

Diagnostic Agreement Between the Personality Disorder Examination and the MCMI-II

Stephen Soldz and Simon Budman
*Mental Health Research Program,
Harvard Community Health Plan and
Department of Psychiatry,
Harvard Medical School*

Annette Demby and Jocelyn Merry
*Mental Health Research Program,
Harvard Community Health Plan*

In an attempt to compare different methods for assessing personality disorder, this study compared the Millon Clinical Multiaxial Inventory-II (MCMI-II; Millon, 1987), a self-report questionnaire, and the Personality Disorder Examination (PDE; Loranger, 1988), a semistructured clinical interview. Subjects ($N = 97$) were mental health outpatients of a health maintenance organization in New England. The instruments were compared in terms of the presence of personality disorder, the number of diagnoses assigned to a patient, and agreement in specific diagnoses and in cluster assignment. Agreement between the two instruments was low; the two instruments exhibited greater agreement in predicting the absence of diagnoses than their presence. Agreement was best for the borderline and avoidant diagnoses. Correlations between scales exhibited somewhat better agreement than was evident for diagnoses. Analyses at the cluster level resulted in moderate correlations between the instruments. Very high intracluster correlations were found for the MCMI-II, but not for the PDE.

As increased attention in recent years has focused on personality disorders, the need for good assessment instruments has grown. If future diagnostic systems are to be based on good research data, the first step is to be able to identify which individuals have which disorder. In order for agreement to exist regarding diagnosis, instruments claiming to measure personality disorders need to exhibit convergent validity.

Since the publication of the *Diagnostic and Statistical Manual of Mental Disorders* (3rd ed. [DSM-III]; American Psychiatric Association, 1980) and the *Diagnostic and Statistical Manual of Mental Disorders* (3rd ed., rev. [DSM-III-R]; American Psychiatric Association, 1987), a number of instruments have been developed to assess personality disorders. These have included both structured interviews, such as the Personality Disorder Examination (PDE; Loranger, 1988), the Structured Interview for DSM-III Personality Disorders (SIDP; Pfohl, Stangl, & Zimmerman, 1982) and the Structured Clinical Interview for DSM-III-R Personality Disorders (SCID-II; Spitzer, Williams, & Gibbon, 1987), as well as questionnaires, among which the most notable are the Personality Diagnostic Questionnaire (PDQ-R; Hyler, Reider, & Williams, 1987) and the Millon Clinical Multiaxial Inventory, which has now entered a second edition (MCMI-II; Millon, 1987).

Several attempts have been made to compare diagnoses generated by these various instruments to assess the degree of agreement. Perry (1991) identified seven studies comparing DSM-III or DSM-III-R personality disorder diagnoses. Across all seven studies, the median kappa for diagnostic agreement in making individual diagnoses was .25. If only studies comparing interview instruments were examined, the median kappa went up to .38.

The PDE is one of the most popular semistructured clinical interviews for diagnosing personality disorders. The PDE is especially important because it is being used by the World Health Organization and has been translated into several languages. This article will examine the diagnostic concordance between the PDE and the MCMI-II, which is probably the most used self-report inventory purporting to diagnose personality disorders.

One study has compared the MCMI with the SIDP in a sample of 40 recent-onset schizophrenic patients after recovery (Hogg, Jackson, Rudd, & Edwards, 1990). Median kappa between instruments for diagnoses that occurred with sufficient frequency was .14, with a range from .00 to .34. Pearson correlations between dimensional scores ranged from -.03 to .60, with a median of .26.

Alternatively, two studies compared the PDE with other instruments. Hyler, Skodol, Kellman, Oldham, and Rosnick (1990) compared the PDE with the PDQ-R in a sample of 87 patients applying for inpatient treatment. The median kappa for agreement between the two instruments was .37, with a range from -.02 to .54. Skodol, Oldham, Rosnick, Kellman, and Hyler (cited in Perry, 1991) compared the PDE with the SCID-II and with consensus diagnoses using the longitudinal expert evaluation using all available data (LEAD) technique (Spitzer, 1983). Agreement between the two interviews was moderate, with a median kappa of .50 and a range from .14 to .66. Dimensional scores showed substantial agreement, with a median Pearson r of .77 (range from .58 to .87). In comparing the PDE with the LEAD standard, the concordance was substantially lower (median kappa of .25, range -.01 to .41).

