes only. Before the study began, the subjects were informed of the general nature of the task and were assured of confidentiality (see Appendix A).

Instrument

Form R of the MMPI, which consists of 566 statements to be endorsed as either true (T) or false (F), was used. Each item (statement) of the inventory contributes to the composition of one or more of 10 clinical scales. Extent of psychopathology is indicated by the configuration of scale elevations. Items also contribute to the composition of three validity scales (L, F, and K). Elevations of these scales and the relationship between the F and K scales have historically been used to index the validity of the clinical scales, with elevations above a certain magnitude precluding interpretation of the clinical scales, as discussed in the previous chapter. The following criteria for invalidation were used: L (raw score > 7) T–score > 70 (Hathaway & McKinley, 1967); K T–score > 65 (Greene, 1980) and F–K raw score < 0 (Gough, 1950) for the detection of "faking good," and F (raw score > 16) T–score > 70 (Hathaway & McKinley, 1967) and F–K raw score > 9 (Gough, 1950) for the detection of "faking bad."

All 200 tests were hand scored and the appropriate scales were K–corrected. In addition, all tests were scored for Gough's (1957) Ds–r scale, Wiener's (1948) subtle and obvious scales (Wo and Ws), Cofer's (1949) Mp scale, Buechley and Ball's (1952) TR index and Greene's (1978) Carelessness scale. The following criteria for invalidation, recommended by

Greene (1983) as a new procedure for assessing MMPI validity, were utilized: Ds–r T–score ≥ 70 and Wo minus Ws T–scores ≥ 149 for the detection of "faking bad"; Mp T–score ≥ 65 and Wo minus Ws T–scores ≤ –59 for the detection of "faking good"; and TR raw score ≥ 6 and Carelessness raw score ≥ 6 for the detection of random responding.

Procedure

The MMPI was initially administered to 60 University undergraduate students and 40 psychiatric inpatients following standard instructions. The test was administered to the students in small groups whereas the psychiatric patients were assessed individually. One week later a subgroup of students (N=40) completed the MMPI under "fake bad" instructions as follows:

> You will recall doing this inventory recently. This time when you are answering, you are to complete the test as if you are a very maladjusted or mentally ill person. Answer those items true or false which you feel are descriptive of such a person. Remember, all you have to do is answer as if you were maladjusted or mentally ill.

The other student subgroup (N=20) completed the MMPI under instructions to answer in a random fashion, as follows:

> When completing the test I want you to respond in a careless, haphazard manner as if you were bored or uninterested in the test. Remember, all you have to do is complete the form in an uninterested and careless manner.

The psychiatric group (N=40), on their second administration of the MMPI, were requested to respond under a "fake good" instructional set, as follows:

> Today I would like you to do the test again, only this time I would like you to answer as if you have no problems, as if there is nothing wrong with you. That is, try to give a very good impression of yourself on the test. Remember, answer as if you were completely well and had no problems.

After each set of instructions, clarification and extended detail were provided upon request. Upon completion of each test, answer sheets were checked to ensure that not more than 5 items were left unanswered. Clopton and Neuringer (1977) and Greene (1983) maintain that the omission of 5 items or less is not likely to lead to significant profile distortion. None of the answer sheets had more than 5 unanswered items.

The order in which the tests were given was not varied, as studies (Cofer, Chance & Judson, 1949; Gough, 1950; Gendreau, Irvine & Knight, 1973; Hunt, 1948) have shown that the order of administration has no significant effect on results.

Statistical Design

1. Comparisons of mean scores on each of the "traditional" and "new" validity measures appropriate for each group were made between "honest" conditions and "faking" conditions.
2. The McNemar test was used to determine if the number of people (frequencies) who received acceptable and unacceptable scores on the appropriate validity measures differed under "honest" and "faking" conditions. The McNemar test was used because the scores derived under the two conditions were not independent. The number of misclassifications under "honest" and "faking" conditions were determined.
3. All the validity indices were intercorrelated in order to ascertain the strengths and directions of their relationships.

4. A Wilks' Stepwise Discriminant Function Analysis was used to determine which validity scales could best distinguish "honest" responders from "fakers" on the MMPI. This procedure maximizes the differences between groups while minimizing differences within groups.

CHAPTER IV

RESULTS

Comparisons were made between mean T–scores on the L, K, Mp and Wo–Ws scales and between mean raw scores on the F–K index for the "fake good" group. Results of these t–tests, shown in Table 1, indicate a highly significant difference ($p<.001$) between mean scores on all validity scales under "honest" and "fake good" instructions. Results of t–tests comparing mean scores on the F, F–K, Ds–r and Wo–Ws scales for the "fake bad" group shows a highly significant difference ($p<.001$) between mean scores on all validity scales under "honest" and "fake bad" instructions (see Table 2). Results of t–tests comparing mean scores on the F, F–K, TR and Carelessness (C) scales for the random group show a highly significant difference for F ($p=.003$), F–K ($p=.002$) and TR ($p=.003$) scales under "honest" and "random" instructions. A significant difference ($p<.05$) is shown between C scale scores (see Table 3).

Results of the McNemar Test for the "fake good" group are presented in Table 4. Highly significant differences between classification under "honest" and "faking" conditions was found on the L, K, F–K, Mp and Wo–Ws scales. The L scale ($L>70$) did not misclassify any of the honest responders while misclassifying 28 of the "fake good" responders for a total of 28 (35%) misclassifications.

TABLE 1

Results of t–Tests for the "Fake Good" Group

Variables	Number of cases	Mean	SD	t Value	df	2 Tail Probability
L (1)		51.83	7.99			
	40			−4.99	39	<.001
L (2)		62.30	11.87			
K (1)		50.10	8.38			
	40			−5.09	39	<.001
K (2)		60.13	10.68			
F–K (1)		2.38	10.72			
	40			6.46	39	<.001
F–K (2)		−10.36	10.90			
Mp (1)		50.65	9.41			
	40			−5.69	39	<.001
Mp (2)		65.56	13.03			
Wo–Ws (1)		74.28	83.28			
	40			6.78	39	<.001
Wo–Ws (2)		−40.80	81.74			

