JOURNAL OF PERSONALITY ASSESSMENT, *75*(1), 46–81

ARTICLES

On the Science of Rorschach Research

Gregory J. Meyer
Department of Psychology
University of Alaska Anchorage

I describe problems in an article by Wood, Nezworski, Stejskal, Garven, and West (1999b). These include (a) claims that researchers found or said things they did not, (b) an assertion that my data did not support the incremental validity of the Rorschach over the MMPI–2 when the opposite was true, (c) complications with their recommended incremental validity procedures, (d) unwarranted criticism of Burns and Viglione's (1996) statistical analyses, (e) oversimplifying issues associated with extreme groups research, (f) misleading criticisms of composite measures, and (g) faulty criticisms of Burns and Viglione's composite scale that overlooked relevant evidence. Wood et al. also asserted that Burns and Viglione's primary Rorschach variable was faulty and created a formula that seemed to show how Burns and Viglione's scores were "incompatible" and "not … even very close" to those obtained from the proper formula. These criticisms were made even though Wood et al. had been told that their formula was incorrect and shown that it was almost perfectly correlated with the proper formula in 8 large samples (rs > .998). Sound criticism of Rorschach research will advance science and practice, but the Wood et al. article did not provide sufficient guidance.

In the process of achieving useful scientific knowledge, it is very common for researchers to hold opposing viewpoints, dispute one another's findings, and criticize each other's methodology. Some recent examples drawn from the assessment literature include debates on the clinical utility of the Rorschach (Meyer, 1999c), the integration of the Rorschach and the Minnesota Multiphasic Personality Inventory (MMPI; Hathaway & McKinley, 1951; e.g., Archer, 1996; Ganellen, 1996b), similarities and differences between the MMPI and the MMPI–2 (Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989; e.g., Dahlstrom, 1996; Tellegen & Ben-Porath, 1996), and the utility of normal personality assessment measures for clinical practice (e.g., Ben-Porath & Waller, 1992; Costa & McCrae, 1992). Thoughtful critiques that bring critical issues into focused relief or appropriately warn about the dangers associated with particular methodological or statistical designs are embraced because they, ultimately, advance genuine knowledge.

Recently, Wood et al. (1999b) critiqued three Rorschach studies published in 1996 (Burns & Viglione, 1996; Ganellen, 1996a; Weiner, 1996). Wood et al.'s critique has merit on a number of points. For instance, studying extreme groups does lead to larger than normal effect sizes, appropriate control groups are important for any study that wishes to shed light on an experimental group, and diagnostic efficiency statistics drawn from studies that have been conducted by many investigators across numerous settings should be given more credence than those drawn from a single investigator's work.

Although these points are sound, Wood et al.'s (1999b) article also contained many inaccurate and misleading statements. Most troubling, there is reason to believe that Wood et al. knew some of their assertions were incorrect and misleading even before they submitted the article for publication.

Raising the latter is not something I do lightly. James Wood has sharpened my thinking on a number of issues and has made valuable contributions to my own research (see Meyer, 1997b). In addition, the critiques that he and his colleagues have published on the Rorschach (e.g., Nezworski & Wood, 1995; Wood, Nezworski, & Stejskal, 1996a, 1997) have, in my view, led to a heightened awareness of certain methodological issues and have spurred authors to conduct sound research that disputes the criticism (e.g., Hilsenroth, Fowler, Padawer, & Handler, 1997; Meyer, 1997a, 1997c). Nonetheless, because of the seriousness of the issues and because available research indicates published retractions have little to no impact on decreasing the frequency with which an originally problematic article gets cited (e.g., Whitely, Rennie, & Hafner, 1994), this article details some of the salient problems found in Wood et al. (1999b).

Wood et al. (1999b) devoted the majority of their article to criticizing the study by Burns and Viglione (1996). Before addressing issues that relate to Burns and Viglione, I briefly consider points raised about Weiner (1996) and Ganellen (1996a) and then discuss and correct several improper citations from the literature.

ISSUES RELATED TO WEINER (1996) AND GANELLEN (1996a)

Wood et al. (1999b) criticized one point in a lengthy article by Weiner (1996). Specifically, they faulted Weiner for a logical argument. Weiner noted that three samples of war veterans diagnosed with posttraumatic stress disorder had Rorschach scores that differed from normative values in a theoretically expected manner. Wood et al. maintained that Weiner's logic was problematic because the three studies did not collect their own control groups, and thus, the logical comparison with normative data may have been confounded by other factors. This may be true, and it certainly would have been optimal if each of the original studies had been able to solicit, schedule, test, score, and analyze Rorschach findings from their own nonpatient groups. However, doing so essentially doubles the expense of a study and may not always be feasible to accomplish in the early stages of research. For instance, the technical manual (Psychological Corporation, 1997) for the third editions of the Wechsler Adult Intelligence Scale and Wechsler Memory Scale presented validity data for each test. The manual reported data for 13 criterion samples (e.g., Alzheimer's disease and learning disabilities). None of the clinical samples were accompanied by their own control group. Instead, in every instance, the manual provided the same logic as Weiner: If the clinical samples have scale scores that differ from the normative sample in theoretically expected ways, the differences support the validity of the test scales. Although this kind of evidence does not provide airtight validity, it is still meaningful and much better than no evidence at all.

With respect to Ganellen (1996a), Wood et al.'s (1999b) main purpose was to expound on limitations in his database. However, Ganellen himself repeatedly articulated the same limitations. For instance, there are 17 pages of text in Ganellen's article. Of these, 5½ pages are devoted to discussing limitations with the data, and 4½ pages deal with general issues that could be applied to all the tests, whereas 1 full page was specifically devoted to limitations in the Rorschach evidence base. In addition, there were at least 11 other places in the manuscript in which Ganellen referred to his findings as "preliminary," affected by "methodological limitations," or "tentative." Thus, although Wood et al. (1999b, p. 125) said that Ganellen did not seem to consider the limitations in his review to be as significant as they did, it is hard to imagine how Ganellen could have been more clear on this issue.

ERRONEOUS CITATIONS

Wood et al. (1999b) inaccurately cited two articles of mine. In the process of addressing faulty citations to my work, I identify a number of other problematic references in Wood et al.'s article.

In the first erroneous citation of my work, Wood et al. (1999b, p. 125) stated that Ganellen's (1996a) "problematic conclusions about the DEPI [Rorschach Depression Index] … have been cited as justification for using the Rorschach in forensic and other contexts (McCann, 1998; Meyer et al., 1998)." Although the Meyer et al. report did cite Ganellen's (1996a) article, Ganellen's research was only cited with respect to the MMPI and Millon Clinical Multiaxial Inventory (see Meyer et al., pp. 24–25) because these were tests for which Ganellen provided a review of the published literature. James Wood (personal communication, June 1, 1999) apologized for this mistake and published a correction (Wood, Nezworski, Stejskal, Garven, &

West, 1999a). Although this is the optimal corrective action to take, such a specific misstatement does not inspire confidence.

Wood et al.'s (1999b) citation of McCann (1998) in the quote given previously was also inappropriate. In a single sentence, McCann mentioned two classification rates that Ganellen had reported for the *DEPI*. However, in the following sentence, McCann indicated that variations in Rorschach response frequency can confound the Rorschach's classification accuracy. Next, McCann stated:

> Wood et al. (1996a, 1996b) pointed out that the DEPI has shown rather poor diagnostic power in cross-validation studies and falls prone to what is termed shrinkage during cross-validation. The results of independent studies have shown that the DEPI does not have a strong relation with self-report measures of depression (Ball, Archer, Gordon, & French, 1991; Meyer, 1993). Moreover, the Rorschach indexes need to be investigated further in independent research. (p. 137)

Taken in its full context, it is hard to see how McCann's statements provided a justification for using the Rorschach in forensic practice, as Wood et al. (1999b) suggested. Rather, McCann provided a brief overview of DEPI evidence in the literature, both positive and negative. Thus, even though Wood et al. cited McCann and Meyer et al. (1998) as a way to criticize Ganellen (1996a), neither citation actually supported Wood et al.'s criticism.

In a second set of inaccurate citations, Wood et al. (1999b) stated:

> Studies that have compared the first and second versions of the SCZI [Rorschach Schizophrenia Index] with the MMPI (Archer & Gordon, 1988; Meyer, 1993) have found that the SCZI does not add incremental validity to the prediction of schizophrenia diagnoses, beyond what can be obtained using the MMPI. (p. 125)

Both of the studies cited in the previous quote are problematic, and each is discussed in turn.

It was inaaccurate for Wood et al. (1999b) to cite Meyer (1993) because my study neither examined incremental validity nor even mentioned the topic. Although James Wood (personal communication, June 1, 1999) apologized for this mistake and published a correction (Wood et al., 1999a), this is another direct misattribution that could have been avoided by conscientious effort to portray the Rorschach literature accurately.

It was also misleading for Wood et al. (1999b) to cite Archer and Gordon's (1988) study in the way they did. Archer and Gordon did not present detailed findings on the combined use of Rorschach and MMPI scores. They briefly presented information about a discriminant function analysis that combined Scale *8* from the MMPI and the Schizophrenia Index from the Rorschach. However, this appeared in the Discussion section of their article and not in the Results section. Furthermore, when presenting this information, Archer and Gordon did not indicate whether both, either, or neither of these variables reached statistical significance in their multivariate equation. As such, they never provided a formal test of the Rorschach's ability to add incrementally to the classification of psychotic diagnoses over the MMPI.

After examining the results in their sample, Archer and Gordon (1988) selected a nontraditional MMPI cutoff score (*T* score ≥ 75) that would maximize diagnostic prediction for Scale *8* of the MMPI in that particular sample. Against this baseline, it appears that the Schizophrenia Index may not have contributed additional statistically significant unique information to diagnostic prediction (although see the following). However, by consulting the sample results to select an optimal cutoff for the MMPI scale, the design did not provide a uniform test of incremental validity. A uniform test could have been obtained if both scales were evaluated blindly on their own merits.

Several other features of Archer and Gordon's (1988) study deserve attention because they suggest that the *SCZI* actually outperformed the MMPI. First, the a priori planned comparisons reported in Table 1 of Archer and Gordon's article clearly indicated that Scale *8* from the MMPI did not differ significantly across diagnostic groups, although the *SCZI* did discriminate in an expected pattern. Second, although Archer and Gordon concluded that Scale *8* had "a slightly better overall hit rate among adolescents than the SCZI" (p. 284), their data did not bear out this conclusion. From their Table 2, it can be seen that the optimal overall hit rate for the *SCZI* was .80 (using a cutoff of 5), whereas the optimal overall hit rate for Scale *8* was .76 (using a cutoff ≥ 75). Third, when using traditional cutoffs for the Rorschach and the MMPI, the *SCZI* showed better overall classification accuracy than Scale *8*. Specifically, the *SCZI* had a hit rate of .69 using the traditional cutoff of greater than or equal to 4, whereas Scale *8* had hit rates of .48 and .60 using the more traditional cutoffs of *T* ≥ 65 and *T* ≥ 70, respectively. Finally, in their discussion, Archer and Gordon computed two multivariate models that combined the MMPI and Rorschach. These equations had classification accuracy rates of .60 and .73. The authors noted how these results did not improve upon the optimal hit rate of .76 that was obtained from using Scale *8* alone. However, if overall classification accuracy was the criterion for determining test adequacy, any univariate or multivariate equation that used the MMPI actually performed worse than the *SCZI* on its own. Recall that the *SCZI* had an optimal hit rate of .80, which exceeded any value found for the MMPI.