This article compares the PDE and MCMI-II in an outpatient sample referred for treatment for personality disorders. Both of these instruments have recently been revised using information from previous experience in order to more closely match DSM-III-R (Loranger, 1988; Millon, 1985). In addition to providing the first

comparison of these new instruments, our study also provides some of the first data on the new Sadistic and Self-Defeating personality disorders that have been included in the *DSM-III-R* appendix.

How agreement between instruments is assessed will be affected by how one conceptualizes personality disorders. Although *DSM-III-R* conceptualizes personality disorders as discrete categories that are either present or absent in a given person, others have argued that personality disorders are better conceptualized as being on a continuum with normality, such that personality disordered individuals possess normal characteristics to an unusual degree (e.g., Costa & McCrae, 1990; Frances & Widiger, 1986; Soldz, Budman, Demby, & Merry, 1993; Widiger, 1991; Wiggins & Pincus, 1989). The categorical approach leads to the question as to whether two personality disorder assessment tools place people in the same categories. In contrast, the dimensional approach leads to the conclusion that agreement between instruments is better approached by means of correlations between comparable dimensional scores generated by the two instruments.

In line with this thinking, we compared the PDE and the MCMI-II, examining them both in terms of the categorical diagnoses generated and in terms of the correlations between similar dimensions that are assessed by the two instruments. Because, as is well known (Widiger, 1991), individuals who are personality disordered often receive more than one personality disorder diagnosis from most diagnostic instruments, it can be hard to assess the nature of disagreement that exists between instruments at the level of individual diagnoses. We therefore also compared the PDE and the MCMI-II in terms of the clusters that are proposed in *DSM-III-R* (American Psychiatric Association, 1987) for grouping personality disorders and examined the two instruments' degree of agreement in assigning any personality disorder to a patient.

METHODS

Subjects

Subjects were 97 consecutive patients referred for possible inclusion in a study of group psychotherapy for personality disorders at Harvard Community Health Plan, the largest health maintenance organization in New England. The exclusion criteria for the groups included current suicidality, psychosis, or any other Axis I condition that would need to be the predominant focus of the treatment. Because of reservations about the utility of interactional group psychotherapy with Antisocial and Cluster A patients, they were also excluded, although, in a couple of cases where it was judged on clinical grounds that the primary personality pathology was not in Cluster A, despite the presence of a diagnosis in this category, these patients were included. We were hoping by these criteria to produce a patient sample that would benefit from group therapy focused on long-standing personality problems. The patients in this sample were subjected to preliminary screening through phone calls between research staff and their referring clinicians. Patients who obviously did not meet our criteria were excluded at this stage, although, in the case of doubt, the

patient was interviewed by research staff and included in the sample. This sample included 52 females and 45 males with a mean age of 36.8 ($SD = 8.04$). They were predominantly White (89 were White, 1 Asian, 5 African American, 1 Hispanic, and 1 Native American). Fifty-nine were single, 4 cohabiting, 13 married, 3 separated, 17 divorced, and 1 was widowed. They were a highly educated group, with 44 having at least some graduate education (29 with graduate degrees); another 32 had college degrees, and 18 had some college education. Because of the nature of the referrals, all patients were presumed by their clinicians to be personality disordered.

Instruments

PDE. The PDE, second edition, is a semistructured clinical interview for diagnosing personality disorders. It consists of 126 items arranged in six categories (e.g., Self or Work), along with a detailed scoring manual. Each item assesses part or all of a *DSM-III-R* personality disorder criterion and is rated on a 3-point scale: *absent or normal* (0), *exaggerated or accentuated* (1), *meets criteria or pathological* (2). Items consist of one or several primary questions, with follow-up questions. All positive responses are followed by requests for examples. After the provided questions are exhausted, the clinical interviewer is free to ask additional questions until he or she is able to score the item. A sample item is "Do you often feel bored or empty inside?" for the *DSM-III-R* Borderline criterion "Chronic feelings of emptiness or boredom". A positive response to this question would be followed by "Does that upset you or cause any problems for you?", followed by "Tell me about it." The PDE generates probable (subthreshold number of *DSM-III-R* criteria met) and definite diagnoses for each of the *DSM-III-R* diagnoses, including those diagnoses included in the appendix, namely Sadistic and Self-Defeating, and NOS (Not Otherwise Specified) for patients meeting a large number of personality disorder criteria, without meeting the criteria for any particular diagnosis. Dimensional scores can also be generated for each diagnosis by adding the ratings on all the criteria making up a diagnosis. No interrater reliabilities for the revised PDE have yet been reported; interrater reliabilities for the first version range from .54 to .93, with a median kappa of .78 (Loranger, 1988). The PDE takes between 2 and 3.5 hr to administer.

