MILLON®
INVENTORIES

The Millon® Clinical Inventories Research Critical of Their Forensic Application, and Daubert Criteria

Frank J. Dyer—Private Practice, Montclair, New Jersey
Joseph T. McCann—United Health Services Hospitals, Binghamton, New York

ALWAYS LEARNING
PEARSON

The Millon® Clinical Inventories Research Critical of Their Forensic Application, and Daubert Criteria

Frank J. Dyer—Private Practice, Montclair, New Jersey
Joseph T. McCann—United Health Services Hospitals, Binghamton, New York

The Rogers et al. Critique

In their recent article criticising forensic use of the Millon Inventories under the criteria for scientific evidence articulated in *Daubert v. Merrell Dow Pharmaceuticals (1993)*, Rogers et al. (1999) assert that the Millon Clinical Multiaxial Inventory-II (MCMI-II®) is fit only for circumscribed use and that the MCMI-III® should not be used at all in such settings because of poor convergent and discriminant validity. Rogers et al. list a number of additional complaints, including the fact that neither the MCMI-II nor MCMI-III has been validated against specific legal criteria such as legal insanity or competence to proceed to trial, that the MCMI-II was validated against *DSM-III-R*™ criteria rather than *DSM-IV*® criteria for personality disorders, and that the MCMI-III test manual did not sufficiently describe procedures used in determining content validity against *DSM-IV*.

In this paper we respond to the issues raised by Rogers et al. (1999) in their analysis of forensic applications and admissibility of the Millon Clinical Multiaxial Inventory under Daubert criteria. More specifically, we detail several methodological shortcomings in their study and detail misleading conclusions that they render. Additionally, we address criticisms related to the MCMI-III as a measure of *DSM-IV* disorders, including concerns raised about the content validity of the MCMI-III, as well as the contention that the MCMI® "cannot be employed to address elements of legal standards" (Rogers et al., 1999, p. 439). Finally, we point out that the research compiled by Rogers et al. (1999) on the MCMI-III is incomplete, as they failed to cite the most recent edition of the MCMI-III manual (Millon, Davis, & Millon, 1997), which includes an expanded validation study by Davis, Wenger, & Guzman (1997) that was published and available two years prior to the publication of the Rogers et al. (1999) paper.

Arbitrary Assignment of MCMI as Predictor and Other Instruments as Criterion

Rogers et al. employ a raft of other personality disorder measures as criteria against which to validate the MCMI without comment as to the appropriateness of such criteria. This is an important consideration, as the criterion for establishing the validity of a test is the measure that is considered the benchmark against which the test is gauged. The designation by Rogers et al. of these other measures as the criteria reflects the presumption that if the MCMI fails to demonstrate convergent and discriminant validity against them, then it is the MCMI that lacks validity and that the other measures are all adequate benchmarks by which to assess the validity of the MCMI. Wise (1994) notes that in the development of MMPI® personality disorder scales, the authors and other researchers chose the MCMI as their criterion, as it was the only established measure of personality disorders at the time. Similarly, Renneberg et al. (1992) in their study of the SCID-II explicitly state that the MCMI is an appropriate benchmark against which to evaluate that new measure.

McCann (1991), in commenting on his convergent-discriminant study of the MCMI-II and Morey's MMPI personality disorder scales, concludes that his findings demonstrate adequate convergent and discriminant validity for both instruments, with the exception of the MCMI-II and MMPI Compulsive Scales, which assess that disorder according to the differing conceptualizations of their respective models. McCann's perspective was that of a simultaneous assessment of the validity of the Morey MMPI personality disorder scales and those of the MCMI-II rather than designating either one as the criterion variable. Moreover, one of the major purposes of the studies by McCann (1990, 1991) was to examine the effect of item overlap (one of the often cited weaknesses of the MCMI) on convergent and discriminant validity of the MCMI-II. Results from these studies indicated that the impact of excessive item overlap on validity of the MCMI-II scales is negligible and that factor analyses supported the construct validity of the instrument.