To summarize, (a) Archer and Gordon (1988) never provided a formal test of the Rorschach's ability to add incrementally to the classification of schizophrenic diagnoses over the MMPI; (b) in preplanned univariate tests, the *SCZI* outperformed the MMPI Scale *8;* (c) in optimal univariate classification accuracy, the *SCZI* outperformed Scale *8*; (d) in analyses using traditional scale cutoffs to determine classification accuracy, the *SCZI* outperformed Scale *8;* and (e) in terms of optimal classification, the *SCZI* by itself outperformed all multivariate or univariate equations that used Scale *8*. Thus, in contrast to Wood et al.'s (1999b) assertion, Archer and Gordon's results are silent on the issue of the statistically significant incremental contribution of the *SCZI* to diagnostic classification. However, at a minimum, their findings indicate that the *SCZI* is a better univariate predictor than Scale *8*. Given that the *SCZI* was superior to Scale *8* in every head-to-head comparison, Wood et al.'s conclusion that the opposite was true suggests either a lack of attention to the facts or a propensity to hold the Rorschach to a different and more demanding standard of evidence than the MMPI.

FURTHER EVIDENCE ON THE INCREMENTAL VALIDITY OF THE RORSCHACH OVER THE MMPI–2

Wood et al.'s (1999b) citation of Meyer (1993) indicated there would not be evidence of incremental validity for the *SCZI* in my data. Because diagnostic categories form crude criteria that are often not tied to specific behaviors (Persons, 1986), I have not been disposed to explore this issue in detail. However, now that the question has been raised, a failure to address it directly with evidence may leave some readers assuming that my data must implicitly support Wood et al.'s assertion. To prevent this from happening, relevant findings are presented next.

Over a 4-year period, I collected Rorschach and MMPI–2 data from 362 patients at the University of Chicago Medical Center. A description of this sample can be found in other reports (Meyer, 1993, 1997b, 1999a). For 265 of these patients, I was able to obtain diagnoses that had been assigned for billing purposes by outpatient therapists or inpatient treatment teams. Although these diagnoses cannot be considered a gold standard criterion because (a) unwanted sources of error can affect billing records and (b) patients were frequently referred for testing to clarify diagnostic uncertainties, the diagnoses were fully independent of the testing results and can serve as an approximate criterion—one that should underestimate the true validity of test scales. Research assistants and support personnel obtained diagnoses from the computerized billing records of the hospital. For the diagnoses to be recorded, they had to have a posting date that was prior to the onset of my testing evaluation and, hence, independent of it (the people who obtained the diagnoses were only given the patient's name, medical number, and date of referral). Three of the 265 patients with diagnostic information did not have scores available for the MMPI–2 Bizarre Mentation scale. Because Bizarre Mentation is an important variable to include in any regression analyses designed to predict psychotic disturbances (Ben-Porath, Butcher, & Graham, 1991), analyses were based on a final sample of 262 patients.

Patients were classified into two different sets of diagnostic criterion groups. The first was a narrow classification contrasting patients with schizophrenia, schizophreniform disorder, or schizoaffective disorder (n = 30; i.e., any *Diagnostic and Statistical Manual of Mental Disorders* code in the 295.xx series) to patients with all other diagnoses (n = 232). The second was a broad classification that compared patients with any psychotic disorder (n = 120) to those with all other diagnoses (n = 142). Psychotic disorders included schizophrenia, affective disorder with psychotic features, delusional disorder, brief psychotic disorder, psychotic disorder not otherwise specified, schizotypal personality disorder, or borderline personality disorder. This classification corresponded to the criterion used in the Meyer (1993) study cited by Wood et al. (1999b). With respect to the narrow diagnostic category of schizophrenia, one patient received a diagnosis of residual schizophrenia. Excluding this patient did not materially alter the findings, so results are reported for all 30 schizophrenia patients.

The analyses used three theoretically derived predictors: Scale *8* (Schizophrenia) and Bizarre Mentation from the MMPI–2 and the *SCZI*. Hierarchical linear regression[1] was used with stepwise forward entry and backward removal of variables within blocks. Stepwise analysis is an iterative procedure and within blocks the regression equation is built sequentially according to the specified criteria. *Forward variable entry* means the most significant predictor enters the regression equation first, followed by the next most significant predictor after controlling for the scale (or scales) already in the equation. This iterative process terminates when there are no other significant predictors to enter the equation. If more than one variable has entered the equation within a block, a variable entered on an earlier step may no longer be a significant predictor because its variance may have been subsumed by variables entered later. Consequently, the removal criterion evaluates the equation after each step and allows for the backward elimination of variables if they are no longer statistically significant. Statistical modeling has demonstrated that the traditional alpha level of .05 is inadequate when building stepwise regression models because it often excludes important variables from the final equation. Consequently, experts recommend that the alpha level for a variable to enter a regression equation be set between $p = .15$ and $p = .20$ (see Hosmer & Lemeshow, 1989). However, in large samples like this one, even relatively small contributions can reach alpha levels in the range between .15 to .20. To simultaneously take into account the research on optimal alpha levels while being mindful of the relatively large sample for this analyses, I set alpha at .10 for variables to enter the regression equation and .15 for variables to be removed.

Regression analyses most often report findings using squared correlational values. However, because r^2 (or R^2) is a poor index for gauging the magnitude of an effect, prominent statisticians recommend instead that emphasis be placed on nonsquared correlational values (i.e., *r* or *R;* see Cohen, 1988; Hunter & Schmidt, 1990; Rosenthal, 1991). I follow the latter recommendations when discussing the importance or impact of a variable, although both sets of values are still reported.

For my analyses, there were two data entry blocks in the hierarchy. The first block evaluated both MMPI–2 scores according to the stepwise criteria. The sec-

[1] Although multiple linear regression is familiar to many and is a standard procedure used in incremental validity analyses, because the diagnostic criterion is dichotomous, it would be more appropriate to use logistic regression analyses. Equivalent results were obtained when logistic regression was used instead of linear regression. For familiarity and ease of communication, only the linear regression results are presented.

ond block evaluated the *SCZI* according to the stepwise criteria. With this design, the *SCZI* was only able to enter the regression equation after the MMPI–2 variables had been considered and only if the *SCZI* made a statistically significant incremental contribution to diagnostic classification beyond that which could be obtained from the MMPI–2. The analyses were conducted twice: once using MMPI–2 raw scores and once using *T* scores. The outcomes were virtually identical, so only the raw score results are reported.

Table 1 presents findings for both the narrow and broad diagnostic criteria. In Block 1 of each analysis, Bizarre Mentation from the MMPI–2 significantly predicted the diagnostic criterion. Scale *8* added no additional self-reported information beyond that available from Bizarre Mentation. In Block 2, the *SCZI* contributed to diagnostic prediction over and above that which could be obtained from the MMPI–2 ($p < .01$ in both analyses). This clear evidence of Rorschach incremental validity was demonstrated using both sets of diagnostic criteria. Thus, in contrast to Wood et al.'s (1999b) assertion, my data unambiguously reveal how the Rorschach provides useful clinical knowledge that cannot be gleaned from the MMPI–2. This is true when predicting the relatively narrow spectrum of schizophrenic disorders, and it is also true when predicting the broader category of psychotic disorders.

Given the nature of the questions asked on the MMPI and the nature of the overt behaviors quantified by the Rorschach, the Bizarre Mentation Scale and the *SCZI* both quantify psychotic characteristics that may be manifest across a variety of diagnostic categories. Hallucinations, thought disorder, and perceptual distortions are not restricted to schizophrenia. Consequently, both the MMPI and the *SCZI* produced their strongest validity coefficients in relation to the "any psychotic disorder" criterion. With this criterion, when the *SCZI* was added to the MMPI–2, the magnitude of the validity coefficient increased from $R = .48$ to $R = .49$, indicating the *SCZI* made a substantial and unique contribution to this clinical task.

Although Wood et al. (1999b) did not comment on the incremental validity of the DEPI in my data set, this issue has been addressed before in adolescent samples (Archer & Gordon, 1988; Archer & Krishnamurthy, 1997). For the sake of completeness, I ran an incremental validity analysis to predict depressive disorders. Patients were classified as having a depressive disorder ($n = 212$) if they had a diagnosis of major depression, bipolar disorder with mixed or depressed features, dysthymia, or depressive disorder not otherwise specified. The remaining patients were classified as having a nondepressive disorder ($n = 50$). The hierarchical analysis again had two data entry blocks. The first evaluated two MMPI–2 variables, Scale *2*, and the Depression content scale, according to the stepwise criteria. The second block evaluated the *DEPI* according to the stepwise criteria. As before, the Depression Index was only able to enter the regression equation after the MMPI–2 variables had been considered and only if it made a statistically significant incremental contribution to diagnostic classification beyond that which could be obtained from the MMPI–2. The analysis only examined MMPI–2 raw scores.

TABLE 1
Incremental Validity of the Rorschach Schizophrenia Index (*SCZI*) Over Scale *8* and Bizarre Mentation (*BIZ*) From the MMPI–2 for Predicting Schizophrenic Disorders and Any Psychotic Disorders: Hierarchical Regression Using Stepwise Entry Within Blocks

Criterion and Block	*Step*	*Variable Coefficients at Block Termination*	*B*	*SE B*	β	R^2	Δ R^2	*R*	Δ *R*
Schizophrenic disorders									
Block 1	1	*BIZ*	.0198	.0043	.2740*	.0751*	.0751*	.2740*	.2740*
		(Scale *8*)[a]	—	—	—				
Block 2		*BIZ*	.0193	.0043	.2662*				
	2	*SCZI*	.0291	.0107	.1612*	.1010*	.0259*	.3178*	.1610*
Any psychotic disorders									
Block 1	1	*BIZ*	.0423	.0065	.3735*	.1395*	.1395*	.3735*	.3735*
		(Scale *8*)[a]	—	—	—				
Block 2		*BIZ*	.0405	.0061	.3593*				
	2	*SCZI*	.0890	.0153	.3146*	.2382*	.0987*	.4881*	.3142*

Note. *N* = 262. MMPI–2 = Minnesota Multiphasic Personality Inventory–2.
[a] Variable was considered, but it did not enter the equation.
*$p < .01$.

TABLE 2
Incremental Validity of the Rorschach Depression Index (*DEPI*) Over Scale *2* and Depression (*DEP*) From the MMPI–2 for Predicting Depressive Disorders: Hierarchical Regression Using Stepwise Entry Within Blocks

Block	Step	*Variable Coefficients at Block Termination*	*B*	*SE B*	β	R^2	ΔR^2	R	ΔR
1	1	*DEP*	.0095	.0041	.2051**	.1013**	.1013**	.3183**	.3183**
	2	Scale *2*	.0076	.0045	.1505*	.1111**	.0098*	.3334**	.0992*
2		*DEP*	.0084	.0041	.1826**				
		Scale *2*	.0089	.0045	.1771**				
	3	*DEPI*	.0353	.0172	.1209**	.1254**	.0143**	.3542**	.1196**

Note. $N = 262$. MMPI–2 = Minnesota Multiphasic Personality Inventory–2.
*$p < .10$. **$p < .05$.

Table 2 presents results from this analysis. On Block 1, both MMPI–2 variables entered the regression equation, although the contribution from Scale *2* was less important ($\Delta R = .10$, $p < .10$) than from *DEP* ($\Delta R = .32$, $p < .001$). On Block 2, the Rorschach DEPI entered the regression equation ($\Delta R = .12$, $p < .05$), indicating that it contributed meaningful information to the prediction of depressive disorders over that which could be obtained from the MMPI–2. Although the latter is important validation data, it should be recognized that the contribution from the *DEPI* was modest.

GENERAL CONSIDERATION FOR INCREMENTAL VALIDITY ANALYSES

Wood et al. (1999b) present just one definition of incremental validity (see Meyer, 1999b, for a brief overview of alternative definitions), and they do not favor the analyses presented in Tables 1 and 2. Specifically, they do not believe that researchers should use stepwise procedures (using either forward entry or backward elimination) when building a regression model to test incremental validity. Wood et al. stated:

> As Cohen and Cohen (1983, pp. 120–125) explain, a hierarchical selection procedure is appropriate to test a hypothesis … regarding incremental validity. In a hierarchical procedure, the researcher (a) specifies beforehand that variables will be forced into the regression equation in discrete stages in a particular order, and (b) predicts that when a particular variable (or set of variables) is forced into the equation, the predictive power of the equation will incrementally and significantly improve over the previous stage. (p. 121)

Furthermore, Wood et al. asserted that (a) “stepwise and hierarchical selection procedures have entirely different purposes and interpretations” (p. 121), (b) experts on regression describe hierarchical and stepwise procedures “as two separate approaches with completely different uses” (p. 122), (c) stepwise selection procedures are used “primarily for data reduction” and “are not appropriate for … hypothesis testing” (p. 121), and (d) the approach used by Burns and Viglione (1996), which combined hierarchical and stepwise techniques, reflected “an unusual combination of two different procedures” (p. 122).