MCMI-II. The MCMI-II is the second edition of a self-report questionnaire designed to provide comprehensive diagnostic information about a patient (Millon, 1987). It consists of 175 items scored using a true/false format. It generates scale scores for 12 primary symptom scales, 10 personality disorder scales (including scales for the 2 diagnoses included in the *DSM-III-R* appendix), and 3 severe personality disorder scales (Borderline, Paranoid, and Schizotypal). The MCMI-II conceptualizes personality disorders using a compromise between Millon's (1981, 1990; Millon & Everly, 1985) ideas and those represented in *DSM-III-R*. In the two cases (Aggressive-Sadistic and Compulsive) in which the MCMI-II uses a name for a scale different from the corresponding *DSM-III-R* disorder, we have

used the *DSM-III-R* name for consistency. Personality disorder diagnoses were made when a patient had base rate adjusted scale scores greater than or equal to 85 (Millon, 1987) on that personality disorder scale.

Interviewers

Interviewers for the PDE were one clinical psychologist and four clinicians with master's degrees in a mental health discipline. All interviewers had extensive previous experience administering structured clinical interviews. Three of the interviewers were trained by Loranger, the developer of the PDE; the other two interviewers were trained by two of the initial three.

RESULTS

Table 1 lists the frequency of diagnoses resulting from the MCMI-II and the PDE. For the PDE we have reported both the definite diagnoses, meeting *DSM-III-R* criteria, and the probable diagnoses that may be subthreshold for the *DSM-III-R* diagnosis. As can be seen, the number of patients (66) assigned personality disorders by the MCMI-II is closer to the number of patients assigned PDE probable and definite diagnoses (75) than to the number of given to the definite diagnoses alone (43). Because of this convergence in frequencies, we report results for both PDE definite and probable plus definite diagnoses. The PDE resulted in a

TABLE 1
Frequencies of Personality Diagnoses With the PDE and MCMI-II

MCMI-II	Definite Diagnosis		Probable or Definite Diagnosis		
	MCMI-II	PDE	PDE and MCMI	PDE	PDE and MCMI
Paranoid	0	2	0	11	0
Schizoid	12	0	0	3	1
Schizotypal	12	1	0	1	0
Histrionic	8	4	2	6	2
Narcissistic	11	2	0	11	2
Antisocial	8	0	0	2	2
Borderline	9	16	6	23	7
Avoidant	40	18	14	34	22
Dependent	20	5	4	12	7
Obsessive Compulsive	8	5	0	9	0
Passive Aggressive	29	6	1	14	7
Sadistic	14	1	1	3	2
Self-defeating	23	7	4	13	6
Cluster A	18	3	0	14	2
Cluster B	23	18	8	30	12
Cluster C	58	31	21	52	37
Any personality disorder diagnosis	66	43	35	75	56

Note. Total $N = 97$.

TABLE 2
Diagnostic Agreement Between PDE and MCMI-II

MCMI-II	PDE	
	Definite Diagnosis, Kappa	Definite + Probable Diagnosis, Kappa
Paranoid	.00	.00
Schizoid	.00	.09
Schizotypal	-.02	-.02
Histrionic	.29***	.23*
Narcissistic	-.04	.08
Antisocial	.00	.38***
Borderline	.41***	.35***
Avoidant	.30***	.35***
Dependent	.26***	.33***
Obsessive Compulsive	-.07	-.10
Passive Aggressive	-.05	.16*
Sadistic	.12**	.19**
Self-defeating	.18*	.20*
Cluster A	-.06	-.04
Cluster B	.23**	.25**
Cluster C	.09	.25**
Any personality disorder diagnosis	.23**	.26**

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

mean of .69 ($SD = .94$) definite and 1.46 ($SD = 1.57$) probable plus definite diagnoses per patient,¹ whereas the mean for the MCMI-II was 2.00 ($SD = 2.08$).