Some research, such as that of Renneberg et al. (1992) demonstrates that the MCMI-II was superior to other instruments such as the SCID-II. Renneberg et al. also point out that there is little convergence across various instruments of assessment of personality disorders. Hunt & Andrews (1992) found poor agreement between the PDQ-R and the PDE, both of which are used in the aggregated correlations in the Rogers et al. meta-analysis as though they were equivalent and interchangeable personality disorder measures. We find it curious that, given the fact that many of the studies used by Rogers et al. validated the newer measures against the MCMI (in other words, if the other test correlated well with the corresponding MCMI scale then the other test was considered valid and if it correlated poorly, then the other test was considered invalid), Rogers et al. see fit to make these newer measures the criterion without even acknowledging that they were developed in part through the use of the MCMI as a benchmark. In our view, the logic of the Rogers et al. study rests

on an arbitrary reversal of predictor and criterion. A more reasonable view of the Rogers et al. findings, if we can ascribe any validity to them at all, is that studies that fail to demonstrate convergent and/or discriminant validity against the MCMI illustrate a failure of the newer, atheoretical measures to show validity against a proven measure of personality disorder grounded in a coherent theory of psychopathology.

The *DSM-III-R* vs. *DSM-IV* Objection

As to the criticism advanced by Rogers et al. that the MCMI-II cannot be used to assess *DSM-IV* disorders because it had only been assessed against the *DSM-III-R* criteria current when that instrument was developed, we wish to point out that both the *DSM-III-R* and *DSM-IV* reflect to a very great extent the theories of Theodore Millon (1981; Millon & Davis, 1996), who served as a committee member of the task force on personality disorders for both the DSM-III and *DSM-IV*. We also note that the criteria for personality disorders differ very little from *DSM-III-R* to *DSM-IV*, with the exception of Depressive, Self-Defeating, and Sadistic Personality Disorders. The various MCMI versions are the only personality tests grounded in a coherent theory of personality that stresses personality disorders rather than Axis I conditions, and which are specifically keyed to the DSM, and that have base rate considerations built into the scoring system, improving the hit rate, and presenting validity results in a format that explicitly states error rate for both sensitivity and positive predictive power.

The Content Validity Objection

McCann & Dyer (1996) characterise the *DSM-IV* as the ultimate learned treatise in forensic psychological work, (i.e. the accepted standard in court against which the credibility of experts' conclusions is evaluated). Dyer (1997) concludes that the MCMI-III has content validity against the *DSM-IV* that is superior to any other major personality instrument. This opinion is based on the presentation of content validity in the second edition of the MCMI-III manual, in which the *DSM-IV* diagnostic criteria for each disorder are listed on one side of the page and the corresponding MCMI-III item(s) on the other.

One of the major criticisms made by Rogers et al. concerning content validity of the MCMI-III is weakened by their failure to cite the most current version of the MCMI-III manual (Million, Davis, & Millon, 1997). In their analysis, Rogers et al. state that the procedures for selecting items and obtaining expert judgments as to item content are not adequately described. On the contrary, the current edition of the MCMI-III manual describes the procedures for creating item pools and assigning items to scales based on expert judgments (p. 24–26). Moreover, the current edition of the manual details constructs that were used to write specific items, the number of items initially written for various scales, and the fact that six out of eight clinicians (i.e., mental health professionals who were well versed in the theory and nosology" upon which the MCMI-III is based, Millon, Davis, & Millon, 1997, p. 25) had to independently agree on the assignment of items to scales while also being unaware of the constructs that originally guided creation of each item. The MCMI-III manual outlines the specific MCMI-III items that parallel individual *DSM-IV* personality disorder criteria (See, Millon, Davis, & Millon, 1997, pp. 27–48). Therefore, the procedures for selecting items and assigning them to scales on the MCMI-III are described in the test manual and presented in a manner that permits direct comparisons against *DSM-IV* personality disorder criteria. In our view, the provision of validity evidence in the form of recent criterion-related research (Davis, Wenger, & Guzman, 1997) and the aforementioned content validity presentation offers a comprehensive and persuasive demonstration of the test's validity for forensic purposes.