Thus, Wood et al. (1999b) maintained that the proper way to conduct an incremental validity analysis is to use hierarchical analysis in which all variables in an initial block (or blocks) are forced into the equation prior to evaluating the statistically significant incremental contribution of a key variable on the final block. In terms of the analyses reported in Table 1, Wood et al.’s position was that Scale *8* should have been forced into the regression equation on Block 1, although it did not have a statistically significant association with the criterion after Bizarre Mentation had entered the equation. According to Wood et al. , only after Scale *8* and Bizarre Mentation had both been forced into the equation would it be appropriate to consider whether the *SCZI* made a unique contribution to diagnostic prediction.

Note how in this formulation the MMPI scales would not be held to a standard of statistical significance, yet the *SCZI* would be. Undoubtedly, this is a strong and exacting test for demonstrating the incremental validity of a Rorschach score because all the sampling error that is associated with variables used on the initial block gets forced into the equation, and only the Rorschach score must demonstrate a statistically significant level of prediction. Although this alone can be problematic, there are several other difficulties with Wood et al.’s (1999b) claim that this is the only proper way to conduct an incremental validity analysis.

First, even though Wood et al. (1999b) invoked pages 120 through 125 of Cohen and Cohen’s (1983) text on multiple regression to support their position, Cohen and Cohen are not as clear-cut on this point as Wood et al. maintained. For instance, after describing the basic format of forward or backward stepwise procedures for multiple-regression/correlation analysis, Cohen and Cohen indicated how stepwise algorithms proceed until no other independent variables (IVs) have a statistically significant contribution. Cohen and Cohen went on to state:

> An investigator may be clear that some groups of variables are logically, causally, or structurally prior to others, and yet not have a basis of ordering variables within such groups. Under such conditions, variables may be labeled for entering in the equation as one of the first, second, or up to *h*th group [i.e., block] of variables, and the sequence of variables within each group is determined by the computer in the usual stepwise manner [i.e., by statistical significance]. This type of analysis is likely to be primarily hierarchical (between classes of IVs) and only incidentally stepwise (within classes), and computer programs so organized may be effectively used to accomplish hierarchical MRC [multiple-regression/correlation analysis] by sets of IVs. (p. 124)

Given that Wood et al. specifically cited this section of Cohen and Cohen's text, it is surprising that this point about the combined use of hierarchical and stepwise procedures was overlooked. As the quote makes clear, Cohen and Cohen did not see hierarchical and stepwise procedures as incompatible, nor did they see their combination as something "unusual" like Wood et al. maintain. Instead, Cohen and Cohen viewed this combined approach, in which hierarchical blocks were specified a priori and statistical significance was used to determine the entry of variables within a block, as a legitimate way to test incremental validity.

A second problem with Wood et al.'s (1999b) recommended incremental validity procedures is its exclusive reliance on statistical significance as the arbiter that determines success. This overlooks critical issues related to statistical power and effect size. Recently, there have been serious discussions in psychology focused on "banning" statistical significance tests (e.g., Shrout, 1997; Wilkinson & Task Force on Statistical Inference, 1999). The basic problem is that trivial effects (e.g., $r = .001$) will be "significant" in sufficiently large samples whereas important effects (e.g., $r = .30$) will be "insignificant" if the sample is rather small (see Cohen, 1994). When applied to regression analyses, this means it will be much easier to demonstrate "significant" incremental validity in a large sample than in a small sample, although such a demonstration may have nothing to do with the practical importance of the variable under consideration.

Third, Wood et al.'s (1999b) insistence that incremental validity analyses should only be conducted using forced entry of variables on the initial step is countered by what is typically done in the published literature. Although some investigators do force variables into a regression equation on an initial block regardless of statistical significance (e.g., Barthlow, Graham, Ben-Porath, & McNulty, 1999; Schinka, LaLone, & Greene, 1998), most do not. The following personality assessment studies have all recently examined questions of incremental validity using the "unusual combination" of hierarchical and stepwise procedures: Archer, Aiduk, Griffin, and Elkins (1996); Archer, Elkins, Aiduk, and Griffin (1997); Ben-Porath et al., (1991); Ben-Porath, McCully, and Almagor (1993); Munley, Busby, and Jaynes (1997); Paunonen (1998); and Wetzler, Khadivi, and Moser (1998).

Finally, the forced entry of many variables can produce misleading conclusions. Cohen and Cohen (1983) explained that when many IVs are used in a block, "the partialling process highlights their uniqueness, tends to dissipate whatever common factor they share, may produce paradoxical suppression effects, and is thus also likely to create severe difficulties in interpretation" (p. 171). Similarly, when discussing forced entry, Hosmer and Lemeshow (1989) stated, "The major problem with this approach is that the model may be overfitted and produce numerically unstable estimates" (p. 83). These points are explicated by returning to the question addressed in Table 1. Even though Bizarre Mentation and Scale *8* are the most pertinent MMPI–2 variables to use when predicting a psychotic disorder criterion, a researcher could try to discredit the findings in Table 1 by pointing out how the *SCZI* was only evaluated

TABLE 3
Illustrating Problems From the Forced Entry of Variables Within Blocks: Incremental Validity of the Rorschach *SCZI* Over 28 MMPI–2 Basic and Content Scales for Predicting Any Psychotic Disorder Using Hierarchical Regression and Forced Entry of Block 1 Variables

Variable in the Equation	*Block 1*[a] β	*Block 2*[b] β
L	–.002	.021
F	–.084	–.090
K	.034	.008
Scale *1*	–.061	.021
Scale *2*	.088	.147
Scale *3*	.116	.133
Scale *4*	–.029	–.079
Scale *5*	.015	–.024
Scale *6*	.205	.201
Scale *7*	–.275	–.137
Scale *8*	.335	.234
Scale *9*	–.012	.003
Scale *0*	–.162	–.216
Anxiety	–.135	–.197
Fears	.238**	.220**
Obsessiveness	–.133	–.162
Depression	.207	.171
Health Concerns	–.264	–.303
Bizarre Mentation	.206	.233*
Anger	–.108	–.068
Cynicism	.327*	.322*
Antisocial Practices	–.158	–.108
Type A	.059	–.025
Low Self-Esteem	.148	.176
Social Discomfort	.030	.080
Family Problems	.096	.067
Work Interference	.126	.121
Negative Treatment Indicators	–.184	–.153
SCZI		.306***

Note. $N = 262$. MMPI–2 = Minnesota Multiphasic Personality Inventory–2. Because of suppressor effects, some variables are not statitically significant predictors even though they have larger beta values than do variables that are significant predictors.

[a]$R^2 = .2531$***; $\Delta R^2 = .2531$***; $R = .5031$***; $\Delta R = .5031$***. [b]$R^2 = .3337$***; $\Delta R^2 = .0806$***; $R = .5777$***; $\Delta R = .2839$***.

*$p < .05$. **$p < .01$. ***$p < .0001$.

against two of the many MMPI–2 scales that are available. As such, one could suppose that the *SCZI* would not demonstrate incremental validity if a larger number of scales from the MMPI–2 had been considered. This question will be addressed by examining the 13 Basic scales and 15 Content scales from the MMPI–2.

Following the recommendations of Wood et al. (1999b), all 28 MMPI–2 scales were forced into the regression equation on Block 1 of the analysis, regardless of their statistical significance and regardless of whether the analysis produced conceptually meaningful results. Next, on Block 2, the *SCZI* was held to the traditional statistical significance hurdle ($p < .05$), such that it had to provide a statistically significant incremental contribution to the prediction before it could enter the equation.

Table 3 presents simplified results from this analysis. From the left side of the table, it can be seen that when all 28 MMPI–2 scales were entered on Block 1, despite the beta values (some of which are inflated by suppression effects), there were only two statistically significant MMPI–2 predictors. The first was a scale for assessing specific phobic-like fears (Fears) and the second was a scale for assessing cynical attitudes (Cynicism). Both of these findings are paradoxical, particularly because the traditional scales for assessing psychotic processes (i.e., Scale *8* and Bizarre Mentation) no longer had statistically significant associations with the criterion. Despite these oddities, the right side of Table 3 indicates that the *SCZI* entered the equation on the second block because it still had statistically significant incremental validity for predicting the criterion after all 28 MMPI–2 scales had been forced into the equation ($\Delta R = .28$, $p < .0001$; $\Delta R_{\text{adjusted}} = .30$). The right side of Table 3 also shows how adding the *SCZI* into the mix of predictors altered the relative contribution of other variables, such that Bizarre Mentation now became a statistically significant predictor. However, Bizarre Mentation remained a less powerful predictor (in terms of *p* values) than the Fears and Cynicism scales.[2]

What should be made of the results in Table 3? On the one hand, perhaps the value of this analysis is that it makes it abundantly clear how the Rorschach has incremental validity over all the primary MMPI–2 scales.[3] On the other hand, the MMPI–2 predictors are a mess; the results are not theoretically coherent because a cacophony of scales was forced into the regression equation and the final results capitalized on nonreplicable sampling error (despite the fact that this sample was

[2] Similar findings were observed when all 28 MMPI–2 scales were forced into the regression equation as predictors of the depressive disorder criterion prior to evaluating the incremental contribution made by the *DEPI*. Despite paradoxical MMPI–2 scale contributions, on Block 2 of the analysis the *DEPI* contributed statistically significant unique information that could not be obtained from the 28 MMPI–2 scales ($\Delta R = .15$, $p = .0069$; $\Delta R_{\text{adjusted}} = .15$).

[3] For those who may still think the Rorschach Schizophrenic Index (*SCZI*) would have faltered if "better" MMPI–2 variables had been considered, the analysis was also run using 113 MMPI–2 scales, which is the total number of scales I have calculated for the MMPI. The scales include all those listed in Table 3 plus all the Harris and Lingoes subscales, all the supplementary scales, all the content component scales, Ben-Porath's Si subscales, the Personality Psychopathology Five, the Goldberg Psychotic–Neurotic Index, the mean elevation of MMPI–2 basic scales, the Peterson psychosis signs, and the Taulbee-Sisson psychosis signs (for a description of scales, see Ben-Porath & Sherwood, 1993; Greene, 1991; Harkness, McNulty, & Ben-Porath, 1995). These 113 MMPI–2 scales were forced into the regression equation on the first block. On the second block, the *SCZI* was allowed entry into the equation if it made a significant and unique contribution. As before, the *SCZI* contributed significant information to the diagnostic task that could not be obtained from all 113 MMPI–2 scales ($\Delta R = .26$, $p < .0001$; $\Delta R_{\text{adjusted}} = .33$). When the same analysis was run using the depressive disorder criterion and the Rorschach Depression Index (*DEPI*) on Block 2, the incremental contribution of the *DEPI* was of marginal statistical significance ($\Delta R = .10$, $p = .0884$; $\Delta R_{\text{adjusted}} = .11$).

fairly large; *N* = 262). Although others may see the results differently, it is hard to fathom how the forced entry of all these variables presents a reasonable model of the capacity of the MMPI–2 to predict psychotic diagnoses in the real world. As such, it also does not seem reasonable to require that this be the standard against which the *SCZI* (or any other Rorschach scale) is evaluated for incremental validity.