The degree of agreement between the two instruments for each disorder, and for the presence of any personality disorder diagnosis, is reported in Table 2. Using definite diagnoses, the kappa for agreement between the PDE and the MCMI-II as to the presence of any personality disorder was .23, whereas for PDE probable and definite diagnoses combined, the kappa was .26, indicating, in each case, that there was only moderate convergence of the scales at this level.

In contrast, if we calculate a total personality disorder score from each scale by adding all the scale scores and correlating the two scores (in the case of the PDE we used the dimensional scores; similar findings result if the total number of criteria met is used), the correlation is .64, indicating substantial quantitative agreement as to total degree of personality pathology.

Diagnostic Agreement

As can be seen in Table 2, for four specific disorders, Antisocial, Borderline, Avoidant, and Dependent, the kappas for diagnostic agreement were above .30 for the definite plus probable PDE diagnoses, whereas with definite PDE diagnoses only, only two kappas (Borderline and Avoidant) were above .30. The kappas for

¹If NOS diagnoses are considered, the comparable figures are .80 ($SD = .91$) and 1.73 ($SD = 1.56$).

several disorders were zero, near zero, or slightly negative, indicating no better than chance agreement. This problem was most notable for Paranoid, Schizoid, Schizotypal, Narcissistic, Obsessive Compulsive, and (PDE-definite Passive Aggressive). In fact, the median kappa for PDE-definite diagnoses was 0.0, whereas that for definite plus probable diagnoses was .19. If we restrict the calculation to diagnoses that had frequencies of at least 4 for both instruments, the median kappas increase to .26 for definite and .29 for probable plus definite diagnosis; using a minimum frequency of 5 leads to even lower minimum kappas (.22 for definite and .215 for probable plus definite diagnosis).

Another way of comparing two diagnostic instruments is to use diagnostic efficiency statistics such as sensitivity, specificity, positive predictive power, and negative predictive power (Baldessarini, Finkelstein, & Arana, 1983; Fliess, 1981), which can pinpoint the nature of differences. Assuming one has access to the "correct" disorder diagnosis, the sensitivity of a test is the frequency with which the test gives the diagnosis, given that the disorder is present; specificity is the frequency of the absence of a diagnosis, given that the disorder is not present; positive predictive power is the frequency with which the disorder occurs, given that the test gives the diagnosis; and negative predictive power is the frequency that the disorder is not present, given that the test does not give the diagnosis. In order to apply these statistics it is necessary that one instrument be designated as the standard that the other is trying to match. We have arbitrarily chosen the PDE to be the standard.²

These four sets of statistics are provided in Table 3 for PDE-positive and for PDE-positive plus probable diagnosis patients. We see that the MCMI-II has high specificity and negative predictive power for all diagnoses; the scales tend to agree on who does not have a diagnosis. In contrast, the sensitivities are low for most scales, except for Avoidant, Dependent, Histrionic, and Self-Defeating, whereas the positive predictive power is quite low or even 0 for all scales except Borderline. It is thus clear that the scales tend to agree on who does not have a diagnosis, while frequently disagreeing on who does have a certain diagnosis. Similar results hold for the PDE-definite plus probable diagnoses.

Correlations

In order to examine the agreement between the PDE and the MCMI-II if each scale is treated as a dimensional variable, we correlated the PDE dimensional scores for each *DSM-III-R* disorder with the MCMI-II Axis II scales (Table 4). With the exception of the Obsessive Compulsive dimension, which was negative, nonsignificant, and near zero, the correlations between matching scales were all highly significant. The other correlations range from a low of .34 for Antisocial to a high of .60 for Borderline (median of .41), providing some evidence of convergent

²It is important to realize that if the reader would prefer to regard the MCMI-II as the standard to be predicted by the PDE, the same statistics are applicable; they need only to be renamed: sensitivity becomes predictive power; specificity becomes negative predictive power; positive predictive power becomes sensitivity; and finally, negative predictive power becomes specificity.