The Assessment of Elements of Legal Standards Objection

As for the assertion by Rogers et al. that the MCMI cannot be used to address elements of legal standards, there is some case law and legal authority that point to precisely such applications. McCann & Dyer (1996) note that in criminal cases where mental state is raised as a defense, a common objection by prosecutors is that personality disorders are not mental disorders and therefore cannot be used to support such a defense. McCann & Dyer cite a New Jersey Supreme Court case, *State v. Galloway (1993)* in which the Court held that a defendant's Borderline Personality Disorder was a mental disorder capable of affecting his cognitive functioning in such a way as to negate the knowing and purposeful elements of the mental state required for the crime of murder. In New York, the defense of extreme emotional disturbance requires that courts view circumstances from the defendant's perspective and "such disabilities as borderline retardation, *underlying personality disorders,* and long term depression must be taken into account" (Greenberg, 1996, p. 186–187, emphasis added). These are clearly instances in which the diagnosis of personality disorder has relevance to the ultimate legal issue before the court and in which such a diagnosis may directly address an element of a legal standard.

There are numerous other examples of forensic applications for the MCMI-III to be found in the literature that dispute the contention by Rogers et al. (1999) that "its applicability to forensic issues remains virtually untested" (p. 439). In fact, this statement misses a point underscored by Heilbrun (1992) that "[i]t is…misguided to criticize psychological tests for being only weakly or indirectly re-

lated to legal issues" (p. 269). Heilbrun noted that psychological test date serve as a source of data to formulate, confirm, or disconfirm hypotheses about psychological constructs that are relevant to the forensic issues under consideration in a particular case. This is especially applicable to situations in which the guiding legal standard includes "mental disease", "mental disorder", or similar language as an element.

There are several areas where the MCMI-III provides very useful data that can inform consideration of forensically related issues, including substance abuse (Craig, 1997; Flynn & McMahon, 1997), posttraumatic stress disorder (Craig & Olson, 1997; Hyer, Brandsma, & Boyd, 1997), domestic violence (Gondolf, 1999), violence risk assessment (Kelln, Dozois, & McKenzie, 1998), and those outlined by McCann & Dyer (1996). Thus, the criticism by Rogers et al. that the MCMI lacks direct applicability to forensic issues completely overlooks that fact that any psychological test is but on source of data in a comprehensive assessment and is precisely the misguided criticism that Heilbrun (1992) cautions against.

We note that neither the MCMI-III nor the MCMI-II can be used as a direct measure of such functional legal criteria as the ability to adequately assist counsel, to rationally and factually understand proceedings, and the like. Indeed, Millon has never claimed that his Inventories have the sort of specific applicability, and we endorse the use of other measures focused on those specific legal standards where such referral questions are to be addressed.

The Logic of the Multitrait-Multimethod Model

Before considering the procedures employed by Rogers et al. in their analysis, it is instructive to review the basic concepts of Campbell and Fiske's multitrait-multimethod model that served as a basis for their study. Suen (1990) states that Campbell and Fiske (1959) identified three points along the correlation continuum: a) the correlation between maximally similar measures, which corresponds to reliability, b) the correlation between two maximally dissimilar measures of the same construct, which corresponds to convergent validity, and c) the difference between the convergent validity coefficient and the correlation between maximally dissimilar measures of different constructs, which demonstrates discriminant validity. Wiggins (1973) in commenting on this design, stresses the necessity for different methods of assessment and offers the example of self-report, peer ratings, and situational tests as the three methods contributing to the heteromethod aspect of the analysis. Wiggins states "…clear-cut evidence for construct validity requires a demonstration that construct (trait) variance exceeds method variance in the situation under consideration." (p. 409) In other words, common method variance is a significant source of error that is addressed by true multimethod techniques that involve something other than self-report as the alternative method in the validation of a self-report measure. Further, Wiggins discusses applications of multitrait-multimethod designs in the context of his stance that "it is important that relationships between the personality scale and outside variables be demonstrated and that at least some of these outside variables be nontest behaviors." (p. 406, italics in the original). In regard to the requirement that monotrait-heteromethod correlations be higher than heterotrait-monomethod correlations, Wiggins notes "Although this requirement is an ideal one, it may be inappropriate to apply it too rigidly." (p. 408) Again, this is in the context of a true heteromethod model involving nontest alternative measures. The point is that common method variance can elevate the heterotrait correlations artificially, comprising discriminant validity. We note that the Rogers et al. requirements for discriminant validity display the rigidity that Wiggins warns against. We acknowledge that there is a long history of using multitrait-multimethod designs to assess self-report instruments against other self-report instruments and, in fact, the McCann (1991) study employed such a design. We accept this as a useful way of exploring similarities and differences among scales of self-report instruments through examining the patterns and magnitudes of the scale intercorrelations. On the other hand, we regard it as a blatant misuse of the method to subject a test-test analysis to rigid "comparison violation" standards, especially when the study involves a 13 by 13 matrix rather than the usual comparison of only 2 or 3 scales.