(1) refers to "honest" condition
(2) refers to "fake good" condition

TABLE 2

Results of t–Tests for the "Fake Bad" Group

Variables	Number of Cases	Mean	SD	t Value	df	2 Tail Probability
F (1)		54.38	8.63			
	40			−22.16	39	<.001
F (2)		112.75	15.12			
F–K (1)		−10.40	8.79			
	40			−17.00	39	<.001
F–K (2)		38.93	16.32			
Ds–r(1)		49.56	9.39			
	40			−20.14	39	<.001
Ds–r(2)		107.28	16.20			
Wo–Ws (1)		−32.63	58.06			
	40			−21.25	39	<.001
Wo–Ws (2)		234.58	61.62			

(1) refers to "honest" condition
(2) refers to "fake bad" condition

TABLE 3

Results of t–Tests for the "Random" Group

Variables	Number of cases	Mean	SD	t Value	df	2 Tail Probability
F (1)		54.95	8.51			
	20			−3.35	19	.003
F (2)		73.95	26.46			
F–K (1)		−10.65	7.21			
	20			−3.47	19	.002
F–K (2)		1.95	15.95			
TR (1)		0.80	0.77			
	20			−3.47	19	.003
TR (2)		3.60	3.49			
C (1)		1.45	1.32			
	20			−2.73	19	.013
C (2)		2.75	2.02			

(1) refers to "honest" condition
(2) refers to "random" condition

TABLE 4

Results of McNemar Tests for the "Fake Good" Group

	acceptable(Valid) (if L ≤ 70)	unacceptable (Invalid) (if L > 70)	
Honest	40	0	$\chi^2(1)=10.08$
Fake good	28	12	p=.001
	acceptable(Valid) (if K ≤ 65)	**unacceptable(Invalid) (if K > 65)**	
Honest	36	4	$\chi^2(1)=9.60$
Fake good	23	17	p=.002
	acceptable(Valid) (if F−K ≥ 0)	**unacceptable(Invalid) (if F−K < 0)**	
Honest	24	16	$\chi^2(1)=14.45$
Fake good	6	34	p<.001
	acceptable(Valid) (if Mp < 65)	**unacceptable(Invalid) (if Mp ≥ 65)**	
Honest	38	2	$\chi^2(1)=15.43$
Fake good	19	21	p<.001
	acceptable(Valid) (if Wo−Ws > −59)	**unacceptable(Invalid) (if Wo−Ws ≤ −59)**	
Honest	37	3	$\chi^2(1)=16.06$
Fake good	19	21	p<.001

The K scale (K > 65) misclassified 4 honest and 23 "fake good" responders for a total of 27 (34%) misclassifications, while the F–K scale (F–K<0) misclassified 16 honest and 6 "fake good" responders for a total of 22 (28%) misclassifications. The Mp scale (Mp ≥ 65) misclassified 2 honest and 19 "fake good" responders for a total of 21 (26%) misclassifications and the Wo–Ws scale (Wo–Ws ≤ 59) misclassified 3 honest and 19 "fake good" responders indicating a total of 22 (28%) misclassifications.

Table 5 presents results of the McNemar Test for the "fake bad" group. A highly significant difference between classification under honest and "fake bad" instructions was found on the F, F–K, Ds–r and Wo–Ws scales ($p<.001$). The F scale (F > 70) misclassified 2 honest and 2 "fake bad" responders for a total of 4 (5%) misclassifications while the F–K scale (F–K > 9) misclassified 2 honest and 3 "fake bad" responders for a total of 5 (6%) misclassifications. The Ds–r scale (Ds ≥ 70) misclassified 2 honest and 1 "fake bad" responder for a total of 3 (4%) misclassifications and the Wo–Ws scale (Wo–Ws ≤ 149) misclassified 1 honest and 2 "fake bad" responders for a total of 3 (4%) misclassifications.

Table 6 presents the results of the McNemar Test for the random group. The difference between classification under honest and random instructions was not found to be significant on the TR and C

TABLE 5

Results of McNemar Tests for the "Fake Bad" Group

	acceptable(Valid) (if F ≤ 70)	unacceptable(Invalid) (if F > 70)	
Honest	38	2	$\chi^2(1)=34.03$
Fake bad	2	38	$p<.001$
	acceptable(Valid) (if F−K ≤ 9)	**unacceptable(Invalid) (if F−K > 9)**	
Honest	38	2	$\chi^2(1)=33.03$
Fake bad	3	37	$p<.001$
	acceptable(Valid) (if Ds−r < 70)	**unacceptable(Invalid) (if Ds−r ≥ 70)**	
Honest	38	2	$\chi^2(1)=35.03$
Fake bad	1	39	$p<.001$
	acceptable(Valid) (if Wo−Ws < 149)	**unacceptable(Invalid) (if Wo−Ws ≥ 149)**	
	39	1	$\chi^2(1)=35.03$
Fake bad	2	38	$p<.001$

TABLE 6

Results of McNemar Tests for the "Carelessness" Group

	acceptable(Valid) (if C < 6)	unacceptable(Invalid) (if C ≥ 6)	
Honest	20	0	Binomial
Random	18	2	p=.50

	acceptable(Valid) (if TR < 6)	unacceptable(Invalid) (if TR ≥ 6)	
Honest	20	0	Binomial
Random	16	4	p=.13

	acceptable(Valid) (if F ≤ 70)	unacceptable(Invalid) (if F > 70)	
Honest	19	1	Binomial
Random	7	13	p<.04

	acceptable(Valid) (if F−K ≤ 9)	unacceptable(Invalid) (if F−K > 9)	
Honest	20	0	Binomial
Random	7	13	p<.02

scales; however, a significant difference ($p<.05$) was found on the F and F–K scales. The F scale ($F > 70$) misclassified 1 honest and 7 random responders for a total of 8 (20%) misclassifications while the F–K scale ($F-K > 9$) did not misclassify any honest responders but misclassified 7 of the random responders for a total of 7 (18%) misclassifications. The TR index ($TR \geq 6$) did not misclassify any honest responders but misclassified 16 of the random responders for a total of 16 (40%) misclassifications and the C scale ($C \geq 6$) did not misclassify any honest responders but misclassified 18 of the random responders, for a total of 18 (45%) misclassifications.

Analysis of the data in Table 6 indicated the TR index and Carelessness scales were ineffective in identifying random responders as such, therefore disconfirming the second hypothesis that these scales are significantly better at detecting randomly answered MMPI's than the traditional validity measures. Therefore, the random group was not included in further analysis.