TABLE 3
Performance of MCMI-II Relative to PDE

MCMI-II	Sensitivity		Specificity		Positive Predictive Power		Negative Predictive Power	
	DEF	DEF & PRB	DEF	DEF & PRB	DEF	DEF & PRB	DEF	DEF & PRB
Paranoid	0.00	0.00	1.00	1.00	0.00	0.00	0.98	0.89
Schizoid	0.00	0.33	0.88	0.88	0.00	0.08	1.00	0.98
Schizotypal	0.00	0.00	0.88	0.88	0.00	0.00	0.99	0.99
Histrionic	0.50	0.33	0.94	0.93	0.25	0.25	0.98	0.96
Narcissistic	0.00	0.18	0.88	0.90	0.00	0.18	0.98	0.90
Antisocial	0.00	1.00	0.92	0.94	0.00	0.25	1.00	1.00
Borderline	0.38	0.30	0.96	0.97	0.67	0.78	0.89	0.82
Avoidant	0.78	0.65	0.67	0.71	0.35	0.55	0.93	0.79
Dependent	0.80	0.58	0.83	0.85	0.20	0.35	0.99	0.94
Obsessive Compulsive	0.00	0.00	0.91	0.91	0.00	0.00	0.94	0.90
Passive Aggressive	0.17	0.50	0.69	0.74	0.03	0.24	0.93	0.90
Sadistic	1.00	0.67	0.86	0.87	0.07	0.14	1.00	0.99
Self-defeating	0.57	0.46	0.79	0.80	0.17	0.26	0.96	0.90
Cluster A	0.00	0.14	0.81	0.81	0.00	0.11	0.96	0.85
Cluster B	0.44	0.40	0.81	0.84	0.35	0.52	0.86	0.76
Cluster C	0.68	0.71	0.44	0.53	0.36	0.64	0.74	0.62
Any personality disorder diagnosis	0.81	0.75	0.35	0.54	0.53	0.85	0.74	0.39

Note. N = 97. DEF = PDE definite diagnosis. DEF & PRB = PDE definite or probably diagnosis.

TABLE 4
Correlations Between MCMI-II and PDE

MCMI-II	PDE												
	Paranoid	Schizoid	Schizotypal	Histrionic	Narcissistic	Antisocial	Borderline	Avoidant	Dependent	Obsessive Compulsive	Passive Aggressive	Sadistic	Self-defeating
Paranoid	0.38**	-0.01	0.15	0.39**	0.44**	0.14	0.35**	-0.01	0.13	0.15	0.15	0.31**	0.33**
Schizoid	0.15	0.48**	0.39**	-0.27	-0.19	-0.09	-0.04	0.58**	0.11	0.15	0.02	-0.15	-0.07
Schizotypal	0.35**	0.28**	0.39**	0.07	0.27**	0.01	0.23*	0.47**	0.35**	0.26*	0.14	0.11	0.22*
Histrionic	0.10	-0.38**	-0.34**	0.56**	0.39**	0.23*	0.9**	-0.57**	0.00	-0.07	0.21*	0.29**	0.26**
Narcissistic	0.31**	-0.17	-0.13	0.39**	0.41**	0.32**	0.23*	-0.38**	-0.14	0.03	0.20	0.38**	0.23*
Antisocial	0.30**	-0.14	-0.03	0.54**	0.51**	0.37**	0.36**	-0.17	0.02	0.11	0.29**	0.36**	0.27**
Borderline	0.36**	-0.05	0.09	0.57**	0.49**	0.30**	0.60**	-0.09	0.40**	0.25*	0.47**	0.24*	0.43**
Avoidant	0.35**	0.32**	0.46**	0.02	0.06	-0.04	0.21	0.51**	0.29**	0.22	0.25*	0.01	0.22*
Dependent	-0.13	0.11	0.14	-0.01	-0.08	-0.25*	0.08	0.16	0.38**	-0.02	0.05	-0.17	0.07
Obsessive Compulsive	0.03	0.25*	0.23*	-0.19	-0.08	-0.17	-0.06	0.29**	0.00	-0.05	-0.30**	-0.09	-0.09
Passive Aggressive	0.45**	0.04	0.22*	0.45**	0.41**	0.22*	0.34**	0.17	0.22*	0.34**	0.41**	0.32**	0.38**
Sadistic	0.35**	-0.03	0.03	0.38**	0.48**	0.31**	0.31**	-0.14	-0.10	0.11	0.18	0.44**	0.25*
Self-defeating	0.26**	0.03	0.18	0.36	0.34**	0.03	0.45**	0.21*	0.44**	0.26*	0.49**	0.05	0.34**

Note. N = 97.
*p < .05. **p < .01.
***p < .001.

validity. Four PDE scales have their highest correlation with the corresponding MCMI-II scale: Schizoid, Antisocial, Borderline, and Sadistic. Similarly, four MCMI-II scales have their highest correlation with the matching PDE scale: Borderline, Avoidant, Dependent, and Narcissistic. Only for Borderline do we thus see clear convergent and discriminant validity for the two measures.