As Dyer (1994) points out, the scales of the MCMI-II are affected by response set as measured by Scales Y (Desirability) and Z (Debasement). It is not a great stretch of the imagination to conclude that scale scores on other major personality assessment instruments including the MMPI are similarly affected. This characteristic of self-report tests, in fact, is one of the problems that the multitrait-multimethod approach to test validation was devised to circumvent. The fact that some subjects tend to register high scores on most or all scales of an instrument because of a pathology-oriented response set and that others register low scores on most or all scales because of a socially desirable response set contributes to spuriously high correlations among scales measuring both similar and dissimilar constructs across instruments. This is one aspect of what the originators of the multitrait-multimethod design term "common method variance", which they regard as a significant source of error. On the other hand, common method variance, including response style distortion, is not a factor in true multimethod data sets such as those that correlate self-report data with ratings by others.

Mechanics of the Rogers et al. Procedure

The criteria for determining whether convergent and discriminant validity have been demonstrated must be chosen in a responsible and balanced manner. Rogers et al. have designated as their criteria the following test: that scales must correlate the highest with the corresponding scales of the counterpart instruments according to a strict count of numbers of monotrait versus heterotrait coefficients, with instances of greater heterotrait than monotrait correlations constituting "comparison violations" after Bagozzi and Yi (1991) and Byrne and Gofin (1993). This procedure is inappropriate for the Rogers et al. meta-analysis for two other reasons besides the common method variance problem cited above. In the first place, the sheer number of comparisons in a 13 by 13 matrix represents a formidable hurdle for any personality measure to surmount in meeting this standard. Under the Rogers et al. procedure, if an MCMI scale correlates even slightly more highly with any of the 12 other composite measures than the counterpart personality disorder, a "comparison violation" is recorded. Additionally, the Rogers et al. study did not detail the individual comparisons, but presented tabled results in terms of averaged convergent validity (cv), heterotrait-monomethod (dv1), and heterotrait-heteromethod (dv2) correlations across studies. In fact, close scrutiny of the method discloses that Rogers et al.'s dv1 and dv2 composite variables are actually *averages of averages*. Thus it is not possible to glean from the study what the patterns of intercorrelations were for the various MCMI-II scales with their counterparts in the studies employed in the Rogers et al. analysis. We have no idea whether the comparison violations were by less than a few points and whether the non-counterpart measure with which the MCMI-II scale correlated more highly was logically related to it, as for example, Schizotypal correlating better with Schizoid than with the counterpart Schizotypal measure, or Antisocial correlating better with Aggressive (Sadistic), or Self-Defeating correlating better with Dependent. The familiar multitrait-multimethod matrix in which such results are typically presented is notable absent from the Rogers et al. study. From the perspective of understanding how the scales are functioning, it is the magnitudes of the differences between coefficients and the patterns of correlations for each of the MCMI-II scales that are of much greater interest than a mechanical counting up of "comparison violations" that could conceivably amount to as little as one point.

