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 in-

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. Further-

tion, 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.

cluded 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 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 d been9F4 1 Tfı˝164.464 0 TDı˝(TDw˝(STTjı˝/F2 1 Tfı˝20.772

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 analy-

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 stepagainst 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 their statistical significance and regardless of whether the analysis produced conceptually meaningful results. Next, on Block 2, the *SCZI*

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.

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

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

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 vance, 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

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 in-

bers 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.

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 differ-

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:

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

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

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

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:

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

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.”

> 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

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. NmI–2 contr-160(Nmcr-16P″(diagnoses among .376 -1(dio5rrent)-238(va

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.

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.

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