Clusters

We also examined diagnostic agreement between the two instruments at the cluster level. Agreement for Cluster A was extremely poor and at no better than chance levels, with kappas below zero. Agreement was poor whether we used the PDE-definite or PDE-definite plus probable diagnosis. In both instances there was very low sensitivity and positive predictive power and high specificity and negative predictive power. Agreement was somewhat better for Cluster B (kappa = .23 for PDE-definite diagnosis, and kappa = .25 for PDE-definite plus probable), though it was still poor. The specificity and negative predictive power remained high, and the sensitivity and positive predictive power were higher than those for Cluster A. Cluster C was somewhat unusual in that there was a large discrepancy between the agreement indices with the different ways of treating the PDE. With PDE-definite diagnosis the kappa was near 0 (.09) and nonsignificant, whereas the kappa was .25 for PDE-definite plus probable diagnosis. The sensitivities and positive predictive powers were greater than those for the other clusters, whereas the specificities and negative predictive powers were lower. In general, there did not appear to be appreciably better diagnostic agreement when diagnosis was at the cluster level rather than for individual diagnostic categories.

We also wanted to look at agreement between scores on the two instruments at the cluster level. In order to do this we added all the MCMI-II items composing the scales for the personality disorders in each cluster using unit weighting of the items (and appropriate adjustments of sign). For the PDE we added together all scores on items measuring the DSM-III-R criteria for the diagnoses in a given cluster to form a cluster score. Table 5 shows the correlations between these scale

TABLE 5
MCMI-II and PDE Cluster Scores Correlated

	MCMI-II Cluster			PDE Cluster		
	A	B	C	A	B	C
MCMI-II cluster						
A	-	0.56***	0.93***	-0.51***	-0.31**	-0.55***
B		-	0.47***	-0.29**	-0.67***	-0.28***
C			-	-0.49***	-0.26**	-0.61***
PDE cluster						
A				-	0.25*	0.33***
B					-	0.29**
C						-

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

scores. In every case, the correlations are significant, and the correlations between corresponding scales on the two instruments are always the highest correlations for a given cluster, though for Cluster A the matching correlation is only trivially higher than that between MCMI-II Cluster A and PDE Cluster C. The best convergent and discriminant validity is for Cluster B where the correlation between the matching scales is .65, whereas the highest correlation between either of the scales and other scales on the other instrument is .30.

We can also see in Table 5 that the PDE has much better discrimination between clusters than the MCMI-II, the highest intercluster correlation being .36 between Clusters A and C. For the MCMI-II, there was considerably greater overlap between clusters. The lowest intercluster correlation was .51 (between Clusters B and C), whereas the highest was .94 between Clusters A and C.

DISCUSSION

This study compared the latest revisions of two commonly used instruments for diagnosing personality disorders. It was found that the PDE and the MCMI-II had poor to moderate agreement in assigning diagnoses to patients. This was true when evaluating whether patients are given particular diagnoses, or whether patients are considered to possess any personality disorder at all. If attention is restricted to only those diagnoses that were assigned to patients at least four times on both instruments, then the median kappas of .26 and .29 (for definite and definite plus probable PDE diagnoses respectively) are very similar to the median kappa of .25 reported by Perry (1991) for studies comparing personality disorder diagnostic instruments. We can thus conclude that the revised instruments examined here are in agreement no more than were a previous generation of instruments.

In general, the instruments were in better agreement in identifying who did not have a personality disorder than who did possess a particular disorder, as is indicated by the fact that the specificities and negative predictive powers were substantially greater than the sensitivities and positive predictive powers. To some degree this is a result of the low base rates for many of the disorders in our sample. Given that the patients in our sample were all presumed by their clinicians to have personality disorders, the base rates in the current study were probably about as high as will be obtained in most studies with outpatients. Furthermore, examining only disorders that were diagnosed with a minimum frequency of 4 or 5 did not lead to acceptable levels of agreement.