The Rogers et al. procedure is, in our opinion, an ill-informed pastiche of meta-analysis and the convergent-discriminant validity model that reflects a fundamental misunderstanding of some of the basic concepts of each technique. In contrast to the original Campbell and Fiske formulation, in which the term multimethod means using truly different methods and not merely test-test comparisons, common method variance resulting from Rogers et al.'s test-test design produces artificially high heterotrait-"heteromethod" correlations. The basic rationale for development of the multimethod design was to eliminate the effects of such common method variance to see whether it is the trait and not merely the assessment method that is responsible for the correlation. The Campbell and Fiske examples cited in psychometrics tests typically involve a 3 by 3 matrix where error associated with the number of variables is miniscule compared to the Rogers procedure, a 13 by 13 matrix. Thus, to propose a monomethd (self-report correlated with self-report) study as an appropriate paradigm for assessing discriminant validity of one self-report measure designated as predictor and other self-report measures designated as criteria is to load the case against finding discriminant validity for the predictor. The compound the matter, the individual scale scores on the instruments (the MCMI included) are not magic trait thermometers, but are also affected by unreliability. While the MCMI-II meets Heilbrun's (1992) .80 criterion for all personality disorder scales and the MCMI-II meets it for most scales, other instruments do not fare as well. As McCann & Dyer (1996) point out, the MMPI-2, which figures so prominently in the Rogers et al. meta-analysis, has several scales (including Paranoia, O-H, M-F) that have a reliability of less than .50. Unreliability limits the capacity of a scale to correlate with another variable, as error variance (which reaches as much as 75% in some MMPI-2 scales) by definition cannot be shared with another variable. The Rogers et al. procedure therefore necessarily results in such an enormous amount of psychometric error due to common method variance and unreliability of criteria that their "findings" of low discriminant validity may be characterised as nothing more than methodological artifacts. Specifically, it cannot be established under this procedure whether high heterotrait correlations, purportedly indicative of poor discriminant validity, are simply the artifactual result of both scales' being affected heavily by social desirability. Similarly, it cannot be established under this procedure whether low monotrait correlations are simply the result of poor reliability of individual scales of the "criterion" instruments. This latter problem is fatally compounded by another feature of the Rogers et al. composite procedure, as described below.

An additional concern is that the Campbell and Fiske method was designed for the assessment of the validity of unidimensional personality traits as assessed by self-report measures. The difference between such trait scales and the polythetic conception of personality disorders in Millon's theory, as reflected in the scales of the MCMI, is that the more factorially pure trait scales will be much more likely to show convergent, and especially discriminant, validity against other factorially pure trait measures. Personality

disorders are diagnostic categories and not unidimensional traits. Thus, methods that were devised to assess the later are less appropriate for assessing the validity of the former than are diagnostic criteria that reflect the complexity of the polythetic disorder. This point is a minor one, however, as the great majority of personality disorder scales of both the MCMI-II and MCMI-III have internal consistency reliabilities above .80 and are therefore not precluded by unreliability from registering high correlations with other measures, provided that those measures are adequately reliable.

All things considered, however, the appropriate criterion for assessing the validity of a personality disorder scale, which is a diagnostic measure, is a nontest diagnosis of the disorder. Similarly, the appropriate statistical paradigm for assessing the validity of a personality disorder scale is a presentation of the diagnostic efficiency of that scale in terms of the usual classification efficiency statistics, and not a series of Pearson product-moment correlation coefficients, which would not provide any information as to the percentage of cases in which scores on the test correctly diagnose the condition or the percentage of cases in which the test picks up individuals who have the disorder.