Tables 7 and 8 present the matrix of correlations between the validity indices for the "fake good" and "fake bad" groups. Results indicate the following highly significant relationships between validity scale scores in the "honest" condition: K and F–K, –.78; K and Wo–Ws, –.79; F–K and Wo–Ws, .87. The following highly significant

TABLE 7

Pearson Product Moment Correlations

Correlations for "Fake Good" Group

	L(2)	K(1)	K(2)	F−K(1)	F−K(2)	Mp(1)	Mp(2)	Wo−Ws(1)	Wo−Ws(2)
L(1)	.15	.63***	.13	−.53***	.003	.45**	−.09	−.46***	.01
L(2)		−.05	.62***	.03	−.48***	.14	.75***	.03	−.54***
K(1)			.16	−.78***	−.06	.25	−.24	−.79***	−.10
K(2)				−.27*	−.85***	−.21	.50***	−.16	−.91***
F−K(1)					.34*	−.21	.27	.87***	.27*
F−K(2)						.25	−.48***	.19	.92***
Mp(1)							−.07	−.32*	.29*
Mp(2)								.32*	−.53***
Wo−Ws(1)									.15
Wo−Ws(2)									

Note: (1) refers to "honest" condition

(2) refers to "fake good" condition

*p<.05 **p<.01 ***p<.001

TABLE 8

Results of Pearson Product Moment Correlations

Correlations for "Fake Bad" Group

	F(2)	F–K(1)	F–K(2)	Ds–r(1)	Ds–r(2)	Wo–Ws(1)	Wo–Ws(2)
F(1)	.10	.73***	.12	.69***	.06	.64***	.20
F(2)		.01	.87***	.27**	.72***	.25	.70***
F–K(1)			.02	.60***	–.07	.69***	.10
F–K(2)				.19	.85***	.20	.87***
Ds(a)					.07	.80***	.14
Ds(2)						.08	.87***
Wo–Ws(1)							.12
Wo–Ws(2)							

Note: (1) refers to "honest" condition
(2) refers to "fake bad" condition

*$p<.05$ **$p<.01$ ***$p<.001$

relationships were found between validity scale scores in the "fake bad" condition: L and Mp, .75; K and F–K, –.85; K and Wo–Ws, –.91, and F–K and Wo–Ws, .92. The very high relationships between K and F–K is understandable due to the commonality of K with F–K. However, the extremely high relationships between the other indices are more difficult to interpret (see discussion).

Table 9 presents the results of Wilks' stepwise discriminant analysis on the "fake good" group. The Mp and Wo–Ws scales, in that order, were most effective in distinguishing honest and "fake good" responders. The L, K and F–K indices were less effective in discriminating between groups. It was found that 85.5% of those in Group 1 (honest condition) were predicted to be in Group 1, and 72.5% of those in Group 2 (faking condition) were predicted to be in that group. This resulted in 77.5% of cases being correctly classified.

Table 10 presents the results of the stepwise discriminant analysis on the "fake bad" group. The F–K index was not included in this analysis because of multicollinearity, i.e., discriminant functions could not be computed when the F–K index was included in the analysis. All the other three scales, (F, Ds–r and Wo–Ws) were highly successful in differentiating honest and "fake bad" responders on their own but only F and Wo–Ws were retained in the discrimant function. It was found that 97.5% of those in Group 1 (honest responders) were predicted to be in that group, while 95% of those in Group 2 (those "faking bad") were predicted to be in Group 2. Overall, 96.25% of cases were correctly classified.

TABLE 9

Results of Discriminant Analysis on "Fake Good" Group

Significant Variables in Equation

Variable	Wilks' Lamda	p
Mp	.69	<.001
Wo–Ws	.62	<.001

Discriminant Function Equation

Discriminant score (standardized) = .61(Mp) –.58(Wo–Ws)

Canonical correlation = .62

Wilks' Lamda = .62

Chi–square (2) = 36.83, p<.001

Classification Results

Actual Group	No of Cases	Predicted Group Membership 1	2
Group 1	40	33 82.5%	7 17.5%
Group 2	40	11 27.5%	29 72.5%

Percent of "Grouped" cases correctly classified: 77.5%

Group 1 = Honest responders

Group 2 = "Fake good" responders

TABLE 10

Results of Discriminant Analysis on "Fake Bad" Group

Significant Variables in Equation

Variable	Wilks' Lamda	p
F	.15	<.001
Wo–Ws	.13	<.001

Discriminant Function Equation

Discriminant score (standardized) = .63(F) .47 (Wo–Ws)

Canonical correlation = .93

Wilks' Lamda = .13

Chi–square(2) = 155.9, p<.001

Classification Results

Actual Group	No. of Cases	Predicted Group Membership 1	2
Group 1	40	39 97.5%	1 2.5%
Group 2	40	2 5.0%	38 95.0%

Percent of "Grouped" cases correctly classified: 96.25%

Group 1 = Honest responders

Group 2 = "Fake bad" responders

CHAPTER V

DISCUSSION

The usefulness of the MMPI has been dependent upon the scrupulousness of the responder or on the sensitivity of the traditional validity measures to detect dishonest or careless responding. New validity measures have been developed from the MMPI items which have been claimed to be more effective at detecting "random," "fake–bad," and "fake–good" response sets. In this study comparisons of the traditional and "new" validity indices were made to determine which were more effective in detecting these invalidating response sets, and the discriminatory power of combinations of the various indices, both new and old, was examined. Results differed in each of the three experimental conditions.

Random Condition

The random responding instructional set was designed to test the effectiveness of the TR and C indices against the traditional F and F–K measures. Results of the t–tests between honest and random conditions show a highly significant difference for F, F–K and TR scores under honest and random instructions. A significant difference ($p<.02$) was also

found for the C scores under the two different instructional sets. All differences were in the expected direction, strengthening the implicit assumption that responders were answering honestly under standard instructions.

