

Forensic Use and Abuse of Psychological Tests: Multiscale Inventories

RICHARD ROGERS, PhD, ABPP

Properly used, many psychological tests are very useful diagnostic and clinical tools. They often clarify ambiguous information, elicit previously unavailable data, and add objectivity, validity, and reliability to patient interactions and record reviews. Unfortunately, some instruments do not live up to these objectives, especially in forensic cases and/or in the hands of those without specialized psychometric training. This month's guest columnist, one of America's most prominent forensic psychologists, discusses a widely used type of test, multiscale inventories, and focuses on three of the most commonly employed measures for impaired populations.

William H. Reid, MD, MPH

First, what is a multiscale inventory? The term refers to a psychological test, or "inventory," that assesses different aspects of a person's functioning or personality based on his or her answers to a standard series of questions presented in a paper-and-pencil format. Patterns of psychopathology are reported as raw or standardized scores on individual scales. Answers to individual questions are generally much less important than these scale scores or elevations, or the patterns among them. On many inventories, specific scales are designed to detect response styles and other test-taking behaviors that decrease the validity of the results. Response styles include malingering (e.g., markedly exaggerated impairment), defensiveness (e.g., grossly minimized impairment), and irrelevant responses (e.g., haphazard or inconsistent answering of questions).

For the purposes of this column, I will focus only on multiscale inventories that assess patterns of adult psychopathology. This discussion will be limited to the three most popular measures: the Minnesota Multiphasic Personality Inventory-2 (MMPI-2),¹ the Personality Assessment Inventory (PAI),² and the latest version of the Millon Clinical Multiaxial Inventory (MCMI-III).³

Multiscale inventories have sometimes been referred to as "objective tests," an expression still used periodically. The term "objective tests" is a misnomer that may create a false impression among mental health professionals. Although the *scoring* is objective, the *interpretation* is not. In practice, clinicians generally select specific interpretations from a panoply of already published possible interpretations, such as those found in an MMPI-2 handbook or a computerized report. This issue of selective interpretations will be discussed in more detail below with reference to "cherry-picking."

Psychiatrists' Options with Multiscale Inventories

Psychiatrists have three basic options in how they use—or possibly misuse—multiscale inventories: 1) dismiss and discount, 2) use directly, and 3) use through consultation.

1. *Dismiss and discount.* The first option is to dismiss multiscale inventories as unnecessary or superfluous to psychiatric evaluations. Although this alternative may sometimes be tempting, multiscale inventories should not be discounted entirely. These measures are commonly used by the majority of psychiatrists, psychologists, and other mental health professionals, especially in forensic consultations. Multiscale inventories are often an integral component of diagnostic evaluations.
2. *Use directly.* The second option is to purchase testing services directly from test firms. This option appears to be favored (implicitly at least) by the American Psychiatric Association in the *APPI*

Guest columnist RICHARD ROGERS, PhD, ABPP, is professor of psychology at the University of North Texas and maintains a part-time consultation practice specializing in clinical-forensic assessments and malingering. His books include *Clinical Assessment of Malingering and Deception* (winner of the American Psychiatric Association's Guttmacher Award), *Conducting Insanity Evaluations* (with Daniel W. Shuman, JD), and *Diagnostic and Structured Interviewing* (including diagnostic measures usable by psychiatrists), all published by Guilford Press.

Handbook of Psychiatric Measures.⁴ That text details practical issues, such as costs and contact information, for multiscale inventories. The basic limitation of this option lies in the necessarily generic nature of computerized reports; their encyclopedic approach to interpretation is overly inclusive, combining empirically tested findings with weak correlates and theoretically driven suppositions about the individual being tested. How does one separate the comparatively sparse wheat from the overly abundant chaff?

3. *Use through consultation.* The third option, consistent with discipline-based practice, is to use an expert consultant for multiscale inventory interpretations. The psychiatrist would rely on another professional, typically a clinical psychologist, for the interpretation of profiles and concomitant response styles. A forensic psychologist may be recommended when legal matters are involved. The main limitation of this option is in the selection of a consultant; a substantial minority of doctoral-level psychologists have only a basic knowledge of specific test interpretation. The consulting psychologist should be very familiar with current studies on the applicability, validity, and reliability of the inventories used and ideally should have advanced training in their design and use.

Each alternative appears less than ideal by itself. I recommend that psychiatrists engage in some self-training. For example, Greene's *The MMPI-2: An Interpretive Manual*⁵ is a good introduction to that test. As the next step, the psychiatrist should choose a testing consultant. The consultant should be selected with some care, since less than rigorous test interpretations threaten clinical and forensic evaluations and undermine the psychiatrist's credibility. Look for a consultant with a knowledge of the research (both prevailing and countervailing) on the tests being considered.