The methodological problems in the Rogers et al. analysis of the MCMI's convergent validity are no less serious. The mono-trait-"heteromethod" correlations in the Rogers et al. study are simply a hodgepodge resulting from a misuse of Fisher's z transformation, which was never intended to aggregate correlations between 15 different criterion instruments and a single predictor. The z transformation is properly used as a tool for aggregating correlations between the same two variables calculated on different samples. Combining different outcome measures is an accepted practice in meta-analysis, as for instance in the study of effectiveness of psychotherapy, but even the most die-hard proponents of that technique retain an awareness of the reality that these aggregates are not genuine variables, but composite estimates. Such composites cannot be applied to an eyeball analysis of their relative magnitudes under rigid pass-fail criteria. In fact, discussions of statistical and psychometric considerations in the meta-analysis of test validity (Schmidt, 1988; Hedges, 1988) present the topic in terms of aggregating correlations of a test with the same criterion measure across different samples. Schmidt (1988) points out that variations in the criterion measure across samples in poorly designed validity generalization studies obscure the true validity of a test because of the resulting statistical artifacts.

The Rogers et al. study employs composite correlations that, for each disorder, assess shared variance between a single test, the MCMI, and a hodgepodge of 14 measures plus clinician diagnoses captured in a single *coefficient*. The MCMI is the constant Variable A in these correlations and Variable B consists of all of the different counterpart measures that make up the average. Not only do we have the reliability problem of individual measures, we also have error due to peculiarities in the intercorrelations of the 15 "criterion" measures, none of which are reported. We know, however, that at least two of these measures, the PDQ-R and PDE, correlate poorly (Hunt & Andrews, 1992) and that the lack of convergence among various personality disorder measures, which conceptualise the disorders in different ways, has been recognized as a problem (Renneberg et al., 1992).

At the other pole of the error spectrum, we have artificial correlations between the heterotrait-heteromethod measures because of common method variance in the case of self-report measures, which account for most of the Rogers et al. meta-analysis. Thus the hurdle that Rogers et al. impose is that each personality disorder scale of the MCMI-II must correlate highest with a corresponding composite made up of tests that do not correlate well with each other, which have been shown to perform moor poorly than the MCMI-II in some respects, and are in part artificially correlated with all MCMI-II scales because of common method variance. Rogers et al. conclude that the failure of the MCMI-II to demonstrate convergent and discriminant validity against these peculiar agglomerations indicates that the MCMI-II fails as scientific evidence under Daubert. One wonders how, if some of the components of the Rogers et al. composite criteria (in the form of averaged correlations) do not correlate with each other, the MCMI-II could possible be expected to correlate highly with all of them? As a practical matter, if they are being used as some sort of composite criterion, then their degree of intercorrelation, or lack thereof, represents the reliability (internal consistency) of the criterion.

It is axiomatic in designing criterion-related test validation studies that the reliability of the criterion is a crucial consideration. While an empirical demonstration of criterion reliability is not of paramount importance in studies where the test correlates very highly with criterion measure (in which case the criterion is necessarily a reliable one), studies in which low or moderate test-criterion correlations are interpreted as evidence for the test's lack of validity are in a different class. In such cases, it is incumbent upon the researcher to demonstrate that the moderate or low criterion-related validities that are interpreted as evidence of the test's lack of worth are not merely artifacts that are due to criterion unreliability. In this case, we must take into consideration not only the reliabilities of the individual criterion measures that make up the meta-analytic composites for each personality disorder scale, but also the fact that combining them in this way makes their

intercorrelations, or lack thereof, a further source of criterion unreliability. We note that nowhere in the Rogers et al. article is there to be found any indication of an awareness on the part of the authors that there could conceivably be any problem whatsoever associated with criterion unreliability.

To compound this flaw in Rogers et al. study, they fault the MCMI for not meeting the specific effect size standards proposed by Fiske & Campbell (1992) in its correlations with other variables. It is quite surprising that Rogers et al. do not make any reference to the fact that the effect sizes to which they refer are presented as standards for correlations between a single predictor and a single criterion measure. Thus, Rogers et al. set up a situation in which the MCMI must correlate at a high level simultaneously with 15 different measures that do not have high correlations with each other and whether or not this series of test-criterion correlations is of an acceptable magnitude is gauged against effect sizes that do not apply to such composites. We submit that this is a fatal bias built into the very design of the study and that the fact that the obtained MCMI-composite correlations do not meet the specified effect sizes is of absolutely no significance whatsoever.