Results of the McNemar tests showed a significant difference ($p<.04$) between classification under honest and random instructions on the F scales and F–K index. The F scale misclassified 20% of the responders, with one profile in the honest condition classified as invalid and 7 profiles in the random condition classified as valid, The F–K index misclassified 18%, with seven profiles in the random condition classified as valid. No significant difference was found between classification under honest and random instructions on the TR and C scales. The TR index misclassified 80% of random responders, with 16 profiles in the random condition classified as valid. The C scale misclassified 90% of random responders, with 18 profiles in the random condition classified as valid. This is contrary to expectations and refutes the usefulness of the TR index and the C scale in the detection of random responding. There are two possible ways of interpreting these latter findings. First, following the second (random) MMPI administration, several of the University students verbally expressed considerable difficulty in responding carelessly and haphazardly. It seems possible that a population accustomed to test–

taking in a careful and conscientious manner may have found it impossible to respond randomly. The second consideration is related to the cutoff scores used with the TR index and C scale. Originally, Buechley and Ball (1952) suggested a cutoff score of 3 on the TR index. Later, Dahlstrom and Dahlstrom (1972) and Greene (1979) found TR $\geq$ 4 most effective in detecting random responding. The results in this study showed a mean score of 3.60 (SD = 3.49) in the random condition. A cutoff score of 4 would have identified 4 additional random responders or an additional 20%. Similarly, the cutoff score of 6 on the C scale appears too high. The results showed a mean score of 2.75 (SD = 2.02) in the random condition. A cutoff score of 4 would have identified 3 additional random responders, or an additional 15%. Greene (1979) originally suggested a cutoff score of 4, but recently has revised this to a score of 6 (Greene, 1983). It appears from these results that a cutoff score of 3 or 4 would be more appropriate than a cutoff score of 6.

Fake–Bad Condition

The t–test results showed a highly significant difference ($p<.001$) in the expected direction between all validity measures under honest and "fake bad" conditions. Results of the McNemar tests showed a highly significant difference ($p=.001$) between classification under honest and "fake bad" instructions on all validity measures. The F scale only misclassified

5% of responders, the F–K index misclassified 6% and both the Ds–r and Wo–Ws scales misclassified 4% of responders. All validity indices were therefore highly effective. The general nature of the "fake–bad" instructions resulted in a blatant mode of exaggeration which produced highly maladjusted profiles that were easily detected by all validity indices. Possibly more specific instructions, as in role–playing particular disorders would have produced less exaggerated clinical profiles and lower scores on validity measures. However, there is no consensus regarding the method (general or specific) of faking commonly used by psychiatric patients. In establishing instructional sets to "fake–bad," the assumption was made that psychiatric patients initially dissimulate on the MMPI in a general manner. However, multiple admission patients may become more adept at feigning particular syndromes and utilize a specific role–playing style of dissimulation.

Correlations between F and F–K in both honest and "faking–bad" conditions were positive and highly significant. As the F–K index contains the F scale, a high correlation is to be expected. The other highly related validity measures are F and Ds–r, F–K and Ds–r, F–K and Wo–Ws and Ds and Wo–Ws. Although these validity scales are not measuring exactly the same thing, (otherwise the correlation would be 1.00), the concepts of "unusual response" of the F scale, "stereotypic view

of neuroticism" of the Ds–r scale, and "obviously deviant" view of the Wo scale appear to have a large common element. This element may comprise the stereotype of maladjustment or mental illness held by many lay people, whom the subjects in this study would represent. Since instructions to "fake–bad" requested this stereotypic view of a generally mentally ill person, and all validity measures appear to reflect this view, it is not surprising that all validity indices were extremely effective.

Results of the discriminant analysis indicate that the F scale and Wo–Ws scale, taken together, are the most effective means of detecting "faking–bad." They together correctly classified 96.25% of cases. Overall, results showed all validity indicators are extremely effective in correctly classifying "fake–bad" responders when individuals attempt to feign non–specific mental illness.

Fake–Good Condition

Results of the t–tests of the difference between honest and "fake–good" conditions showed a highly significant difference between all validity measures under both of these conditions. The mean score was within the acceptable (valid) range for all validity measures on the honest administration.

Results of the McNemar tests showed a highly significant difference between classification under honest and "fake–good" conditions on all

validity measures. The L scale correctly classified 100% of honest responders but misclassified 65% of those faking good. These results are consistent with Good and Brantner's view (1974) which suggests that the average individual is willing to admit to the minor social faults implicit in the L scale, and is not likely to invalidate the L scale when "faking–good." The K scale correctly classified 90% of honest responders but misclassified 58% of those "faking–good." The results show the mean K scale score in the "faking–good" condition was within the acceptable range. This indicated that the K scale did not function well in detecting deliberate "fake–good" responding. To the extent that patients were of a lower middle socioeconomic status, results may be explained in those terms. Graham (1977) indicated that lower middle class individuals typically obtain T–scores ranging between 40 and 60 on the K scale, and a T–score greater than 60 should be considered high. In this study, patients increased their mean T–scores from 50.10 to 60.13. Using a cut–off T–score of 65 on the K scale may have been inappropriate for lower middle class persons. The F–K index misclassified 40% of honest responders while correctly classifying 85% of those "faking–good," with 16 of the honest profiles misclassified as "fake–good." These findings are consistent with other studies (Exner et al., 1963; Gough, 1950; Hunt, 1948)

which found the F–K index successful in identifying "fake–good" attempts but, unfortunately, also classifying a large percentage of honest responders as dissimulators.

Results of the Wo–Ws and Mp scales were more encouraging. The Wo–Ws scale correctly classified 93% of honest responders and 52% of those "faking–good," for a total of 72% correctly classified. This is a 6% and 7% improvement over the K and L scales, respectively, in terms of correct classification. The Mp scale correctly classified 95% of honest responders and 52% of those "faking–good," for a total of 76% correctly classified. These results are almost identical to those of Wiggins (1959) and Cofer et al. (1949) regarding identification of honest responders. They are somewhat lower regarding identification of "fake–good" responders but still suggest that the Mp scale is most effective in distinguishing honest from "fake–good" responders, with a 10% and 11% increase in correct classification over the L and K scales, respectively. Possibly, if different cut–off scores were used on the Wo–Ws and Mp scales, a larger percentage of "fake good" responders would have been identified.