OTHER RORSCHACH INCREMENTAL VALIDITY ISSUES

Wood et al. (1999b) cited research by Archer and Gordon (1988) and Archer and Krishnamurthy (1997) to support the notion that the Rorschach does not have incremental validity over the MMPI. For instance, Wood et al. stated: “Archer and Gordon (1988) had found that the DEPI and the Schizophrenia Index (*SCZI*) do not provide incremental validity beyond MMPI scores in prediction of diagnoses” (p. 124). Later, in describing Archer and Krishnamurthy’s study, they stated: “The new DEPI did not significantly predict diagnoses of depression or increase incremental validity beyond the predictive power of MMPI–A scores” (p. 125). Because these two studies also have been cited by other authors (e.g., Hunsley & Bailey, 1999) to support the notion that the Rorschach generally does not have incremental validity over the MMPI, it is worthwhile to clarify what these studies have and have not found.

I already discussed Archer and Gordon’s (1988) findings with respect to the *SCZI* and made two points: (a) In terms of single scales, the *SCZI* was a better diagnostic predictor than the MMPI Schizophrenia scale, and (b) it was impossible to draw direct conclusions regarding incremental validity because this information was never provided. With respect to the *DEPI*, Wood et al.’s (1999b) previous quote could leave some readers with the false impression that the MMPI validly predicted depressive diagnoses but that the *DEPI* added no additional predictive information to the MMPI. However, in Archer and Gordon’s study neither the MMPI nor the Rorschach could predict depressive diagnoses at a statistically significant level.

Also, two points can be noted about Archer and Krishnamurthy (1997). These authors did not evaluate the ability of the Rorschach and MMPI to predict schizophrenia or psychosis. Instead, they focused on predicting depressive disorders and conduct disorders. Wood et al. (1999b) accurately stated that the *DEPI* did not add incrementally to the MMPI’s ability to predict depressive diagnoses. However, what was not indicated was that Archer and Krishnamurthy found two subcomponents to the Rorschach *DEPI* (vista responses and the affective ratio) that did add to the prediction of depressive diagnoses beyond the information that could be obtained from the MMPI. Thus, when considering this study, it would be more accurate to say that even though the DEPI did not have incremental validity, the Rorschach itself contributed unique information that could not be derived from the MMPI.

Second, unlike the MMPI, the Rorschach does not have standard scales that are thought to be related to conduct disorder. Perhaps because of this, it was not surprising that Archer and Krishnamurthy (1997) found no Rorschach scores that would add to the prediction of conduct disorder over scales derived from the MMPI. At the same time, however, Archer and Krishnamurthy's conduct disorder analysis was somewhat compromised because one of the significant MMPI predictors, the Immaturity scale, was allowed into the regression equation even though it "predicted" conduct disorders in the wrong direction (see Butcher et al., 1992, for a description of the scale). This can be seen if one examines the means reported in Archer and Krishnamurthy's Table 1 or if one calculates phi or kappa coefficients from the data in their Table 4. Because the conduct disorder patients were paradoxically lower on the Immaturity scale than the remaining patients, the multivariate classification equation capitalized on nonsensical MMPI findings. Although the Rorschach still may not have fared differently, in fairness to the Rorschach the results should have been recomputed after excluding the Immaturity scale from the multivariate model.

Finally, for those who are seriously interested in questions about the Rorschach's incremental validity, it would be useful to review a broader array of evidence. Viglione (1999) reviewed a number of incremental validity studies from the past 20 years, and Meyer (1999b) provided a focused review and meta-analysis of the incremental validity of the Rorschach Prognostic Rating Scale over self-reported mental health and measured intelligence.

BURNS AND VIGLIONE'S (1996) REGRESSION ANALYSES

Wood et al. (1999b, p. 122) criticized Burns and Viglione (1996) for using relaxed alpha levels ($p < .11$ and $p < .19$) to build two of their three regression models. Although this criticism may seem reasonable to some because it is based upon historical guidelines, it overlooks several relevant facts. First, even though Wood et al. (1999b, p. 121) specifically cited page 106 of Hosmer and Lemeshow's (1989) well-respected text on logistic regression, they did not attend to the material discussed two pages later. Based on relevant research, Hosmer and Lemeshow instruct researchers who are building stepwise regression equations to avoid the traditional .05 alpha level. Instead, they recommend that researchers set the probability level for variables to enter the equation in the range between .15 and .20. This is the range that was used by Burns and Viglione. Unfortunately, because Burns and Viglione did not mention this rationale in their article, they left themselves open to unwarranted criticism.[4]

Second, building an appropriate regression equation is a complicated, multistep process. Hosmer and Lemeshow (1989) indicated that researchers must not only make many decisions regarding how variables should enter an equation, but they ultimately must also examine the adequacy of the resulting model to see how well it fits the original data and discriminates the two targeted criterion groups. About half of the information in Burns and Viglione's (1996) Results section focused on the latter issues. The data they presented demonstrated the value of the Rorschach Human Experience Variable (*HEV*) for maximizing the practical importance and accuracy of the regression model.

Finally, although one could conceivably debate some of the fine points related to Burns and Viglione's (1996) analysis, certain facts remain fixed. The final step in each of their regression equations ultimately indicates the results that would emerge if all the salient predictors and covariates had been forced into the equation and then evaluated for retention based on the backwards elimination of noncontributing variables. In every analysis, the results indicate that the *HEV* was a critically important variable for predicting interpersonal functioning. Furthermore, in every analysis, the results indicate that the *HEV* was a more important predictor than alternative Rorschach or non-Rorschach predictor variables. Thus, contrary to what Wood et al.'s (1999b) criticisms might appear to suggest, it is indisputable that the *HEV* was an important predictor of interpersonal competence.

EXTREME GROUPS

Wood et al. (1999b) devoted more than 15% of their article to a discussion of extreme group designs. They correctly noted how research strategies that only examine the extreme ends of some continuum produce larger than normal effect sizes. However, their comments on this topic did not address the equally problematic factors that cause effect sizes to be smaller than normal (see Meyer & Handler, 1997, or Hunter & Schmidt, 1990, for a discussion of various factors that impact effect size magnitude).

Also, Wood et al. (1999b) closed their article by asserting that qualms about extreme groups designs "do not apply to studies in which group membership is based on diagnostic categories (e.g., schizophrenics vs. non-schizophrenics, Alzhei-

[4] If Wood et al. (1999b) were troubled by analyses that used higher alpha levels when building multivariate models, one would expect them to criticize all studies that use such procedures. However, Wood et al. (1999b; e.g., p. 125) touted the findings by Archer and Krishnamurthy (1997), despite the fact that Archer and Krishnamurthy also relied on a higher alpha level ($p < .15$) when building their multivariate equations. This is another instance that suggests Wood et al. may hold positive Rorschach evidence to a more demanding standard than positive MMPI evidence.

mer's patients vs. normal elderly)" (p. 125). This statement is potentially quite misleading. Because diagnostic criterion groups are used regularly to validate psychological tests, it is worthwhile to consider this issue in some detail.

In general, any factors that produce larger than normal variance in the distribution of criterion scores produces a form of extreme group design. Thus, if one compares patients with a diagnosis of Alzheimer's disease to a group of normal elderly who are selected to ensure they have no more than a limited number of memory complaints, then one has created an extreme groups design because there is a gap in the underlying distribution of criterion scores (i.e., in memory problems). This gap produces increased variance in the diagnostic criterion.[5]

Extreme groups also can be created in even more subtle ways. For instance, Alzheimer's affects about 2 to 4% of the population over age 65 (American Psychiatric Association, 1994). Thus, about 3 in 100 people over this age have the disease. If one selected 30 patients diagnosed with Alzheimer's from a geriatric clinic that had this population base rate and then compared these patients to a random sample of 30 other patients drawn from the same clinic, the researcher would have artificially increased the base rate of Alzheimer's in the study from 3 to 50%. Because variance for a dichotomous variable is just a function of the base rate (i.e., variance = P[1 – P], where P is the base rate) and because variance reaches its maximum when the base rate is 50%, by selecting 30 patients with Alzheimer's and 30 without, the researcher has artificially and dramatically increased the variance in Alzheimer's diagnoses for this study. Doing so produces larger than normal effect sizes (see Cohen & Cohen, 1983; Lijmer et al., 1999).

To exemplify this process, consider Christensen, Hadzi-Pavlovic, and Jacomb's (1991) meta-analysis on the ability of neuropsychological tests to differentiate patients with dementia from normal controls. Christensen et al. did not describe the procedures that were used to select normal controls in the primary studies they reviewed, and they also did not report the base rate of dementia in these studies. Consequently, it is impossible to determine how discontinuities in the underlying distribution of cognitive functioning (e.g., from comparing a group of patients with severe Alzheimer's symptoms to a group of normal controls with no symptoms) or how the artificial equating of patient and control base rates may have influenced the results. Nonetheless, Meyer et al. (1998, p. 24) indicated the average effect size from this meta-analysis was r = .68 if one assumed an equal proportion of patients and controls (i.e., if one assumed the dementia base rate was .50). In contrast, if one assumed a dementia base rate of

[5] This can be visualized by imagining a bell-shaped distribution of memory abilities. Patients with Alzheimer's disease will cluster in one tail of the distribution whereas nonpatients who have no salient memory complaints will be distributed towards the other tail of the distribution. When a study compares these two sets of people, the design effectively reduces the density of those who fall in between these groups. Thus, the study uses an extreme groups design.

10%, the average effect size would drop to r = .49. If one was working in a screening context where the base rate of dementia was just 3%, the average validity of neuropsychological tests to differentiate patients with dementia from normal controls would drop to r = .30. Obviously, a validity coefficient of r = .30 is very different from a validity coefficient of r = .68, although r = .30 is probably the more generalizable coefficient for characterizing the validity of neuropsychological tests in this context.

For an example of this phenomena in the area of personality assessment, one can turn to a meta-analysis examining the ability of the MMPI to detect underreported pathology (Baer, Wetter, & Berry, 1992). Across the 27 independent effect sizes included in this meta-analysis, the average base rate for faking good was .50 (see their Table 1). A typical clinician working in an outpatient or inpatient mental health clinic certainly would not expect half their patients to be faking good on the MMPI. Nonetheless, this meta-analysis determined that the ability of the MMPI to detect faking was r = .46 (Cohen's d = 1.04) when 50% of all patients fake good. However, this effect size is artificial because it is based on a criterion (i.e., faking good vs. not) with inflated variance that does not generalize to most clinical settings. The results are very different when more realistic base rates are used. If we assume that 10% of patients in a typical clinical setting fake good, the effect size would drop to r = .30; if we assume that only 5% of patients in a typical clinical setting fake good, the validity coefficient would drop from r = .46 to r = .22.

The same problem of inflated effect size estimates affects the two meta-analyses that have examined the ability of the MMPI to identify patients who malinger illness (Berry, Baer, & Harris, 1991; Rogers, Sewell, & Salekin, 1994). In these meta-analyses, the average base rate of faking bad was 49%. One certainly should not expect that 50% of the people who complete the MMPI fake good while 49% fake bad. In a parallel fashion, one should not expect MMPI effect sizes to generalize to clinical practice when they are obtained from studies with such artificial base rates.

All three MMPI meta-analyses also produced artificially inflated results because they relied on many studies that used a more traditional extreme groups design. In the typical study, one half of the participants were instructed to fake their MMPI results, and the other half were instructed to respond honestly. Such a design cuts out all the people who would fall in the middle section of a faking continuum (i.e., all those who are neither completely honest nor globally faking). By omitting the central portion of a faking continuum, the MMPI effect sizes are inflated again, beyond the level of inflation that would be due to artificial base rates.

In some ways, it is surprising that Wood et al. (1999b) criticized a single Rorschach study for using an extreme groups design when the 90 studies included in these three MMPI meta-analyses are more severely confounded by this problem. As before, this suggests that Wood et al. are prone to hold the Rorschach to a more demanding standard than other psychological tests.