Common Pitfalls in Test Interpretation

Reading comprehension. The essential prerequisite for multiscale inventories is that patients have adequate reading comprehension. Many patients are embarrassed to admit their limited literacy and may guess at the meaning of difficult words. Psychiatrists and other mental health professionals should not rely on years of completed education as a guide for reading grade level. Forensic research suggests an average disparity of approximately 4 years between school attain-

ment and reading comprehension. Clinicians should ask patients to read the first 10 or so items aloud as a minimal screen for reading comprehension.

Cherry-picking interpretations. How does the clinician select the "correct" interpretation of a test when faced with dozens—sometimes hundreds—of possibilities? Clinicians with sophisticated knowledge of the test validation may be able to select those interpretations that are best validated. A common error for most clinicians, however, is selecting the interpretation that "best fits" the individual patient. This process of "taking the best and leaving behind the rest" is known as *cherry-picking*. It is a dangerous enterprise that involves self-selecting confirming interpretations and discarding interpretations that do not fit one's own clinical view. *Cherry-picking constitutes an extreme form of confirmatory bias and should not be used in clinical or forensic practice.* Psychiatrists should routinely ask their consultants, "Given the number of possible interpretations, how did you arrive at these conclusions?" and "Were there other interpretations that did not fit the patient as well?"

As a practical solution, I suggest that psychiatrists shape their referral questions to minimize cherry-picking. Consider using wording such as the following when requesting psychological testing.

I am treating Mr. Jones and have formulated preliminary ideas regarding his personality, diagnosis, and needed treatment. Your psychological consultation would be most helpful to me if it provided different, even competing, interpretations of Mr. Jones's personality functioning and diagnostic issues.

Translated versions. Multiscale inventories can be translated into different languages with relatively little effort. The critical issue is that *linguistic* equivalence (i.e., similar sentences) cannot be equated to *clinical* equivalence (i.e., similar diagnostic relevance). Clinical equivalence cannot be assumed, but must be objectively tested. Simple comparisons of vocabulary and syntax (e.g., from English to Spanish) are insufficient to establish clinical validity for translated versions.

An approach to translation validation in which a mere lack of significant group differences between two language or ethnic groups is assumed to mean the tests "work the same way" makes little sense. With depressed patients, for example, the clinician needs to

know whether or not depressed persons of different cultures and languages have the appropriate elevations on multiscale inventories. Given our dearth of knowledge regarding translated versions and their cultural differences, psychiatrists and their consultants should be very cautious about using and interpreting translated tests.

Use and Misuse of the MMPI-2

The Minnesota Multiphasic Personality Inventory, a test first introduced in the early 1940s, went through a major revision which created the MMPI-2,¹ with new and revised items and modified scales. Its 567 true-false questions are organized into overlapping scales that create one of the most complex interpretation challenges in psychological measurement.

The MMPI-2 is composed of the following:

- Ten clinical scales addressing patterns of psychopathology. Unfortunately, the names of these scales are misnomers. For example, elevations on the Schizophrenia [Sc] scale are found in a range of disorders and cannot be used as direct evidence of schizophrenic disorders. These misnomers are likely to mislead professionals unfamiliar with the MMPI-2 and its interpretation process.
- Fifteen content scales describing common clinical issues that the patient has endorsed.
- Validity scales (originally three, but now a dozen or more scales and indices) for evaluating whether or not patients are a) malingering (e.g., markedly over-reporting), b) defensive (e.g., markedly under-reporting), or c) grossly inconsistent (e.g., random) in their presentations.
- Hundreds of specialized research scales that often have either focused or limited clinical applications. Some scales have been extensively validated in peer-reviewed research; many others have not.

The MMPI-2 is a well-validated multiscale inventory that can be used with a wide range of inpatients and outpatients. Patients should be screened for adequate education (at least 8th grade reading level) and sufficient concentration to complete this lengthy inventory. The MMPI-2 is especially well-suited for cases in which psychiatrists have concerns about a patient's response styles. Rogers et al.⁶ completed a comprehensive meta-analysis of 65 feigning studies that compared malingered and genuine MMPI-2 protocols. They found that specific validity scales (especially Scale Fp) can assist in determining cases of

possible malingering. In defensive patients, the MMPI-2 has the most sophisticated methods for detecting minimization of psychological impairment of any current test (e.g., Wiggins Social Desirability). Beyond response styles, the MMPI-2 offers clinical descriptions based on code-types (i.e., combination of highest elevations) and individual scale elevations. Carefully interpreted, MMPI-2 profiles provide useful data about patterns of psychopathology.