The correlations between the various validity measures were examined to determine whether a common element is evident in all indices or whether a number of factors were related to "faking good." Correlations between K and F–K were highly significant in both honest

and "fake–good" conditions. This is to be expected as the F–K index contains the K scale. A highly significant relationship was found between the L and Mp scale (r=.75). Similarly, a highly significant relationship was found between the K and Wo–Ws and the F–K and Wo–Ws scales. Wiggins (1959) examined the interrelationships among MMPI dissimulation measures under standard and social desirability conditions and found that the L and Mp scales were highly related to social desirability response sets, and that the Mp scale was more effective than the L scale in identifying social desirability dissemblers. In the present study the strong relationship between the L and Mp scales may therefore be interpreted in terms of social desirability response sets. Wiggins (1959) also found the K scale to be related to an "acquiescence" (problem–denial) response style or bias. He also found the Mp scale to be unrelated to this bias. As well, Wiggins also found that instructions to dissimulate increased the use of the acquiescence response set. In the present study, high correlations were found between K, F–K, and Wo–Ws scales. It is suggested that each of these measures reflect an acquiescence response style, but that this bias is not reflected in the L and Mp scales. The superiority of the Mp scales and the Wo–Ws scales over the L and K scales in detecting "fake–good" responding may reflect their greater effectiveness in measuring social desirability response sets and acquiescence response sets, respectively.

Results of the discriminant analysis further confirms the superiority of the Mp and Wo–Ws scales in distinguishing honest and "fake–good" responders. These two scales correctly classified 77.5% of cases. Overall, the value of the Mp and Wo–Ws scales over the L and K scales in differentiating honest from "fake–good" responders, in a clinical population, is apparent.

Synopsis of Results from Three Experimental Conditions

The present study gave a clear indication that the "new" validity indicators, the Mp and Wo–Ws scales, are better at detecting "fake–good" response sets than the traditional F, K and F–K indices.

The proposed hypotheses for the "fake–bad" and random conditions were not confirmed. All validity measures were extremely effective when students were instructed to generally "fake–bad."

The present study indicated that the traditional measures, F and F–K, were effective in determining profile invalidity when subjects were instructed to respond randomly, while the TR index and C scale, "new" validity indices, were ineffective. Previous studies regarding the effectiveness of the TR and C indices mainly utilized profiles of high F scale scorers. The literature review did not uncover any simulation studies of random responding. The university students in the present study experienced difficulty in responding carelessly. Perhaps a different

population and/or a different instructional set (to respond in an uncooperative manner) may facilitate simulation. The usefulness of the TR index and C scale seem dependent on the arbitrary setting of cutoff scores, with classification results varying accordingly. This appears to have been the case in this study.

With the exception of the F–K index and the Wo–Ws scales in the "fake good" condition, significantly larger standard deviations from mean scores were found in the experimental conditions than in the honest conditions suggesting a large variability in the subjects ability to fake.

CHAPTER VI

IMPLICATIONS AND CONCLUSIONS

Implications

The problem of detecting "fake good" response sets in the test-taking of clinical subjects, e.g., anxious to be discharged, has been an ongoing concern. The present study provided evidence which suggests that it is worthwhile to routinely use the Mp and Wo—Ws scales as an alternative to the L and K scales for the detection of "faking good" in psychiatric populations. Previous studies (Gendreau et al., 1973; Wiggins, 1959) provided similar evidence in both prison and normal populations.

All validity indicators were extremely effective in detecting "fake bad" response sets. Therefore, it seems pointless to substitute "new" validity indices to detect a "general fake bad" response set in a normal population. This experimental condition, however, was limited by both the population and instructional set used. Replication is needed using a psychiatric population to determine if results can be generalized. Research is also needed using "specific" instructional sets, in both normal and psychiatric populations. It is possible that the "new" validity scales may be better able to detect "faking bad" when specific descriptions and

instructions regarding the role to be simulated are given. Further research into this stereotypic style of responding seems warranted before a final conclusion can be reached regarding the overall usefulness of "traditional" and "new" validity measures in the detection of "fake bad" responding.

In the random responding condition, the F and F–K indices were adequate in detecting profile invalidity, but they were unable to differentiate between "fake bad – plea for help" response sets and random ones. This content dependent – content independent distinction is often useful to clinicians. The failure of the TR index and C scale to detect random responding may be largely related to the population used. The instructional set used and the inappropriateness of the cutoff scores may also be contributing factors in the poor performance of the TR index and C scale. Replication of this study utilizing a different population, instructional set and more appropriate cutoff scores may provide some understanding as to whether the TR index and C scale may be a useful adjunct to the F scale with different cutoff scores in other populations when different instructional sets are used.

Limitations of the Study

In summarizing the limitations of this study, two main factors appear to restrict the ability to generalize results, namely, the samples used and the instructional sets given in the three experimental conditions.

Sample

Results of the study must be confined to the populations used. Results of the "faking bad" and "random" conditions are limited to a normal population, while results of the "fake good" condition are restricted to a psychiatric population.

The use of volunteers suggests the possibility of biased samples, with differences existing in both populations between those who did and did not volunteer. The vast majority of psychiatric patients initially volunteered to participate in the study. However, approximately 25% of volunteers were unable to complete the study due to hospital discharge, or an increase in psychopathology which lessened the confidence that subjects could respond in a meaningful way. Patients who completed the study, therefore, exhibited a moderate degree of psychopathology.

With reference to both the results of mean score differences between conditions and the number of misclassifications under the two conditions, generalizing the results to other similar populations appears warranted. However, with respect to the discriminant function analyses, the sample

size is considered small and replication is suggested before generalizations can be made.

Instructional Set

Results are also limited by the particular instructional sets utilized in the three experimental conditions. Instructions in the "fake bad" condition were general. University students given instructions to fake a specific psychopathological disorder may produce quite different results.

Instructions to "fake good" contained requests to "make a good impression" (social desirability) and deny problems, the two primary elements identified in "fake good" response sets. Therefore, results may be generalized to other psychiatric populations.

The use of a specific role–playing instructional set, to "respond carelessly" appeared inappropriate to detect random responding in a university sample. Students were unable to respond carelessly even with considerable effort. The more "general" instructional set, to respond randomly, may have been more appropriate for this population.

Summary

Three validity measures are included in the MMPI to appraise the genuineness and consistency of responding. A number of problems with these "traditional" validity scales have been identified. The L scale, which

was designed to detect "fake good" response sets, was found to be ineffective with brighter and more highly educated persons as they are willing to admit minor social faults included on the L scale (Good and Brantner, 1974). Elevated K scale scores, originally considered to be a measure of defensiveness, were found to be so in particular maladjusted populations. However, normal well–adjusted and well–educated individuals were also found to obtain elevated K scale scores. The F scale was found to be fairly successful in detecting "fake bad" response sets. Unfortunately, elevated F scores were also found to reflect severe disorganization or psychosis and honest responding in delinquent adolescents. Furthermore, elevated F scores cannot distinguish whether the individual was "faking bad" or answering in an uncooperative or random manner. As a result of these problems inherent in the "traditional" validity measures, "new" validity indices were developed.