As we have seen from the MCMI-III fiasco described by Retzlaff (1996) the use of an unreliable criterion dooms any validity study to failure. It is also noted that Rogers et al. interpret the findings of Retzlaff (1996) in a misleading manner when they state that his "reanalysis of Millon's own data indicated diagnostic inaccuracy of the MCMI-III for Axis II disorders" (p. 432). In fact, a careful reading of Retzlaff's conclusions reveals that it is the nature of the original MCMI-III validity study, and not the MCMI-III itself, that was problematic. Retzlaff (1996) noted that several pieces of data in the original manual suggested that the MCMI-III is not invalid, but rather that the original validity study was flawed due to the use of a poor external criterion. He cited the high correlations between the MCMI-III and MCMI-II as evidence that there was not likely to be a sharp drop in validity between the two instruments and that the MCMI-III scales demonstrated strong concurrent validity against other self-report measures of related constructs. Thus Retzlaff's conclusion was not that the MCMI-III was diagnostically inaccurate, but rather that the weak operating characteristics cited in the original MCMI-III manual were "most likely…the result not of a suddenly poorer test but of a weak validity study" (p. 437). This difficulty has been overcome with the publication of a much improved validity study on the MCMI-III (Davis, Wenger, & Guzman, 1997; Millon, Davis, & Millon, 1997)as well as independent research that supports the validity of the MCMI-III (Craig, 1997; Craig & Bivens, 1998; Craig & Olson, 1997; Dyce, O'Connor, Parkins, & Janzen, 1997; Gondolf, 1999; Kelln, Dozois, & McKenzie, 1998). We find it surprising that the newer heteromethod study by Davis, Wenger, & Guzman of the validity of the MCMI-III against a reliable criterion of clinicians' diagnoses structured through an objective rating guide is not even mentioned in Rogers et al., let alone included in their analysis. As dyer (1997) notes, the Davis, Wenger, & Guzman study indicates criterion related validity for the MCMI-III that is as good as or better than the validity of the MCMI-II. In our opinion the omission of that study is a critical flaw that renders the Rogers et al. analysis of the MCMI-III meaningless. It misrepresents the current state of the empirical validity evidence for the MCMI-III and therefore its forensic applicability.

REFERENCES

Bagozzi, R.P. & Yi, Y. (1991) Multitrait-multimethod matrices in consumer research. *Journal of Consumer Research, 17*, 426–439.

Byrne, B.M. & Goffin, R.D. (1993) Modeling MTMM data from additive and multiplicative covariance structures: An audit of construct validity concordance. *Multivariate Behavioral Research, 28*, 67–96.

Campbell, D.T. & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105.

Craig, R.J. (1997). Sensitivity of MCMI-III scales T (drugs) and B (alcohol) in detecting substance abuse. *Substance Use and Misuse, 32,* 1385–11393.

Craig, R.J. (1999). Testimony on the Millon Clinical Multiaxial Inventory: Review, commentary, and guidelines. *Journal of Personality Assessment, 73,* 290–304.

Craig, R.J. & Bivens, A. (1998). Factor structure of the MCMI-III. *Journal of Personality Assessment, 70,* 190–196.

Craig, R.J. & Olson, R.E. (1997). Assessing PTSD with the Millon Clinical Multiaxial Inventory-III. *Journal of Clinical Psychology, 53,* 943–952.

Daubert v. Merrill Dow Pharmaceuticals, Inc., 2786 L.Ed.2d (U.S. 113 S.Ct. 1993).

Davis, R.D., Wenger, A., & Guzman, A. (1997). Validation of the MCMI-III. In T. Millon (Ed.) *The Millon inventories: Clinical and personality assessment* (pp. 327–359). New York: Guilford.

Dyce, J.A., O'Connor, B.P., Parkins, S.Y., & Janzen, H.L. (1997). Correlational structure of the MCMI-III personality disorder scales and comparison with other data sets. *Journal of Personality Assessment, 69,* 568–582.