With respect to Wood et al.'s (1999b) criticism of Burns and Viglione's (1996) extreme groups design, several points should be noted. First, Burns and Viglione explained why they used this design, although the rationale was never noted by Wood et al. Specifically, Burns and Viglione excluded the middle portion of their distribution for two reasons: (a) so they did not have to spend the considerable time required to double or triple score all the midrange Rorschach protocols and (b) because they wished to ensure their participants did truly differ on the criterion (see Burns & Viglione, 1996, pp. 94–95). Although Wood et al. did not mention these reasons, they are the same two reasons that Wood et al. said would justify an extreme groups design (i.e., time savings and an interest in determining the presence of a relation between predictor and criterion, regardless of its size or shape; p. 120).

Second, Wood et al. (1999b) also overlooked the fact that Burns and Viglione (1996) conducted research on a homogeneous group that was likely to produce *smaller* than normal effect sizes. By limiting their sample to healthy, relatively well-functioning women who had been married for at least 1 year, had no psychiatric care in the previous 6 months, suffered no salient interpersonal losses during the same period, were not substance abusers, and had average intelligence, Burns and Viglione were more likely to find smaller effects than if they had examined a sample with a much wider range of functioning. For instance, if they had compared their healthy nonpatient women to women diagnosed with borderline personality disorder, they likely would have observed a much larger effect size for their key Rorschach variable.

Finally, Wood et al.'s (1999b) main criticism was that the effect sizes from Burns and Viglione's (1996) research would not reflect the magnitude of effects encountered in clinical practice. However, this criticism overlooks the fact that Burns and Viglione (a) never calculated effect sizes, (b) identified their results as applying to a homogenous group of female nonpatients, and (c) never suggested the effect sizes that could be calculated from their study would reflect those derived from a typical nonpatient sample, much less a clinical sample.

In general, like any statistical or methodological procedure, an extreme groups design can be used appropriately or not used appropriately. As Cortina and DeShon (1998) recently noted, the choice of an extreme groups design versus a standard observational design that contains all participants should be dictated by the questions a researcher wishes to address. If one is interested in "the existence of an effect or the estimation of its magnitude in raw-score terms, then the extreme-groups design is superior" (p. 804). Conversely, when "one wishes to address issues of the relative importance of continuous predictors, then the observational design, with its superior estimates of population variances, leads to more accurate inferences" (p. 804). However, Cortina and DeShon also noted that even an observational design can produce misleading effect sizes whenever sampling procedures restrict or exaggerate the underlying population parameters.

COMPOSITE MEASURES

Wood et al. (1999b) criticized Burns and Viglione (1996) for creating a composite criterion measure of interpersonal relatedness. Specifically, Wood et al. (1999b, p. 118) stated that it is “reasonable” to form a composite measure when the scales to be combined “correlate highly” with each other. If one’s goal is to maximize internal consistency reliability, this recommendation has some merit (although see Clark & Watson, 1995). However, if one wishes to obtain the most valid criterion measure, this recommendation is not appropriate. To obtain the most valid criterion measure, one seeks to limit the sources of systematic error affecting the criterion. In a previous discussion on criteria to validate Rorschach scores (Meyer, 1996), I described how criterion validity is maximized when each source of data entering into a composite is valid but also nonredundant (see also Clark & Watson, 1995, or the formulas in Rosenthal & Rubin, 1986, and Tsujimoto, Hamilton, & Berger, 1990). That is, validity is maximized when each scale entering the criterion is a decent measure of the construct yet contains unique variance not captured by other scales. In contrast to Wood et al.’s assertion, this means that the validity of a criterion is maximized when each scale in a composite is a reasonable measure of the construct yet has a relatively low correlation with the other variables.

Although this fact may initally seem paradoxical, the phenomenon represents the criterion counterpart to the factors that make optimal predictors in a regression equation. To maximize the predictive validity of a regression equation, one does not want highly correlated predictors because each variable would then provide redundant and unnecessary information. Instead, the optimal predictors in a regression equation are those that have strong correlations with the criterion but relatively low correlations with each other (Cohen & Cohen, 1983). Obtaining such variables allows each predictor to contribute unique information to the assessment. The same principles apply when developing an optimally valid criterion measure (see Little, Lindenberger, & Nesselroade, 1999, for a more sophisticated discussion of these issues).

It is certainly possible to create unreliable and invalid composite measures. However, the scales in Burns and Viglione’s (1996) study were neither. Their three scales had good reliability and an average intercorrelation of .50. The latter is a very substantial average correlation, although readers of the Wood et al. (1999b) critique would not have known this because Wood et al. did not report all the relevant correlations.

Furthermore, even though Wood et al. (1999b) said that Burns and Viglione’s (1996) composite measure “may be nearly impossible to interpret” (p. 118) due to combining distinct constructs, their criticism overlooked other relevant facts they had at hand. Wood et al. cited the original dissertation completed by Burns (1993). However, they never mentioned her pilot study that examined the composite measure of interpersonal relatedness (see Burns, 1993, pp. 59–64). Of greatest relevance, Burns factor analyzed the interpersonal scales and found that a single factor explained 67% of the variance. The three scales used by Burns and Viglione had loadings of .74, .92, and .79 on this factor. Particularly, because one scale was derived from self-report, whereas the others came from observer ratings, this clear evidence should leave no doubt that the scales formed a reasonable composite. Not reporting this very relevant factor-analytic data put Wood et al. in a position to allege weaknesses in Burns and Viglione's study that in fact did not exist.

Wood et al. (1999b) also singled out one measure in Burns and Viglione's (1996) composite for the most criticism. They said the Emotional Maturity Rating Form (EMRF) appeared to be an "inadequate" measure of interpersonal relatedness because it was "lacking demonstrated validity" and more than 25% of its items "appear to assess qualities that bear little relation to interpersonal relatedness" (p. 118). It is hard to reconcile Wood et al.'s personal objections to the item content of the EMRF with the available data. Burns (1993, pp. 55–59) described the development of the EMRF in some detail. She noted how the EMRF emerged from a child version of the test, which was developed and had its item content validated by ratings from 55 teachers. The item content in the adult version was subsequently validated by Tilden (1989), who used a panel of four experts to determine how well each item assessed emotional maturity. As Burns (1993) indicated, Tilden found "extremely high agreement among the judges concerning the content validity of the EMRF" (p. 58). Even though Wood et al. (1999b, pp. 119, 122) cited many pages from Burns (1993), they overlooked or ignored this relevant information about the content validity of the EMRF. Furthermore, although Wood et al. said the EMRF lacked "demonstrated validity," this statement overlooked or dismissed the clear EMRF validity coefficients that Burns reported for her two samples (i.e., the pilot study and the actual study; see Burns, 1993, pp. 55–64). Wood et al.'s criticism also overlooked relevant data from Tilden, who found the EMRF was a valid predictor of marital adjustment in a sample of 111 couples.

As a final point on composite scales, Wood et al.'s (1999b) general criticisms of aggregated, multisource criterion measures neglected or oversimplified some of the classic literature in personality assessment (e.g., Epstein, 1983, 1986; Rushton, Brainerd, & Pressley, 1983). Researchers who wish to implement optimal designs to validate psychological tests should closely review the latter citations and embrace composites when they are feasible to construct.

ERRONEOUS ASSERTIONS REGARDING THE *HEV*

As a final matter of particularly serious concern, Wood et al. (1999b) claimed there were two different formulas in the literature for computing the *HEV*. Although there are two formats for computing the *HEV,* one based on a traditional *z*-score format and the other based on variable weights, the underlying formula is not different.

In a section of their article prominently titled “The Two Versions of the *HEV,*” Wood et al. (1999b) stated:

> We turn next to the *HEV,* the central Rorschach variable in Burns and Viglione’s (1996) study. Here an important problem reveals itself: Two different and incompatible methods were used to compute the *HEV* variable, although this problem was not noted in the original article … The “*z* score method” and “weighting method” are intended to be different versions of the same formula, and are supposed to yield identical results (Burns & Viglione, 1996, p. 92). However, the two methods do *not* yield identical results, as can be found by anyone who performs the calculations. For instance, the Appendix of Burns and Viglione’s article (p. 99) gives a hypothetical example of a protocol with 3 *Poor H* [Poor Human Experience] and 5 *Good H* [Good Human Experience] responses. Using the weighting method, an *HEV* score of 2.18 is derived for the example (though –2.18 seems to have been intended). By contrast, using the *z* score method, with the means and standard deviations taken from Perry and Viglione (1991, p. 495, Table 2), we obtained an *HEV* score of –1.59 for the same example. Our calculations are shown in the Appendix of this article. The *z* score and weighting methods do not yield *HEV* scores that are identical or even very close. Most importantly, the two methods can change the order of *HEV* scores. For example, when the weighting method is used, Person A may have a higher *HEV* score than Person B. But when the *z* score method is used, Person B may have the higher score. Thus, the statistical results of the study by Burns and Viglione (1996) could change depending on which scoring method was actually used. (pp. 118–119)

Although this is a lengthy quote, there are two key issues: (a) the claim that the traditional *z*-score and weighting formats are different and (b) the claim that these formats are “incompatible,” “do *not* yield identical results,” “do not yield *HEV* scores that are identical or even very close,” “most importantly … can change the order of *HEV* scores” such that people shift their relative positions in the distribution of *HEV* scores, and that Burns and Viglione’s (1996) results “could change depending on which scoring method was actually used.” If these points were true, Wood et al. (1999b) would have been on solid ground warning about instability in the Burns and Viglione data. However, neither of Wood et al.’s assertions is true. Each is considered in turn.

THE “INCOMPATIBLE” *HEV* FORMULAS

Incompatible formulas for the *HEV* did not exist until Wood et al. (1999b) created a faulty formula for their article. To ensure clarity on this point, I briefly describe how *z* scores are computed in a traditional format and in an equivalent weighted format. Before doing so, however, it should be noted that the issue boils down to the difference between a formula written as $X = ½(Y)$ and $X = .5(Y)$. Although the numbers are superficially different, regardless of which format one uses, *X* will always be one half of *Y*'s original value.

The traditional formula for a single *z* score is

$$z = (\text{observed score} - M)/SD$$

Although this formula is not too complicated, to express the equation using weights one simply solves for parts of the equation. Specifically, the observed score and the mean are multiplied by the inverse of the sample standard deviation such that

$$z = 1/SD(\text{observed score}) - 1/SD(M)$$

For instance, assume that IQ is distributed in the population with a mean of 100 and a standard deviation of 15. The traditional *z*-score format for IQ is then $z =$ (observed score – 100)/15, whereas the equivalent weighted format is $z =$ .066667(observed score) – .066667(100), which can be simplified further to $z =$ .066667(observed score) – 6.6667.[6] A person with an IQ of 85 obtains a *z* score of –1.0 regardless of whether we use the traditional equation (i.e., [85 – 100]/15 = –1.000) or the weighted format (i.e., .066667[85] – 6.6667 = –1.000). Similarly, if MMPI–2 *T* scores are distributed in the population with a mean of 50 and a standard deviation of 10, then a person with a *T* score of 65 on Scale *F* of the MMPI–2 obtains a *z* score of 1.5 regardless of whether we use the traditional equation (i.e., [65 – 50]/10 = 1.500) or the equivalent weighted format (i.e., .10[65] – 5 = 1.500).

The procedures are similar when one wishes to compute the difference between two variables, as with Burns and Viglione's (1996) *HEV* formula, which computes the difference between *z* scores for *Poor H* and *Good H*. In general, the formula for the difference between two *z* scores is

$$z_{\text{diff}} = [(\text{observed score}_A - M_A)/SD_A] - [(\text{observed score}_B - M_B)/SD_B]$$

where *A* and *B* denote the two variables under consideration. Because this difference formula is slightly more complicated than the single variable formula, simplifying weights are of more value. The equivalent (but unsimplified) weighted formula is

$$\begin{aligned} z_{\text{diff}} = \ & [1/SD_A(\text{observed score}_A) - 1/SD_A(M_A)] - \\ & [1/SD_B(\text{observed score}_B) - 1/SD_B(M_B)] \end{aligned}$$

[6] The precision of the weighting formula depends on how much rounding occurs in the formula. Means, standard deviations, or weights that have been rounded substantially (e.g., from .066667 to .07) will create a degree of imprecision.