This study attempted to compare the effectiveness of the "traditional" and "new" validity measures, when subjects were instructed to "fake good," "fake bad" and respond in a random manner. Results showed the "new" validity indices, the Mp and Wo–Ws scales, were superior to "traditional" measures in the detection of "fake good" response sets in a psychiatric population. All validity indices were extremely effective in detecting a "general fake bad" response set in a university

population. The "new" validity measures, the TR index and C scale were found ineffective in detecting random responding in a university population.

Conclusions based upon these findings of the study indicated that the substitution of the "new" validity indices, Mp and Wo—Ws, for the "traditional" validity indicators would be fruitful in a psychiatric population. The addition or substitution of "new" validity measures to detect a "general fake bad" response set in a normal population is not warranted. Further research is needed to determine if these findings can be generalized to a psychiatric population. Further research is also needed to determine the relative effectiveness of "traditional" and "new" validity measures when a specific, stereotypic "fake bad" response set is utilized. Finally, the experimental condition utilized to detect random responding was ineffective in a university population. Further research in this area using a different instructional set and population may provide insight into the usefulness of the TR index and C scale under different conditions.

REFERENCES

Anastasi, A. (1968). *Psychological testing* (3rd ed.), (p. 28). New York: Macmillan.

Anthony, N. (1971). Comparison of client's standard, exaggerated and matching MMPI profiles. *Journal of Consulting and Clinical Psychology, 36*(1), 100–103.

Archer, R. P., White, J. L., & Orvin, G. H. (1979). *Journal of Clinical Psychology, 35*(3), 498–504.

Berger, E. M. (1955). Relationships among acceptance of self, acceptance of others and MMPI scores. *Journal of Counseling Psychology, 2,* 279–283.

Block, J., & Thomas, H. (1955). Is satisfaction with self a measure of adjustment? *Journal of Abnormal Social Psychology, 51,* 254–259.

Buechley, R., & Ball, H. (1952). A new test of "validity" for the group MMPI. *Journal of Consulting Psychology, 16,* 299–301.

Burish, T. G., & Houston, B. K. (1976). Construct validity of the lie scale as a measure of defensiveness. *Journal of Clinical Psychology, 32,* 310–314.

Burkhart, B. R., Gynther, M. D., & Fromuth, M. E. (1980). The relative predictive validity of subtle vs. obvious items on the MMPI depression scale. *Journal of Clinical Psychology, 36,* 748–751.

Buros, O. K. (Ed.). (1978). *The eighth mental measurements yearbook.* (pp. 938–962). Highland Park, N.J.: Gryphon Press.

Butcher, J. N. (Ed.). (1969). *MMPI: research developments and clinical applications.* New York: McGraw–Hill.

Christian, W. L., Burkhart, B. R., & Gynther, M. D., (1978). Subtle–Obvious ratings of MMPI items: new interest in an old concept. *Journal of Consulting and Clinical Psychology, 46*(6), 1178–1186.

Clopton, J. R., & Neuringer, C. (1977). MMPI cannot say scores: normative data and degree of profile distortion. *Journal of Personality Assessment, 41*(5), 511–513.

Coche', E., & Steer, R. A. (1974). The MMPI response consistencies of normal, neurotic and psychotic women. *Journal of Clinical Psychology, 30,* 194–195.

Cofer, C. N., Chance, J., & Judson, A. J. (1949). A study of malingering on the the MMPI. *Journal of Psychology, 27,* 491–499.

Comrey, A. L. (1958). A factor analysis of items on the F scale of the MMPI. *Educational and Psychological Measurement, 18,* 621–632.

Costa, L. D., London, P., & Levita, E. (1963). A modification of the F scale of the MMPI. *Psychological Reports, 12,* 427–433.

Coyle, F. A. Jr., & Heap, R. F. (1965). Interpreting the MMPI L scale. *Psychological Reports, 17,* 732.

Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology, 24,* 349–354.

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

Dahlstrom, W. G., Welsh, G. S., & Dahlstrom, L. E. (1975). *An MMPI handbook. Vol. II. Research applications.* (Rev. ed.). Minneapolis: Universtiy of Minnesota Press.

Evans, R. G., & Dinning, W. D. (1983). Response consistency among high F scale scorers on the MMPI. *Journal of Clinical Psychology, 39,* 246–248.

Exner, J. E. Jr., McDowell, E., Pabst, J., Stackman, W., & Kirk, L. (1963). On the detection of willful falsification in the MMPI. *Journal of Consulting Psychology, 27,* 91–94.

Gallagher, J. J. (1953). MMPI changes concomitant with client centered therapy. *Journal of Consulting Psychology, 17,* 334–338.

Gallucci, N. T. (1984). Prediction of dissimulation on the MMPI in a clinical field setting. *Journal of Consulting and Clinical Psychology, 52,* 917–918.

Gendreau, P., Irvine, M., & Knight, S. (1973). Evaluating response set styles on the MMPI with prisoners: faking good adjustment and maladjustment. *Canadian Journal of Behavioral Sciences, 5,* 183–194.

Gilberstadt, H. (1970). *Comprehensive MMPI code book for males.* Minneapolis: MMPI Research Laboratory, Veterans Administration Hospital.

Good, P. K., & Brantner, J. P. (1974). *A practical quide to the MMPI.* Minneapolis: University of Minneapolis Press.

Gough, H. G. (1947). Simulated patterns on the MMPI. *Journal of Abnormal & Social Psychology, 42,* 215–225.

Gough, H. G. (1950). The F minus K dissimulation index for the Minnesota Multiphasic Personality Inventory. *Journal of Consulting Psychology, 14,* 408–413.

Gough, H. G. (1954). Some common misconceptions about neuroticism. *Journal of Consulting Psychology, 18,* 287–292.

Gough, H. G. (1957). *California Psychological Inventory Manual.* Palo Alto, California: Consulting Psychologists Press.

Graham, J. R. (1977). *The MMPI: A practical guide.* New York: Oxford University Press.

Gravitz, M. A. (1970). Validity implications of normal adult MMPI "L" scale endorsement. *Journal of Clinical Psychology, 26,* 497–499.

Grayson, H. M., & Olinger, L.B. (1957). Simulation of "normalcy" by psychiatric patients on the MMPI. *Journal of Consulting Psychology, 21,* 73–77.