Practitioners should note the common misinterpretation of "within normal limits" ("WNL") profiles. When none of the clinical scales is elevated (i.e., all scale scores are below a T score of 65), many mental health professionals erroneously equate the lack of clinical elevations with an absence of psychopathology or impairment. *The "WNL" profile is often found in chronic populations and cannot be interpreted as a "healthy" profile.* The "WNL" profile is the most common patient profile, occurring in approximately 30% of referrals.

Psychiatrists are urged to use a psychological consultant to address the myriad interpretations possible for the MMPI-2. As noted previously, computerized reports can be less than helpful to mental health professionals who are not versed in the test itself. The daunting task of separating the wheat from the chaff should only be undertaken by those with specialized MMPI-2 training.

Use and Misuse of the PAI

The Personality Assessment Inventory (PAI)² is a new-generation multiscale inventory which challenges the established primacy of the MMPI-2. The PAI has four major advantages over the MMPI-2.

- Patients are allowed four gradations of response ("false," "slightly true," "mainly true," and "very true") rather than being forced to respond only "true" or "false."
- The PAI items are more easily comprehensible (4th grade reading level), making them practicable for use with many clinical populations.
- The PAI scales and subscales are often more directly interpretable than those of the MMPI-2 because a) items on the PAI scales have excellent internal consistency and b) the PAI scales are non-overlapping (i.e., each item is used for only one scale).
- The PAI clinical scales are more directly relevant to DSM-IV than are the MMPI-2 clinical scales. For example, the PAI Anxiety subscales allow clinicians to evaluate cognitive, affective, and physiological components of anxiety disorders. In contrast, anxiety

impairment is indirectly addressed by the MMPI-2 (via the Psychasthenia Scale).

In addition to these advantages, the PAI (344 items) is much shorter than the MMPI-2, making it more convenient for use in patient settings. The PAI scale descriptions are aligned with the DSM-IV conceptualization of mental disorders. It assists in the evaluation of issues related to treatment success, including potential for treatment rejection, lack of social support, and perceived stressors. Finally, the PAI examines several facets of clinical management, such as suicidal ideation and aggression.

Despite its psychometric superiority over the MMPI-2, the PAI should not be considered a *diagnostic* measure. The PAI assesses useful patterns of psychopathology that are related to DSM-IV diagnoses. However, the PAI does not formally evaluate the DSM-IV inclusion and exclusion criteria. Therefore, its results may only augment DSM-IV diagnoses from structured and clinical interviews (see Rogers⁷).

The PAI is strongly recommended as a first-rate multiscale inventory that will likely address many clinical issues and concerns germane to psychiatrists and highly relevant to clinical conditions. Its interpretations (see Morey⁸) are much less complicated than the MMPI-2, which should lead to greater uniformity in PAI interpretations.

Use and Misuse of the MCMI-III

The Millon Clinical Multiaxial Inventory-III (MCMI-III)³ is an ambitious attempt to evaluate both Axis I clinical syndromes and Axis II personality disorders. The MCMI-III strives to accomplish these goals and evaluate response styles with only 175 true-false items. Psychiatrists should note that the MCMI-III is radically different from its predecessor, the MCMI-II. The majority of the items were replaced, and the scoring and scale composition were dramatically changed. Thus it is important to understand that *earlier versions of the MCMI cannot be used to validate the MCMI-III*.

The MCMI-III is marketed to psychiatrists and other mental health professionals by promoting “MCMI-III generated diagnoses” that are consonant with DSM-IV. Given the MCMI-III’s singular contribution to the diagnosis of personality disorders, this discussion will focus on DSM-IV Axis II disorders.

A meta-analysis of the MCMI-III and Axis II disorders by Rogers et al.⁹ found the MCMI-III was insuffi-

ciently validated. Not surprisingly, several authors associated with the MCMI-III have questioned the findings of this meta-analysis.¹⁰ Here are the two key conclusions of that meta-analysis:

1. The MCMI-III scales lack sufficient “construct validity” to be used in forensic settings. For example, the MCMI-III scale for Schizotypal Personality Disorder showed a negligible relationship ($r = 0.16$) to this diagnosis and appeared more related ($r = 0.38$) to other disorders.
2. MCMI-III scales cannot be used to diagnose DSM-IV personality disorders; the test may generate errors in about 80% of diagnosed cases.