Agreement did not improve when we looked at the cluster level. For Cluster A, agreement was worse than chance, probably because, at least in part, this cluster was deliberately underrepresented in our sample. It may be of concern that the MCMI-II assigned 18 patients to this cluster, whereas the PDE assigned only 3. Given that possession of a diagnosis in this cluster was a rule-out for inclusion in the groups that the patients were applying for and that this exclusion criterion was known to referring clinicians, the PDE figure seems the more appropriate one. We are left with the conclusion that the MCMI-II may be substantially overassigning patients to Cluster A.

Agreement was somewhat better for Clusters B and C when PDE definite and probable diagnoses were taken into account. The highest of these cluster kappas was .25, however, which is no better than that for particular diagnostic categories with frequencies greater than four. Combining patients into clusters does not substantially improve diagnostic agreement. If the clusters are regarded as broad configurations of personality difficulties, these two instruments are not assigning the same patients to these broad configurations.

The instruments did not agree substantially more when we looked simply at whether any personality disorder was assigned to a patient. The kappas of .23 and .26 indicate that there was substantial disagreement at this level as well. We must conclude that at all levels, the MCMI-II and the PDE are often in disagreement as to the presence of personality pathology. The instruments, therefore, cannot be used interchangeably, and attention should be paid in literature reviews of personality disorders to examining effects that may be due to the diagnostic instrument used. Perhaps further research will elucidate characteristics of "PDE personality disorders" or "MCMI-II personality disorders," rather than assuming that personality disorder is a unitary concept across instruments and studies.

Matters are somewhat better when both instruments are considered to be trait measures and are compared correlationally. Only one correlation between matching scales, that for Obsessive Compulsive, was not significant; all others were greater than .30. Only for Borderline, however, was there clear evidence of convergent and discriminant validity in that the Borderline scale for each instrument had its highest correlation with the matching scale on the other instrument. Considering these instruments as trait measures and scoring the PDE quantitatively thus leads to greater evidence of convergence of constructs but little evidence of discriminant validity.

At the cluster level, the story is somewhat better. The correlations between matching scales for the three clusters ranged from .51 for Cluster A to .67 for Cluster B. Both Clusters B and C showed evidence of discriminant validity in that the largest correlations for each scale were with the matching one from the other instrument; this evidence was substantially greater for Cluster B, where the largest off-diagonal correlation was .31, compared with a .67 correlation between the matching scales.

One additional issue raised by a perusal of Table 5 is the considerably higher intercluster correlations for the MCMI-II than for the PDE. The lowest intercluster correlation for the MCMI-II is .57 between Clusters B and C, whereas the greatest is .93 between Clusters A and C. Although it is possible that this latter correlation is partly influenced by the low levels of Cluster A pathology in our sample due to our selection procedures, it should be remembered that the MCMI-II assigned at least one Cluster A diagnosis to 18 patients in our sample. This correlation is probably also due in part to the considerable item overlap between scales that is a feature of the MCMI-II (Millon, 1987; Streiner & Miller, 1989). Clusters A and C share 62 items.

When we questioned whether patients were assigned any personality disorder by each scale, we found results similar to those when cluster membership was

examined. The kappas measuring agreement were .23 and .26 for definite and definite plus probable PDE diagnoses, respectively. In contrast, the correlation between the sum of the MCMI-II personality disorder scores and the sum of the PDE dimensional scores was .64, indicating substantial agreement in assessing degree of personality pathology. This correlation, as well as those for the cluster scores, is as high as or higher than many validity coefficients in the personality literature. At this level as well, the dimensional quantitative approach appears to lead to substantially better agreement between the instruments.

In sum, we have found that the PDE and the MCMI-II have only slight to moderate agreement in assigning diagnoses to patients, no matter what level we examine. The dimensional approach to measuring personality disorder, however, led to considerably greater agreement between the instruments, especially at the cluster and personality disorder/no personality disorder levels. At this time, we must continue to assume that personality disorders diagnosed by one instrument are not identical with those diagnosed by other measures. Future work needs to clarify the nature of personality pathology as assessed by each measure.