To demonstrate the equivalence of these formats, suppose someone wished to know the difference in *z*-score units between a patient's IQ score and his or her score on Scale *F* from the MMPI–2. The traditional *z*-score formula would be z_{diff} = [(observed score$_{\text{IQ}}$ – M_{IQ})/SD_{IQ}] – [(observed score$_{\text{F}}$ – M_F)/SD_F]. Using the population means and standard deviations given previously, the *z*-score formula becomes z_{diff} = [(observed score$_{\text{IQ}}$ – 100)/15] – [(observed score$_F$ – 50)/10]. The same formula expressed with simplifying weights becomes z_{diff} = .066667(observed score$_{\text{IQ}}$) – .10(observed score$_F$) – 1.6667. For a patient with the IQ and *F* scale values mentioned before (i.e., 85 and 65, respectively), one finds that the *z*-score difference is –2.5, regardless of whether one uses the traditional formula (i.e., [85 – 100]/15 – [65 – 50]/10 = –1.0 – 1.5 = –2.500) or the weighted formula (i.e., .066667[85] – .10[65] – 1.6667 = –.833305 – 1.6667 = –2.500).

These data document how there is nothing magical about generating an equivalent weighting formula from a traditional *z*-score formula. For the formulas to work together properly, however, after generating the weighted formula, one must constantly insert the correct means and standard deviations into the traditional *z*-score formula. If one generated a weighted formula but then computed a traditional *z*-score formula using the wrong means and standard deviations, it would appear as if the traditional formula and the weighted formula produced incompatible results. To exemplify, consider the *z* difference formulas from the previous paragraph. If one erroneously used 105 and 13 as the mean and standard deviation for IQ and erroneously used 45 and 8 as the mean and standard deviation for the *F* scale, it would now appear that the traditional *z*-score formula produced results that differed from the weighted formula. Specifically, the traditional *z*-score formula would now produce a value of –4.0385 (i.e., [85 – 105]/13 – [65 – 45]/8 = –1.5385 – 2.5 = –4.0385) rather than the correct value of –2.5 that is still produced by the weighting formula.

Thus, like the *HEV* example provided by Wood et al. (1999b), it can appear as if a traditional *z*-score formula differs from its theoretically equivalent weighted formula when in fact this is not the case. The real error comes from using faulty means and standard deviations in the traditional *z*-score formula.

Wood et al. (1999b) said they generated their traditional *z*-score formula by using "the means and standard deviations taken from Perry and Viglione (1991, p. 495, Table 2)" (p. 119). Unfortunately, these are not the correct means and standard deviations to use when generating *HEV* scores.

Perry and Viglione (1991) were explicit on this point and a careful read of their article should have alerted Wood et al. (1999b) to the error in their decision. The *HEV* and its constituent parts, *Good H* and *Poor H,* are elements of the Ego Impairment Index (EII). In their article, Perry and Viglione explained how the EII and its components were initially developed using Rorschach data from a sample of depressed patients collected by Haller (1982). Haller's sample was, thus, the original sample that was used to generate the means and standard deviations for *Good H* and *Poor H.* Perry and Viglione presented factor analytic findings from Haller's data set in Table 1 of their article. The text states that the information in this table came from Haller's sample (see p. 491), and the table note indicates how *Good H* was "the transformed standardized score of good human experiences," (p. 492) whereas *Poor H* was "the transformed standardized score of poor human experiences" (p. 492). As such, Perry and Viglione's article indicated that Haller's sample had been used to create the *z* scores for these variables. If one wished to calculate a traditional *z* score formula for the *HEV,* it would be necessary to obtain Haller's descriptive data for *Good H* and *Poor H.* Wood et al. did not do this. Instead, they used data from Perry and Viglione's Table 2. The means and standard deviations given in this table dealt with a separate study that was unrelated to Haller's original sample. By using means and standard deviations from the wrong sample, Wood et al. produced a *z*-score formula that seemed to disagree with Burns and Viglione's (1996) weighted formula.

Although it is possible that the information in Perry and Viglione's (1991) article was not sufficiently clear or that Wood et al. (1999b) had not read the article closely, another fact bears on this issue. Early in 1998, I was one of five people who reviewed a version of Wood et al.'s article when it was submitted to a different journal.[7] At that time, one of the reviewers pointed out how Wood et al. were "confused and mistaken" in their computations for the *z*-score formula. The reviewer explicitly described how the *HEV* had been derived from Haller's (1982) sample, not from the data reported in Table 2 of Perry and Viglione.[8] For some reason, Wood et al. did not attend to this corrective input from the peer review process and continued to promulgate a *z*-score formula that they had been told was derived from improper means and standard deviations. Ultimately, even though Wood et al. went to great lengths to convince readers that there were problems with the *HEV* formula, it seems that what they demonstrated was an unwavering capacity to insert the wrong numbers into their calculations.

[7] Like me, some may question the appropriateness or wisdom of revealing information that emerged during the review process. Before deciding to do so, I consulted with the American Psychological Association's Research Ethics Officer and the editor of the other journal (which does not require anonymity from its reviewers).

[8] Reviewer 4's full comment was

> The authors refer to Perry and Viglione (1991). These authors should know, per Perry and Viglione (p. 491) that the EII and the HEV weights are derived from data from Haller (1982). One of the strengths of Perry and Viglione (1991) was that the HEV and EII weights were derived from one sample (Haller, 1982) and cross-validated with another. Thus, without recognizing this clearly stated fact (in Perry & Viglione, 1991), no wonder the authors are confused and mistaken in their calculations.

THE ASSOCIATION BETWEEN WOOD ET AL.'S (1999b) FAULTY FORMULA AND THE CORRECT FORMULA

Setting aside the fact that Wood et al. (1999b) championed a formula that they knew was incorrect, Wood et al. also claimed their faulty formula and the correct formula were "incompatible," "do *not* yield identical results," "do not yield *HEV* scores that are identical or even very close," and "most importantly … can change the order of *HEV* scores" to produce distinct statistical findings (pp. 118–119). Are these claims true? Does the faulty Wood et al. *z*-score formula produce results that are so dramatically at odds with the correct formula? The answer to both questions is *no.* Furthermore, Wood et al. knew their statements were not true before they submitted their final article for publication.

Recall that there are three formulas under consideration. First, there is the correct *HEV* *z*-score formula computed in the traditional format. This formula uses the means and standard deviations derived from Haller's (1982) original sample. Donald Viglione (personal communication, November 20, 1998) supplied these values when I requested them. The mean and standard deviation for *Poor H* are 3.02 and 1.98, respectively, whereas the values for *Good H* are 2.09 and 1.33, respectively. Using this information produces the following *z*-score formula:

$$\text{Correct } \textit{HEV} \text{ Traditional } z \text{ Score} = (\textit{Poor H} - 3.02)/1.98 - (\textit{Good H} - 2.09)/1.33$$

The second *z*-score formula is the weighted formula presented by Burns and Viglione (1996). This formula is computed as follows:

$$\text{Correct } \textit{HEV} \text{ Weighted } z \text{ Score} = .51(\textit{Poor H}) - .75(\textit{Good H}) + .04$$

Finally, there is the *HEV* *z*-score formula created by Wood et al. (1999b). This formula used the wrong means and standard deviations, and it is computed as follows:

$$\text{Faulty Wood et al. (1999b) } \textit{HEV}\ z \text{ Score} = (\textit{Poor H} - 3.8)/2.48 - (\textit{Good H} - 2.63)/1.86$$

The critical question is how these three formulas relate to each other. Table 4 presents results using 232 patients from the sample of mine described earlier. Two facts are obvious from Table 4. First, the correct *HEV* traditional *z*-score formula and the correct *HEV* weighted *z*-score formula have a correlation of 1.0000. Thus, as expected, these formulas produce results that are perfectly correlated with each other (despite rounding error in both formulas). Perhaps most importantly, however, the faulty Wood et al. *HEV* *z*-score formula produces correlations in excess of .9985 with the correct formulas. As a result, when considered to 2 decimal

TABLE 4
Pearson Correlations Indexing the Degree of Association Among Wood et al.'s (1999b) Faulty Formula and the Correct Formulas for Computing the Human Experience Variable (*HEV*)

HEV Formula	*1*	*2*	*3*
1. Correct *HEV* traditional *z* score[a]	—	—	—
2. Correct *HEV* weighted *z* score[b]	1.0000	—	—
3. Faulty Wood et al. *HEV z* score[c]	.9986	.9989	—

Note. *N* = 232.
[a] [(*Poor H* – 3.02)/1.98] – [(*Good H* – 2.09)/1.33]. [b] .51(*Poor H*) – .75(*Good H*) + .04. [c] [(*Poor H* – 3.8)/2.48] – [(*Good H* – 2.63)/1.86].

places, Wood et al.'s faulty formula rounds up to a correlation of 1.00 with each of the correct formulas.

Given the remarkable association between these formulas, it is troubling to consider that Wood et al. (1999b) were aware of these findings before they submitted their article for final publication. That is, before going to press, asserting that these formulas were "incompatible," "do *not* yield identical results," "do not yield *HEV* scores that are identical or even very close," and "most importantly … can change the order of *HEV* scores," the authors had been told that, at worst, they were describing correlations greater than .9985. The following two facts document this point.

First, when I reviewed the prior version of Wood et al.'s (1999b) manuscript, my written review contained results from seven simulation studies that documented the extent of association between the faulty Wood et al. *HEV z*-score formula and the correct *HEV* formula.[9] I chose to use simulation studies because James Wood (Wood, Tataryn, & Gorsuch, 1996) published research using these techniques and because he has facilitated my own research (see Meyer, 1997b) using these procedures. Thus, I anticipated the simulation evidence would be clear to him. If not, I knew he had the skills to redo the analyses himself. Each of the seven simulation samples relied on data from 500 cases, and they modeled results that would emerge when different means and standard deviations were used for the *Good H* and *Poor H* variables. Across the seven samples, the correlation between Wood et al.'s (1999b) faulty *HEV* formula and the correct *HEV* formula ranged from a low of .9989 to a high of .9991. Wood and his colleagues received this written feedback in late April or early May of 1998—well before they submitted their manuscript to *Assessment.*

[9] At the time, I inappropriately assumed that the means and standard deviations used in the faulty Wood et al. (1999b) *HEV z*-score formula were correct. Although I should have returned to Perry and Viglione's (1991) original article to double-check this point, I did not. Thus, my simulation samples documented the extent of association between the faulty Wood et al. *HEV z*-score formula and the correct weighted formula but not the correct traditional formula.

Table 5 presents the results of two similar simulation samples. Each sample contains 1,000 computer generated cases with scores for *Good H* and *Poor H*. The first sample was constrained to have means and standard deviations equal to those used in the faulty Wood et al. (1999b) *HEV z*-score formula. The second sample was constrained to have distributions equal to those used in the correct *HEV z*-score formula. From Table 5, one can see how the correct *HEV* formulas produce perfect correlations of 1.0000 in each sample. As before, the incorrect formula created by Wood et al. produces correlations in excess of .9984 with each of the correct formulas.

Second, if this simulation data were not sufficient, on October 14 and 15, 1998, James Wood and I discussed these issues on the Rorschach Discussion List, a professional listserver located at rorschach@maelstrom.stjohns.edu. At the time, I presented the data from my patient sample (see Table 4) to Wood and the several hundred other members of the list. Thus, about 8 months before Wood et al. (1999b) published their article, the first author had seen clear evidence that, at worst, his faulty *HEV* formula produced a near-perfect correlation with the correct formula in a large sample of genuine patients. Despite this, Wood et al. still went to print claiming that the *HEV* formulas were "incompatible," "do *not* yield identical results," "do not yield *HEV* scores that are identical or even very close," and "most importantly … can change the order of the *HEV* scores."