The meta-analysis overlooked an additional study which was cited in the slightly revised 1997 MCMI-III test manual.¹¹ However, the 1997 study and its more positive results violated a fundamental principle of test validation because of criterion contamination. Some clinicians rendering diagnoses in the study were previously exposed to the MCMI-III results, thereby contaminating the results. Rogers et al. also noted other methodological concerns with regard to the 1997 study.¹²

Many mental health professionals continue to use the MCMI-III because of its brevity and ease of administration; it provides interpretative data with a minimum of professional time. Unfortunately, close examination suggests that the MCMI-III may promise more than it delivers. The test manual is helpful in this regard, advising that conclusions from the MCMI-III should be viewed as “a series of tentative and probabilistic judgments.” Given the admitted tentativeness of its conclusions, coupled with methodological limitations, its role in forensic cases is, at best, very circumscribed.

Forensic Applications

Multiscale inventories are something like utility players in baseball. They perform well for a variety of purposes but do not excel at any highly specialized task. Psychiatrists are therefore likely to value multiscale inventories for the general clinical information they provide; however, clinicians should not try to make direct linkages between test interpretations and specific diagnoses or legal capacities. For example, each of the three multiscale inventories discussed here has individual scales designed to discover and assess antisocial features. Should elevations on these scales—even marked ones—be seen as evidence of antisocial personality disorder?

The answer is *definitely not*. Although they measure antisocial characteristics, these scales are not effective in establishing a DSM-IV diagnosis of antisocial personality disorder. Moreover, these scales on the different tests are not highly correlated with each other, which indicates that they are measuring different facets of antisocial, asocial, and delinquent characteristics. Like the metaphor of the utility player, multiscale inventories can provide general information about antisocial characteristics but are ineffective at furnishing detailed data regarding diagnosis or risk management.

Forensic professionals should carefully consider whether or not the multiscale inventories will pass muster under the *Daubert* standard and related case law that limits expert testimony to scientifically established data. The U.S. Supreme Court's landmark decision in *Daubert v. Merrell Dow Pharmaceuticals, Inc.*¹³ established general parameters for the admissibility of scientific and expert testimony. Conclusions from tests and other assessment methods must be empirically testable with a known or knowable error rate. As gatekeepers, trial courts have been discriminating in their decisions about when multiscale inventories are admissible. For example, the Supreme Court of New Hampshire refused to allow MCMI and MMPI-2 results to be used for profiling sex abusers.¹⁴

The Last Word

I greatly value multiscale inventories for what they are able to accomplish, but strongly recommend that practitioners clearly understand their limitations. By discussing some of those limitations, psychiatrists and other professionals can become more sophisticated users of these measures. To accept all of the marketing claims supporting multiscale inventories would be a

serious error; to dismiss these inventories categorically would be equally misguided.

References

1. Butcher JN, Dahlstrom WG, Graham JR, et al. Minnesota Multiphasic Personality-2 (MMPI-2): Manual for administration and scoring. Minneapolis, MN: University of Minnesota Press; 1989.
2. Morey LC. Personality Assessment Inventory: Professional manual. Tampa, FL: Psychological Assessment Resources; 1991.
3. Millon T. The Millon Clinical Multiaxial Inventory-III manual. Minneapolis, MN: National Computer Systems; 1994.
4. Rush AJ, Pincus HJ, First MB, et al, eds. Handbook of psychiatric measures. Washington, DC: American Psychiatric Press; 2000.
5. Greene RL. The MMPI-2: An interpretive manual. Boston: Allyn and Bacon; 2000.
6. Rogers R, Sewell KW, Martin MA, et al. Detection of feigned mental disorders: A meta-analysis of the MMPI-2 and malingering. *Assessment* 2003;10:160-77.
7. Rogers R. Handbook of diagnostic and structured interviewing. New York: Guilford; 2000.
8. Morey LC. An interpretive guide to the Personality Assessment Inventory (PAI). Tampa, FL: Psychological Assessment Resources; 1996.
9. Rogers R, Salekin RT, Sewell KW. Validation of the Millon Multiaxial Inventory for Axis II disorders: Does it meet the Daubert standard? *Law Hum Behav* 1999;23:425-43.
10. Dyer FJ, McCann JT. Millon clinical inventories: Research critical to their forensic application and Daubert criteria. *Law Hum Behav* 2000;24:487-98.
11. Millon T, Davis R, Millon C. The Millon Clinical Multiaxial Inventory-III Manual (2nd ed.). Minneapolis, MN: National Computer Systems; 1997.
12. Rogers R, Salekin RT, Sewell KW. The Millon Clinical Multiaxial Inventory: Separating rhetoric from reality. *Law Hum Behav* 2000;24:501-6.
13. *Daubert v. Merrell Dow Pharmaceuticals*, 113 S.Ct. 2786 (1993).
14. *State v. Cavaliere*, 663 A.2d 96 (N.H. 1995).