ACKNOWLEDGMENTS

Preparation of this grant was aided by a grant from the Harvard Community Health Plan Foundation and by a National Institute of Mental Health Grant #RO1-MH43908-02, Simon H. Budman, PhD, Principal Investigator.

REFERENCES

- American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Washington, DC: Author.
- American Psychiatric Association. (1987). *Diagnostic and statistical manual of mental disorders* (3rd ed., rev.). Washington, DC: Author.
- Baldessarini, R., Finkelstein, S., & Arana, G. (1983). The predictive power of diagnostic tests and the effect of prevalence of illness. *Archives of General Psychiatry*, *40*, 569-573.
- Costa, P., & McCrae, R. (1990). Personality disorders and the five-factor model of personality. *Journal of Personality Disorders*, *4*, 362-371.
- Fliess, J. L. (1981). *Statistical methods for rates and proportions*. New York: Wiley.
- Frances, A. J., & Widiger, T. A. (1986). Methodological issues in personality diagnosis. In T. Millon & G. L. Klerman (Eds.), *Contemporary directions in psychopathology: Toward the DSM-IV* (pp. 381-400). New York: Guilford Press.
- Hogg, B., Jackson, H. J., Rudd, R. P., & Edwards, J. (1990). Diagnosing personality disorders in recent-onset schizophrenia. *The Journal of Nervous and Mental Disease*, *178*, 194-199.
- Hyer, S., Reider, R., & Williams, J. B. W. (1987). *The Personality Diagnostic Questionnaire Revised (PDQ-R)*. New York, NY: New York State Psychiatric Institute.
- Hyer, S., Skodol, A., Kellman, D., Oldham, J., & Rosnick, L. (1990). Validity of the Personality Diagnostic Questionnaire-Revised: Comparison with two structured interviews. *American Journal of Psychiatry*, *147*, 1043-1048.
- Loranger, A. (1988). *Personality Disorder Examination (PDE) Manual*. Westchester, NY: Cornell University Medical College, Department of Psychiatry.

- Millon, T. (1981). *Disorders of personality: DSM-III: Axis II*. New York: Wiley.
- Millon, T. (1985). The MCMI provides a good assessment of the DSM-III disorders: The MCMI-II will prove even better. *Journal of Personality Assessment*, *49*, 379-391.
- Millon, T. (1987). *Manual for the Millon Clinical Multiaxial Inventory II (MCMI-II)*. Minneapolis, MN: National Computer Systems.
- Millon, T. (1990). *Toward a new personality: An evolutionary model*. New York: Wiley.
- Millon, T., & Everly, G. S., Jr. (1985). *Personality and its disorders: A biosocial learning approach*. New York: Wiley.
- Perry, J. C. (1991). *Problems and considerations in the valid assessment of personality disorders*. Cambridge, MA: Cambridge Hospital & Harvard Medical School.
- Pfohl, B., Stangl, D.A., & Zimmerman, M. (1982). *Structured Interview for DSM-III personality disorders (SIDP)*. Iowa City, IA: University of Iowa.
- Soldz, S., Budman, S., Demby, A., & Merry, J. (1993). Representation of personality disorders in circumplex and five-factor space: Explorations with a clinical sample. *Psychological Assessment*, *5*, 41-52.
- Spitzer, R. L. (1983). Psychiatric diagnosis: Are clinicians still necessary? *Comprehensive Psychiatry*, *24*, 399-411.
- Spitzer, R. L., Williams, J. B., & Gibbon, M. (1987). *Structured clinical interview for the DSM-III-R personality disorders (SCID-II)*. New York, NY: Biometrics Research Department, New York State Psychiatric Institute.
- Streiner, D. L., & Miller, H. R. (1989). The MCMI-II: How much better than the MCMI? *Journal of Personality Assessment*, *53*, 81-84.
- Widiger, T. A. (1991). Personality disorder dimensional models proposed for DSM-IV. *Journal of Personality Disorders*, *5*, 386-398.
- Wiggins, J., & Pincus, A. (1989). Conceptions of personality disorders and dimensions of personality. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, *1*, 305-316.

Stephen Soldz
Mental Health Research Program
Harvard Community Health Plan
10 Brookline Place West
Brookline, MA 02146

Received April 29, 1992
Revised July 22, 1992