Shortly before the Wood et al. (1999b) article was published, I requested and received a copy of the manuscript from James Wood. After reading it, I contacted him and said I was confused, baffled, and troubled by the *HEV* statements quoted previously and asked how he could make those assertions in good conscience, knowing the magnitude of the correlation between his faulty formula and the correct *HEV* formula. When he responded to my comments and question (J. Wood, personal communication, June 1, 1999), he summarized his position as follows:

TABLE 5
Pearson Correlations Indexing the Degree of Association Among Wood et al.'s (1999b) Faulty Formula and the Correct Formulas for Computing the Human Experience Variable (*HEV*) Using Two Samples Containing Simulated *Good H* and *Poor H* Scores

HEV Formula	*1*	*2*	*3*
1. Correct *HEV* traditional *z* score	—	1.0000	.9985
2. Correct *HEV* weighted *z* score	1.0000	—	.9988
3. Faulty Wood et al. *HEV z* score	.9985	.9988	—

Note. *N*s = 1,000. Results for the first set of random *Poor H* and *Good H* scores are listed below the diagonal; results for the second set are listed above the diagonal. The first set of variables was constrained to have means and standard deviations equal to those used in the faulty Wood et al. *HEV z*-score formula. The second set was constrained to have means and standard deviations equal to those used in the correct *HEV* traditional *z*-score formula. See Table 4 for each specific formula.

> In your message, you ask how "in good conscience" we could criticize Burns and Viglione on this point, in light of your analyses. Although you seem to see it as an ethical or moral issue, we see it as an intellectual issue: In our view, we are acting reasonably even if we fail to find your analyses as compelling as you do. There is no issue of "conscience" here: You find your numbers highly convincing, but we are still in considerable doubt.

Perhaps some readers will also find the correlations reported in Tables 4 and 5 to be unconvincing evidence on the equivalence of these formulas. Perhaps some will also agree with Wood and his colleagues and find these numbers leave room for "considerable doubt." Perhaps some readers will still believe that Burns and Viglione's (1996) results would have been different even if they used the faulty Wood et al. *HEV z*-score formula, which they did not.[10]

If these are reasonable conclusions to draw from the data, however, one would have expected Wood et al. (1999b) to be quite up front with the degree of correlation between their *z*-score formula and the correct *HEV* formula. They were not. Instead, they told readers that the two formulas were "incompatible," "do *not* yield identical results," "do not yield *HEV* scores that are identical or even very close," and "most importantly … can change the order of the *HEV* scores." I suspect most readers would never envision that such "reasonable" descriptions about incompatibility were being applied to variables that were known to correlate in excess of .998, far above the reliability of any psychological test.

SUMMARY

Wood et al.'s (1999b) article contained several general points that are quite sound. Conducting research with an extreme groups design does produce effect sizes that are larger than those observed in an unselected population. Appropriate control groups are important for any study that wishes to shed light on the characteristics of a targeted experimental group and experimental validity is enhanced when researchers collect data from both groups simultaneously. Diagnostic efficiency statistics—or any summary measures of test validity—should be trusted more when they are drawn from multiple studies conducted by different investigators across numerous settings rather than from a single investigator's work. There should be no question that these points are correct.

[10] An additional point should be noted here. Burns and Viglione (1997) published a correction to their original article that indicated some of their analyses had been generated with a *HEV z*-score formula that used the means and standard deviations derived from their sample rather than from Haller's (1982) original sample. For the data sets that were used to generate Tables 4 and 5, this other formula always had a correlation between .993 and .995 with the correct *HEV* formulas. Although Wood et al. (1999b, p. 119) criticized Burns and Viglione for not being more specific about this degree of correlation, Wood et al. knew the magnitude of these correlations because I had included them in my seven simulation samples and the genuine patient data discussed on the Rorschach listserver.

However, I have pointed out numerous problems with specific aspects of Wood et al.'s (1999b) article. Wood et al. gave improper citations that claimed researchers found or said things that they did not. Wood et al. indicated my data set did not support the incremental validity of the Rorschach over the MMPI–2 when, in fact, my study never reported such an analysis and my data actually reveal that the opposite conclusion is warranted. Wood et al. asserted there was only one proper way to conduct incremental validity analyses even though experts have described how their recommended procedure can lead to significant complications. Wood et al. cited a section of Cohen and Cohen (1983) to bolster their claim that hierarchical and stepwise regression procedures were incompatible and to criticize Burns and Viglione's (1996) regression analysis. However, that section of Cohen and Cohen's text actually contradicted Wood et al.'s argument. Wood et al. tried to convince readers that Burns and Viglione used improper alpha levels and drew improper conclusions from their regression data although Burns and Viglione had followed the research evidence on this topic and the expert recommendations provided in Hosmer and Lemeshow's (1989) classic text. Wood et al. oversimplified issues associated with extreme group research designs and erroneously suggested that diagnostic studies were immune from interpretive confounds that can be associated with this type of design. Wood et al. ignored or dismissed the valid reasons why Burns and Viglione used an extreme groups design, and they never mentioned how Burns and Viglione used a homogeneous sample that actually was likely to find smaller than normal effect sizes. Wood et al. also overlooked the fact that Burns and Viglione identified their results as applying to female nonpatients; they never suggested their findings would characterize those obtained from a clinical sample. Wood et al. criticized composite measures although some of the most important and classic findings in the history of research on personality recommend composite measures as a way to minimize error and maximize validity. Wood et al. also were mistaken about the elements that constitute an optimal composite measure. Wood et al. apparently ignored the factor-analytic evidence that demonstrated how Burns and Viglione created a reasonable composite scale, and Wood et al. similarly ignored the clear evidence that supported the content and criterion related validity of the EMRF. With respect to the *HEV*, Wood et al. created a *z*-score formula that used the wrong means and standard deviations. They continued to use this formula despite being informed that it was incorrect. Subsequently, Wood et al. told readers that their faulty *z*-score formula was "incompatible" with the proper weighted formula and asserted that the two formulas "do *not* yield identical results" and "do not yield *HEV* scores that are identical or even very close." These published claims were made even though Wood et al. had seen the results from eight large samples, all of which demonstrated that their wrong formula had correlations greater than .998 with the correct formula.

At worst, it seems that Wood et al. (1999b) may have intentionally made statements that they knew were incorrect. If so, these statements were then used to make plausible sounding but fallacious arguments about weaknesses in Rorschach validation research. The latter could be seen as an instances of sophist rhetoric, in which arguments are designed to convince readers of a conclusion, regardless of its accuracy. At minimum, whenever sophistry occurs, it stretches the boundaries of proper scientific conduct and trivializes the scientific endeavor into a caricature of the search for knowledge. Such efforts would be particularly striking if they occurred among authors who often refer to ethical principles and professional standards to make a point (Nezworski & Wood, 1995; Wood et al., 1996a, 1996b).

At best, the authors were not sufficiently careful in their scholarship (e.g., the erroneous citations), were not aware of some key literature on a topic (e.g., the composite variables), presented a limited and slanted portrayal of relevant issues and evidence (e.g., overlooking relevant information in Cohen & Cohen, 1983; Hosmer & Lemeshow, 1989; Tilden, 1989; and Burns, 1993), and repeatedly dismissed corrective feedback (e.g., regarding their faulty *z*-score formula and its near-unity correlation with the correct formula). These errors and oversights are reminiscent of issues that have emerged before. For instance, Wood et al. (1996a, 1996b) criticized Comprehensive System scoring reliability and suggested that it may be poor. However, they never presented any evidence to justify that claim, and they disregarded numerous studies that negated it (see Meyer, 1997a, 1997c; Wood et al., 1997).

Given all of this, it seems fair to conclude that even under the most benign interpretation of how Wood et al.'s (1999b) false and misleading statements found their way into print, the authors did not carefully check the accuracy and balance of their assertions and did not correct pivotal mistakes that had been identified for them. Wood et al.'s article was putatively written to offer methodological guidance to Rorschach researchers. They briefly criticized one point in a lengthy article by Weiner (1996), expounded on limitations in Ganellen's (1996a) database although Ganellen had himself repeatedly articulated the same limitations, and devoted the majority of their article to criticizing various aspects of Burns and Viglione's (1996) study. Wood et al. never pointed out a methodological strength in any of the articles they reviewed.

The latter should be a clue to readers. Evidence indicates the same study will be seen as containing more methodological flaws when it produces results that are at odds with preexisting beliefs than when it produces results consistent with existing beliefs (e.g., Koehler, 1993; Lord, Ross, & Lepper, 1979). This effect seems most pronounced when the preexisting beliefs are strongly held (Koehler, 1993). Given that Wood et al. (1999b) ignored important corrective feedback about errors in their *HEV* formula and then found it unconvincing when eight large samples of data produced correlations in excess of .998 between their wrong formula and the correct formula, it is likely that no amount of strong evidence will be sufficient to dislodge their generally negative view of the Rorschach and its research base. Their zeal to criticize the Rorschach does not always seem to be tempered by reason or fact.

Documenting construct validity for test scales is a slow and cumbersome process. Every individual study contains flaws or shortcomings, so it is only through the gradual accumulation of research employing different types of designs, samples, and criteria that one can confidently validate test scales. In my view, the research by Burns and Viglione (1996) was methodologically sophisticated, not deficient as Wood et al. (1999b) would have readers believe. As such, it reflected an important step in the right direction for validating the *HEV*.

As the Rorschach evidence base continues to grow and develop, sound and balanced criticism of the literature will help advance scientific knowledge and applied practice. Conversely, publishing assertions that are known to be wrong or misleading can only serve political purposes that thwart the goals of science and retard genuine evolution in the field. Because of its many problems, the Wood et al. (1999b) article does not provide illuminating guidance. Those who wish to have a balanced understanding of Rorschach limitations and strengths would be wise to consider other sources.

REFERENCES

Archer, R. P. (1996). MMPI–Rorschach interrelationships: Proposed criteria for evaluating explanatory models. *Journal of Personality Assessment, 67,* 504–515.

Archer, R. P., Aiduk, R., Griffin, R., & Elkins, D. E. (1996). Incremental validity of the MMPI–2 content scales in a psychiatric sample. *Assessment, 3,* 79–90.

Archer, R. P., Elkins, D. E., Aiduk, R., & Griffin, R. (1997). The incremental validity of MMPI–2 supplementary scales. *Assessment, 4,* 193–205.

Archer, R. P., & Gordon, R. A. (1988). MMPI and Rorschach indices of schizophrenic and depressive diagnoses among adolescent inpatients. *Journal of Personality Assessment, 52,* 276–287.

Archer, R. P., & Krishnamurthy, R. (1997). MMPI–A and Rorschach indices related to depression and conduct disorder: An evaluation of the incremental validity hypothesis. *Journal of Personality Assessment, 69,* 517–533.

Baer, R. A., Wetter, M. W., & Berry, D. T. R. (1992). Detection of underreporting of psychopathology on the MMPI: A meta-analysis. *Clinical Psychology Review, 12,* 509–525.

Ball, J. D., Archer, R. P., Gordon, R. A., & French, J. (1991). Rorschach Depression indices with children and adolescents: Concurrent validity findings. *Journal of Personality Assessment, 57,* 465–476.

Barthlow, D. L., Graham, J. R., Ben-Porath, Y. S., & McNulty, J. L. (1999). Incremental validity of the MMPI–2 content scales in an outpatient mental health setting. *Psychological Assessment, 11,* 39–47.

Ben-Porath, Y. S., Butcher, J. N., & Graham, J. R. (1991). Contribution of the MMPI–2 content scales to the differential diagnosis of schizophrenia and major depression. *Psychological Assessment, 3,* 634–640.

Ben-Porath, Y. S., McCully, E., & Almagor, M. (1993). Incremental validity of the MMPI–2 Content Scales in the assessment of personality and psychopathology by self-report. *Journal of Personality Assessment, 61,* 557–575.

Ben-Porath, Y. S., & Sherwood, N. E. (1993). *The MMPI–2 content component scales: Development, psychometric characteristics, and clinical application.* Minneapolis: University of Minnesota Press.