Greene, R. L. (1978). An empirically derived MMPI Carelessness Scale. *Journal of Clinical Psychology, 34,* 407–410.

Greene, R. L. (1979). Response consistency on the MMPI: the TR index. *Journal of Personality Assessment, 43*(1), 69–71.

Greene, R. L. (1983). *Workshop: MMPI interpretation.* American Psychological Association. Anaheim, California.

Grow, R., McVaugh, W., & Eno, T. D. (1980). Faking and the MMPI. *Journal of Clinical Psychology, 36*(4), 910–917.

Gynther, M.D. (1961). The clinical utility of "invalid" MMPI F scores. *Journal of Consulting Psychology, 25*(6), 540–542.

Gynther, M. D., Lachar, D., & Dahlstrom, W. G. (1978). Are special norms for minorities needed? Development of an MMPI F scale for blacks. *Journal of Consulting and Clinical Psychology, 46,* 1403–1408.

Gynther, M. D., & Petzel, T. P. (1967). Differential endorsement of MMPI F scale items by psychotics and behavior disorders. *Journal of Clinical Psychology, 23,* 185–188.

Gynther, M. D., & Shimkunas, A. M. (1965). Age, intelligence, and MMPI F scores. *Journal of Consulting Psychology, 29,* 383–388.

Haertzen, C. A., & Hill, H. E. (1963). Assessing subjective effects of drugs: An index of carelessness and confusion for use with the Addiction Research Center Inventory (ARCI). *Journal of Clinical Psychology, 19,* 407–412.

Harris, J. G. Jr., & Baxter, J. C. (1965). Ambiguity in the MMPI. *Journal of Consulting Psychology, 29,* 112–118.

Harrison, R. H., & Kass, E. H. (1968). MMPI correlates of Negro acculturation in a northern city. *Journal of Personality and Social Psychology, 10,* 262–270.

Hartshorne, H., & May, M. A. (1928). *Studies in the nature of character: I. Studies in deceit.* New York: Macmillan.

Harvey, M. A., & Sipprelle, C. N. (1976). Demand characteristic effects on the subtle and obvious subscales of the MMPI. *Journal of Personality Assessment, 40,* 539–544.

Hathaway, S. R. (1947). A coding system for MMPI profiles. *Journal of Consulting Psychology, 11,* 334–337.

Hathaway, S. R., & McKinley, J. C. (1940). A multiphasic personality schedule. Minnesota: 1. Construction of the schedule. *Journal of Psychology, 10,* 249–254.

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

Hathaway, S. R., & McKinley, J. C. (1951). *The MMPI manual.* New York: The Psychological Corporation.

Hathaway, S. R., & McKinley, J. C. (1967). *MMPI manual* (Rev. ed.). New York: Psychological Corporation.

Heilbrun, A. B. (1961). The psychological significance of the MMPI K scale in a normal population. *Journal of consulting Psychology, 25,* 486–491.

Heilbrun, A. B. (1963). Revision of the MMPI K correction procedure for improved detection of maladjustment in a normal college population. *Journal of Consulting Psychology, 27,* 161–165.

Hovanitz, C. A., & Gynther, M. D. (1980). The prediction of impulsive behavior: comparative validities of obvious vs. subtle MMPI hypomania (MA) items. *Journal of Clinical Psychology, 36,* 422–427.

Hovanitz, C. A., Gynther, M. D., & Green, S. B. (1985). Discriminant validity of subtle and obvious items: the MMPI Pa and Ma scales. *Journal of Clinical Psychology, 41,* 42–44.

Hovanitz, C. A., Gynther, M. D., & Marks, P. A. (1983). The prediction of paranoid behavior: comparative validities of obvious vs. subtle MMPI paranoid (Pa) items. *Journal of Clinical Psychology, 39,* 407–411.

Hovanitz, C. A., & Jordan–Brown, L. (1986). The validity of MMPI subtle and obvious items in psychiatric patients. *Journal of Clinical Psychology, 42,* 100–108.

Hunt, H. F. (1948). The effect of deliberate deception on MMPI performance. *Journal of Consulting Psychology, 12,* 396–402.

Hunt, H. F., Carp, A., Cass, W. A., Winder, C. L., & Kantor, R. (1948). A study of the differential diagnostic efficiency of the MMPI. *Journal of Consulting Psychology, 12,* 331–336.

Jones, F. W., Neuringer, C., & Patterson, T. W. (1976). An evaluation of an MMPI response consistency measure. *Journal of Personality Assessment, 40*(4), 419–421.

Kazan, A. T., & Sheinberg, I. M. (1945). Clinical note on the significance of the validity score (F) in the MMPI. *American Journal of Psychiatry, 102,* 181–183.

Landis, C., & Katz, S. E. (1934). The validity of certain questions which purport to measure neurotic tendencies. *Journal of Applied Psychology, 18,* 343–356.

Lanyon, R. I., & Lutz, R. W. (1984). MMPI discrimination of defensive and nondefensive felony sex offenders. *Journal of Consulting and Clinical Psychology, 52,* 841–843.

Leary, T. (1957). *Interpersonal diagnosis of personality: A functional theory and methodology for personality evaluation.* New York: Ronald.

Little, K. B., & Fisher, J. (1958). Two new experimental scales of the the MMPI. *Journal of Consulting Psychology, 22,* 305–306.

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

McDonald, R. L., & Gynther, M. D. (1963). MMPI differences associated with sex, race and class in two adolescent samples. *Journal of Consulting Psychology, 27,* 112–116.

McKegney, F. P. (1965). An item analysis of the MMPI F scale in juvenile delinquents. *Journal of Clinical Psychology, 21,* 201–205.

McKinley, J. C., Hathaway, S. R., & Meehl, P. E. (1948). The MMPI: VI. The K scale. *Journal of Consulting Psychology, 12,* 20–31.

Meehl, P. E., & Hathaway, S. R. (1946). The K factor as a suppressor variable in the MMPI. *Journal of Applied Psychology, 30,* 525–564.

Montgomery, G. T., & Orozco, S. (1985). Mexican American's performance on the MMPI as a function of level of acculturation. *Journal of Clinical Psychology, 41,* 203–212.

Osborne, D., Colligan, R. C., & Offord, K. P. (1986). Normative tables for the F–K index of the MMPI based on a contemporary, normal sample. *Journal of Clinical Psychology, 42,* 593–595.