Dyer, F.J. (1994). Factorial trait variance and response bias in MCMI-III personality disorder scale scores. *Journal of Personality Disorders, 8,* 121–130.

Dyer, F.J. (1997). Application of the Millon inventories in forensic psychology. In T. Millon (Ed.) *The Millon inventories: Clinical and personality assessment* (pp. 124–139) New York: Guilford.

Flynn, P.M. & McMahon, R.C. (1997). MCMI applications in substance abuse. In T. Millon (Ed.), *The Millon inventories: Clinical and personality assessment* (pp. 1873–190) New York: Guilford.

Gondolf, E.W. (1999). MCMI-III results for batterer program participants in four cities: Less "pathological than expected". *Journal of Family Violence, 14,* 1–17.

Greenberg, R.A. (Ed.) (1996) New York criminal law. St. Paul, MN: West.

Hedges, L.V. (1988) The meta-analysis of test validity studies: Some new approaches. In H. Wainer & H.I. Braun (eds.) *Test validity.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Heilbrun, K. (1992). The role of psychological testing in forensic assessment. *Law and Human Behavior, 16,* 257–272.

Hunt, C. & Andrews, G. (1992) Measuring personality disorder: The use of self-report questionnaires. *Journal of Personality Disorders, 6,* 125–133.

Hyer, L., Brandsma, J., & Boyd, S. (1997) The MCMI's and posttraumatic stress disorder. In T. Millon (Ed.), *The Millon inventories: Clinical and personality assessment* (pp. 191–216). New York: Guilford.

Kelln, B.R.C., Dozois, D.J.A., & McKenzie, I.E. (1998). An MCMI-III discriminant function analysis of incarcerated felons: Prediction of subsequent institutional misconduct. *Criminal Justice and Behavior, 25,* 177–189.

McCann, J.T. (1990). A multitrait-multimethod analysis of the MCMI-II clinical syndrome scales. *Journal of Personality Assessment, 55,* 465–476.

McCann, J.T. (1991). Convergent and discriminant validity of the MCMI-II and MMPI personality disorder scales. *Psychological Assessment, 3,* 9–18.

McCann, J.T. & Dyer, F.J. (1996). *Forensic assessment with the Millon inventories.* New York: Guilford.

Millon, T., Davis, R.D., & Millon, C. (1997). *MCMI-III manual (2nd ed.)* Minneapolis, MN: National Computer Systems.

Renneberg, B., Chambless, D.L., Dowdall, D.J., Fauerbach, J.A., & Gracely, E.J. (1992) The Structured Clinical Interview for *DSM-III-R*, Axis II and the Millon Clinical Multiaxial Inventory: A concurrent validity study of personality disorders among anxious outpatients. *Journal of Personality Disorders, 6,* 117–124.

Retzlaff, P. (1996). MCMI-III diagnostic validity: Bad test or bad validity study. *Journal of Personality Assessment, 66,* 431–437.

REFERENCES *(continued)*

Rogers, R., Salekin, R.T., & Sewell, K.W. (1999). Validation of the Millon Clinical Multiaxial Inventory for Axis II disorders: Does it meet the Daubert standard? *Law and Human Behavior, 23,* 425–443.

Schmidt, F.L. (1988). Validity generalization and the future of criterion-related validity. In H. Wainer & H.I. Braun (eds.) *Test validity.* Hillsdale, NJ: Lawrence Erlbaum Associates.

State v. Galloway, 133 N.J. 631, 628 A.2d 735 (1993).

Suen, H.K. (1990). *Principles of test theories.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Wiggins, J.S. (1973). Personality and prediction: *Principles of personality assessment.* Reading, MA: Addison-Wesley.

Wise, E.A. (1994). Managed care and the psychometric validity of the MMPI and MCMI personality disorder scales. *Psychotherapy in Private Practice, 13,* 18–97.

PsychCorp | Pearsonclinical.co.uk

Version 1.0 | Clinical | November 2021 | DCL1: Public