Ben-Porath, Y. S., & Waller, N. G. (1992). 'Normal' personality inventories in clinical assessment: General requirements and the potential for using the NEO Personality Inventory. *Psychological Assessment, 4,* 14–19.

Berry, D. T. R., Baer, R. A., & Harris, M. J. (1991). Detection of malingering on the MMPI: A meta-analysis. *Clinical Psychology Review, 11,* 585–598.

Burns, B. (1993). An object relations study: Relationship between the Rorschach Human Experience Variable and interpersonal relatedness among nonpatients (Doctoral dissertation, California School of Professional Psychology, San Diego, 1993). *Dissertation Abstracts International, 51–04B,* 3847.

Burns, B., & Viglione, D. J. (1996). The Rorschach Human Experience variable, interpersonal relatedness, and object representation in nonpatients. *Psychological Assessment, 8,* 92–99.

Burns, B., & Viglione, D. J. (1997). Correction to Burns and Viglione (1996). *Psychological Assessment, 9,* 82.

Butcher, J. N., Dahlstrom, W. G., Graham, J. R., Tellegen, A., & Kaemmer, B. (1989). *Manual for the restandardized Minnesota Multiphasic Personality Inventory: MMPI–2. An administrative and interpretive guide.* Minneapolis: University of Minnesota Press.

Butcher, J. N., Williams, C. L., Graham, J. R., Archer, R. P., Tellegen, A., Ben-Porath, Y. S., & Kaemmer, B. (1992). *Manual for administration, scoring, and interpretation of the Minnesota Multiphasic Personality Inventory for Adolescents: MMPI–A.* Minneapolis: University of Minnesota Press.

Christensen, D., Hadzi-Pavlovic, D., & Jacomb, P. (1991). The psychometric differentiation of dementia from normal aging: A meta-analysis. *Psychological Assessment, 3,* 147–155.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7,* 309–319.

Cohen, J. (1988). *Statistical power for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49,* 997–1003.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Cortina, J. M., & DeShon, R. P. (1998). Determining relative importance of predictors with the observational design. *Journal of Applied Psychology, 83,* 798–804.

Costa, P. T., Jr., & McCrae, R. R. (1992). Normal personality assessment in clinical practice: The NEO Personality Inventory. *Psychological Assessment, 4,* 5–13.

Dahlstrom, W. G. (1996). Comparability of MMPI and MMPI–2 profile patterns: Ben-Porath and Tellegen's inappropriate invocation of Mahalanobis's D_2 function. *Journal of Personality Assessment, 66,* 350–354.

Epstein, S. (1983). Aggregation and beyond: Some basic issues on the prediction of behavior. *Journal of Personality, 51,* 360–392.

Epstein, S. (1986). Does aggregation produce spuriously high estimates of behavioral stability? *Journal of Personality and Social Psychology, 50,* 1199–1210.

Ganellen, R. J. (1996a). Comparing the diagnostic efficiency of the MMPI, MCMI–II, and Rorschach: A review. *Journal of Personality Assessment, 67,* 219–243.

Ganellen, R. J. (1996b). Integrating the Rorschach and MMPI–2: Adding apples and oranges?: Introduction. *Journal of Personality Assessment, 67,* 501–503.

Greene, R. L. (1991). *The MMPI–2/MMPI: An interpretive manual.* Boston: Allyn & Bacon.

Haller, N. (1982). *The reliability of Rorschach depressive indices in major depressive disorder.* Unpublished doctoral dissertation, United States International University, San Diego, CA.

Harkness, A., McNulty, J., & Ben-Porath, Y. (1995). The Personality Psychopathology Five (*PSY–5*): Constructs and MMPI–2 scales. *Psychological Assessment, 7,* 104–114.

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

Hilsenroth, M. J., Fowler, J. C., Padawer, J. R., & Handler, L. (1997). Narcissism in the Rorschach revisited: Some reflections on empirical data. *Psychological Assessment, 9,* 113–121.

Hosmer, D. W., & Lemeshow, S. (1989). *Applied logistic regression.* New York: Wiley.

Hunsley, J., & Bailey, J. M. (1999). The clinical utility of the Rorschach: Unfulfilled promises and an uncertain future. *Psychological Assessment, 11,* 266–277.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings.* Newbury Park, CA: Sage.

Koehler, J. J. (1993). The influence of prior beliefs on scientific judgments of evidence quality. *Organizational Behavior and Human Decision Processes, 56,* 28–55.

Lijmer, J. C., Mol, B. W., Heisterkamp, S., Bonsel, G. J., Prins, M. H., van der Meulen, J. H. P., & Bossuyt, P. M. M. (1999). Empirical evidence of design-related bias in studies of diagnostic tests. *Journal of the American Medical Association, 282,* 1061–1066.

Little, T. D., Lindenberger, U., & Nesselroade, J. R. (1999). On selecting indicators for multivariate measurement and modeling with latent variables: When "good" indicators are bad and "bad" indicators are good. *Psychological Methods, 4,* 192–211.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology, 37,* 2098–2109.

McCann, J. T. (1998). Defending the Rorschach in court: An analysis of admissibility using legal and professional standards. *Journal of Personality Assessment, 70,* 125–144.

Meyer, G. J. (1993). The impact of response frequency on Rorschach constellation indices and on their validity with diagnostic and MMPI–2 criteria. *Journal of Personality Assessment, 60,* 153–180.

Meyer, G. J. (1996). Construct validation of scales derived from the Rorschach method: A review of issues and introduction to the Rorschach rating scale. *Journal of Personality Assessment, 67,* 598–628.

Meyer, G. J. (1997a). Assessing reliability: Critical corrections for a critical examination of the Rorschach Comprehensive System. *Psychological Assessment, 9,* 480–489.

Meyer, G. J. (1997b). On the integration of personality assessment methods: The Rorschach and MMPI–2. *Journal of Personality Assessment, 68,* 297–330.

Meyer, G. J. (1997c). Thinking clearly about reliability: More critical corrections regarding the Rorschach Comprehensive System. *Psychological Assessment, 9,* 495–498.

Meyer, G. J. (1999a). The convergent validity of MMPI and Rorschach scales: An extension using profile scores to define response and character styles on both methods and a reexamination of simple Rorschach response frequency. *Journal of Personality Assessment, 72,* 1–35.

Meyer, G. J. (1999b). Incremental validity of the Rorschach Prognostic Rating Scale over the MMPI Ego Strength Scale and IQ. *Journal of Personality Assessment, 74,* 356–370.

Meyer, G. J. (1999c). Introduction to the special series on the utility of the Rorschach for clinical assessment. *Psychological Assessment, 11,* 235–239.

Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Kubiszyn, T. W., Moreland, K. L., Eisman, E. J., & Dies, R. R. (1998). *Benefits and costs of psychological assessment in healthcare delivery: Report of the Board of Professional Affairs Psychological Assessment Work Group, Part I.* Washington, DC: American Psychological Association.

Meyer, G. J., & Handler, L. (1997). The ability of the Rorschach to predict subsequent outcome: A meta-analysis of the Rorschach Prognostic Rating Scale. *Journal of Personality Assessment, 69,* 1–38.

Munley, P. H., Busby, R. M., & Jaynes, G. (1997). MMPI–2 findings in schizophrenia and depression. *Psychological Assessment, 9,* 508–511.

Nezworski, M. T., & Wood, J. M. (1995). Narcissism in the comprehensive system for the Rorschach. *Clinical Psychology: Science and Practice, 2,* 179–199.

Paunonen, S. V. (1998). Hierarchical organization of personality and prediction of behavior. *Journal of Personality and Social Psychology, 74,* 538–556.

Perry, W., & Viglione, D. J. (1991). The Ego Impairment Index as a predictor of outcome in melancholic depressed patients treated with tricyclic antidepressants. *Journal of Personality Assessment, 56,* 487–501.

Persons, J. B. (1986). The advantages of studying psychological phenomena rather than psychiatric diagnoses. *American Psychologist, 41,* 1252–1260.

Psychological Corporation. (1997). *WAIS–III and WMS–III technical manual.* San Antonio, TX: Author.

Rogers, R., Sewell, K. W., & Salekin, R. T. (1994). A meta-analysis of malingering on the MMPI–2. *Assessment, 1,* 227–237.

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (Rev. ed.). Newbury Park, CA: Sage.

Rosenthal, R., & Rubin, D. B. (1986). Meta-analytic procedures for combining studies with multiple effect sizes. *Psychological Bulletin, 99,* 400–406.

Rushton, J. P., Brainerd, C. J., & Pressley, M. (1983). Behavioral development and construct validity: The principle of aggregation. *Psychological Bulletin, 94,* 18–38.

Schinka, J. A., LaLone, L., & Greene, R. L. (1998). Effects of psychopathology and demographic characteristics on MMPI–2 scale scores. *Journal of Personality Assessment, 70,* 197–211.

Shrout, P. E. (1997). Should significance tests be banned? Introduction to a special section exploring the pros and cons. *Psychological Science, 8,* 1–20.

Tellegen, A., & Ben-Porath, Y. S. (1996). Evaluating the similarity of MMPI–2 and MMPI profiles: Reply to Dahlstrom and Humphrey. *Journal of Personality Assessment, 66,* 640–644.

Tilden, R. S. (1989). Predicting marital adjustment with level of object relations, romantic love, and emotional maturity (Doctoral dissertation, United States International University, 1989). *Dissertation Abstracts International, 51–04B,* 2088.

Tsujimoto, R. N., Hamilton, M., & Berger, D. E. (1990). Averaging multiple judges to improve validity: Aid to planning cost-effective clinical research. *Psychological Assessment, 2,* 432–437.

Viglione, D. J. (1999). A review of recent research addressing the utility of the Rorschach. *Psychological Assessment, 11,* 251–265.

Weiner, I. B. (1996). Some observations on the validity of the Rorschach Inkblot Method. *Psychological Assessment, 8,* 206–213.

Wetzler, S., Khadivi, A., & Moser, R. K. (1998). The use of the MMPI–2 for the assessment of depressive and psychotic disorders. *Assessment, 5,* 249–261.

Whitely, W. P., Rennie, D., & Hafner, A. W. (1994). The scientific community's response to evidence of fraudulent publication: The Robert Slutsky case. *Journal of the American Medical Association, 272,* 170–173.

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.

Wood, J. M., Nezworski, M. T., & Stejskal, W. J. (1996a). The Comprehensive System for the Rorschach: A critical examination. *Psychological Science, 7,* 3–10.

Wood, J. M., Nezworski, M. T., & Stejskal, W. J. (1996b). Thinking critically about the Comprehensive System for the Rorschach: A reply to Exner. *Psychological Science, 7,* 14–17.

Wood, J. M., Nezworski, M. T., & Stejskal, W. J. (1997). The reliability of the Comprehensive System for the Rorschach: A Comment on Meyer (1997). *Psychological Assessment, 9,* 490–494.

Wood, J. M., Nezworski, M. T., Stejskal, W. J., Garven, S., & West, S. G. (1999a). Erratum for "methodological issues in evaluating Rorschach validity: A comment on Burns and Viglione (1996), Weiner (1996), and Ganellen (1996)." *Assessment, 6,* 305.

Wood, J. M., Nezworski, M. T., Stejskal, W. J., Garven, S., & West, S. G. (1999b). Methodological issues in evaluating Rorschach validity: A comment on Burns and Viglione (1996), Weiner (1996), and Ganellen (1996). *Assessment, 6,* 115–129.

Wood, J. M., Tataryn, D. J., & Gorsuch, R. L. (1996). Effects of under- and overextraction on principal axis factor analysis with Varimax rotation. *Psychological Methods, 1,* 354–365.

Gregory J. Meyer
Department of Psychology
University of Alaska Anchorage
3211 Providence Drive
Anchorage, AK 99508
E-mail: afgjm@uaa.alaska.edu

Received November 28, 1999
Revised January 19, 2000