Page, J., Landis, C., & Katz, S. E. (1934). Schizophrenic traits in the functional psychoses and in normal individuals. *American Journal of Psychiatry, 90,* 1213–1225.

Plemons, G. (1977). A comparison of MMPI scores of Anglo– and Mexican–American psychiatirc patients. *Journal of Consulting and Clinical Psychology, 45,* 149–150.

Post, R. D., & Gasparikova–Krasnec, M. (1979). MMPI validity scales and behavioral disturbance in psychiatric patients. *Journal of Personality Assessment, 43*(2), 155–159.

Ries, H. A. (1966). The MMPI K scale as a predictor of prognosis. *Journal of Clinical Psychology, 22,* 212–213.

Rogers, R., Dolmetsch, R., & Cavanaugh, J. L. (1983). Identification of random responders on MMPI protocols. *Journal of Personality Assessment, 47*(4), 365–368.

Rosen, A. (1952). Reliability of MMPI scales. *American Psychologist, 7,* 341.

Rothaizer, J. M. (1980). A typological study of substance abusers using the MMPI. *Journal of Clinical Psychology, 36*(4), 1019–1021.

Schenkenberg, T., Gottfredson, D. K., & Christensen, P. (1984). Age differences in MMPI scale scores from 1,189 psychiatric patients. *Journal of Clinical Psychology, 40,* 1420–1426.

Schmidt, H. O. (1948). Notes on MMPI: The K factor. *Journal of Consulting Psychology, 12,* 337–342.

Scneck, J. M. (1948). Clinical evaluation of the F scale on the MMPI. *American Journal of Psychiatry, 104,* 440–442.

Schofield, W. (1953). A further study of the effects of therapies on MMPI responses. *Journal of Abnormal and Social Psychology, 48,* 67–77.

Silver, R. J., & Sines, L. K. (1962). Diagnostic efficiency of the MMPI with and without the K correction. *Journal of Clinical Psychology, 18,* 312–314.

Sines, L. K., Baucom, D. H., & Gruba, G. H. (1979). A validity scale sign calling for caution in the interpretation of MMPIs among psychiatric inpatients. *Journal of Personality Assessment, 43,* 604–607.

Smith, E. E. (1959). Defensiveness, insight and the K scale. *Journal of Consulting Psychology, 23,* 275–277.

Sue, S., & Sue, D. W. (1974). MMPI comparisons between Asian–American and non–Asian students utilizing a student health psychiatric clinic. *Journal of Counseling Psychology, 21,* 423–427.

Sweetland, A., & Quay, H. (1953). A note on the K scale of the MMPI. *Journal of Consulting Psychology, 17,* 314–316.

Tyler, F. T., & Michaelis, J. U. (1953). K scores applied to MMPI scales for college women. *Educational and Psychological Measurement, 13,* 459–466.

Vincent, N. M. P., Linsz, N. L., & Greene, M. I. (1966). The L scale of the MMPI as an index of falsification. *Journal of Clinical Psychology, 22,* 214–215.

Wales, B., & Seeman, W. (1968). A new method for detecting the fake good response set on the MMPI. *Journal of Clinical Psychology, 24,* 211–216.

Welsh, G. S. (1948). An extension of Hathaway's MMPI profile coding system. *Journal of Consulting Psychology, 12,* 343–344.

Wheeler, W. M., Little, K. B., & Lehner, G. F. J. (1951). The internal structure of the MMPI. *Journal of Consulting Psychology, 15,* 134–141.

Wiener, D. N. (1948). Subtle and obvious keys for the MMPI. *Journal of Consulting Psychology, 12,* 164–170.

Wiggins, J. S. (1959). Interrelationship among MMPI measures of dissimulation under standard and social desirability instructions. *Journal of Consulting Psychology, 23,* 419–427.

Wiggins, J. S., & Rumrill, C. (1959). Social desirability in the MMPI and Welsh's factor scales A and R. *Journal of Consulting Psychology, 23,* 100–106.

Wilcox, P., & Dawson, J. G. (1977). Role–played and hypnotically induced simulation of psychopathology on the MMPI. *Journal of Clinical Psychology, 33,* 743–745.

Yonge, G. D. (1966). Certain consequences of applying the K factor to MMPI scores. *Educational & Psychological Measurement, 26,* 887–893.

APPENDIX A

CONSENT FORM FOR PATIENTS

CONSENT FORM

I hereby consent to participate in a research project designed to study the validity of a personality test, the Minnesota Multiphasic Personality Inventory (MMPI). This will involve my completing the MMPI for research purposes, a task which normally takes about an hour to an hour and a half per occasion. The project is conducted by Mrs. Carole Woychyshyn in partial fulfillment of the M.Sc. degree in the Department of Educational Psychology, University of Calgary and is being supervised by Dr. William G. McElheran, Director, Department of Psychology, Holy Cross Hospital.

All information provided will be held strictly confidential; the questionnaires will be labelled with an anonymous code number. Only the researchers and my doctor will be allowed to know my individual results. Upon completion of the project, the questionnaires will be destroyed.

I understand that my participation in this project is purely voluntary and that I may withdraw from the project at any time. Should I decline from participating in the project, my opportunity to receive assistance at the Mental Health Centre will in no way be jeopardized. I also understand that I have the right to request a summary of the results of this project.

Signed: _____

Dated: _____

Witness: _____

CONSENT FORM FOR UNIVERSITY STUDENTS:

CONSENT FORM

I, _____, hereby consent to participate in the research study designed to investigate the validity of a personality test, the Minnesota Multiphasic Personality Inventory (MMPI). This will involve my completing the MMPI for research purposes, a task which normally takes about an hour to an hour and a half per occasion. The project is conducted by Mrs. Carole Woychyshyn in partial fulfillment of the M.Sc. degree in the Department of Educational Psychology, University of Calgary.

All aspects of the study which might influence my decision to participate have been fully explained to me, including the purpose and method of the research, the nature of my involvement and the absence of any personal risks. I understand that I may withdraw my participation, without penalty, at any time and also realize the researcher's right to terminate my involvement at any time.

I understand that all information provided by me in the form of psychological tests will be held strictly confidential; the questionnaire will be labelled with an anonymous code number. Only the researcher will know my individual results. Upon completion of the project, the questionnaires will be destroyed.

Signed: _____

Dated: _____

Witness: _____