Neuropsychology Review, Vol. 14, No. 3, September 2004 (© 2004)

Rohling's Interpretive Method for Neuropsychological Case Data: A Response to Critics

Martin L. Rohling,[1,3] L. Stephen Miller,[2] and Jennifer Langhinrichsen-Rohling[1]

In the September 2001 issue of *Neuropsychology Review* Miller and Rohling published a description of the Rohling Interpretive Method (RIM). These authors indicated that the RIM could be used to analyze an individual patient's test results obtained from a flexible neuropsychological battery. Two critiques of the RIM were submitted (Palmer, Appelbaum, & Heaton, 2004; Willson & Reynolds, 2004), both of which are printed in the current issue. This paper is a response to these two author groups concerns about the clinical and psychometric feasibility of the RIM. We provide both psychometric theory and data analyses to refute each of the two author groups' main objections. We conclude with a recommendation that neuropsychologists adopt the RIM for use in their day-to-day practice to improve their diagnostic accuracy and treatment planning skills. The main reason for use of the RIM is to avoid several common errors in clinical judgment that have been well documented in the literature (e.g., Dawes, Faust, & Meehl, 1989).

KEY WORDS: neuropsychological assessment; flexible battery; clinical judgment; case data analysis; Rohling's Interpretive Method.

In the September 2001 issue of this journal, we published a method of individual data interpretation (Miller and Rohling, 2001) called the Rohling Interpretative Method (RIM). The method has also been described in a recent book chapter (Rohling et al., 2003). As a reminder, the steps of the method are summarized in Table 1. Our main purpose for developing the RIM was to address problems noted in the literature when interpreting data obtained in an assessment of an individual patient.

The preceding two critiques of the RIM by Willson and Reynolds (2004) and Palmer et al. (2004) concluded that the statistical and interpretative problems they have identified are of sufficient magnitude that the method should not be used at this time by clinicians who are assessing individual patients. We do not believe this to be the case, as we are able to adequately respond to each of the concerns raised by these two critiques. This is the purpose of the current paper. For convenience, we respond to both critiques simultaneously, as the main criticisms generated by these two author groups often overlapped. Specifically, there were eight substantive concerns identified. Each is listed below.

1. The RIM's suggested method of calculating standard deviations (*SD*s) for both global summary statistics and cognitive domain scores is in error. Since many of the remaining steps of the RIM depend on the use of these *SD*s, this error is magnified in the subsequent steps.
2. Use of the RIM will result in more diagnostic false-positives then traditional clinical judgment. Specifically, Palmer et al. (2004) expressed concern that we failed to distinguish "statistical significance" from "clinical significance" and that our failure to make such a distinction is a critical error in our system.
3. Clinicians who use the RIM will idiosyncratically assign test scores to cognitive domains, resulting in low interrater reliability in RIM analyses and diagnoses.
4. The RIM recommends factor loadings of test scores on domains be unit weighted, which introduces error to the analysis. Willson and Reynolds (2004) suggested that many test scores load on

[1] Department of Psychology, University of South Alabama, Mobile, Alabama.

[2] Department of Psychology, University of Georgia, Athens, Georgia.

[3] To whom correspondence should be addressed at Department of Psychology, University of South Alabama, 381 Life Sciences Building, Mobile, Alabama 36688-0002; e-mail: mrohling@usouthal.edu.

1040-7308/04/0900-0155/0 © 2004 Plenum Publishing Corporation

Table 1. Steps to Rohling's Interpretive Method (RIM) for Neuropsychological Case Data

Summary statistics: Steps 1–17

1. Design and administer a flexible test battery.
2. Estimate premorbid general ability (*EPGA*).
3. Convert test scores to a common metric.
4. Assign each test's scores to domains.
5. Calculate domain means, standard deviations, and sample sizes.
6. Calculate test battery means.
7. Calculate probabilities of heterogenity.
8. Determine categories of cognitive impairment.
9. Determine the percentage of test scores that fall in the impaired range.
10. Calculate effect sizes for all domains and *TBM* scores.
11. Calculate confidence intervals for all domains and TBM scores.
12. Determine the upper limit necessary for premorbid performance.
13. Conduct one-sample *t* tests on each type of mean generated.
14. Conduct a between-subjects ANOVA with domain means.
15. Conduct a power analysis.
16. Sort test scores in ascending order.
17. Graphically display all of the summary statistics.

Interpretation: Steps 18–24

18. Determine the test battery's validity.
19. Determine if psychopathology influenced test scores.
20. Use test battery means to determine if impairment exists.
21. Determine current strengths and weaknesses.
22. Examine scores from noncognitive domains.
23. Explore low power comparisons for Type II errors.
24. Examine the response operating characteristics of sorted *T*-scores.

multiple cognitive domains and that the assignment of scores to domains, as well as the appropriate weights used on those domains, is dependent on the battery of tests administered to the patients whose test scores are being examined.

5. The RIM recommends that multiple measures be used to generate composite scores, which, according to our critics, will result in less rather than more accurate estimates of the cognitive domains of interest.
6. The RIM uses a general ability factor (i.e., Estimate of Premorbid General Ability or EPGA) to represent premorbid functioning for all cognitive domains. According to our critics, this recommendation is not supported by the literature. As a result, it will result in inaccurate conclusions regarding the degree of impairment suffered by a patient in each of the cognitive domains assessed.
7. Norms used to generate *T* scores will come from samples that are of undocumented comparability. Furthermore, even when norms are used that were generated from different but roughly comparable samples, their format may prohibit ready comparisons.
8. Use of the RIM will result in an undue inflation of clinicians' confidence. Such overconfidence will result in more error in a clinician's interpretation and not less.

We address each of these concerns in turn. We reply with psychometric theory, as well as by conducting data analyses. These additional analyses are generated from four datasets summarized below and described in detail at the end of this manuscript.[4]

Dataset 1 consisted of 607 psychiatric inpatients aged 25–34 years, which was a subset of a much larger dataset of 2,395 inpatients from Edmonton, Alberta. Several analyses from this dataset have been published previously (e.g., Iverson et al., 1999, 2001; Iverson and Green, 2002). The dependent variables were all subtests and summary scores of either the WAIS (150 patients) or WAIS-R (457 patients).

[4] Dataset 1: Used by permission from Dr Paul Green, Edmonton, Alberta, Canada. Two-thousand three hundred and ninety five psychiatric inpatients were generally suffering from significant psychopathology. Two subgroups from this sample consisting of the age group 25–34 years were used in our analyses —457 patients with complete WAIS-R protocols, and 150 patients with complete WAIS protocols. Restriction to this age group allowed subtests recorded in the dataset to be age-adjusted.

Dataset 2: Used by permission from Dr Paul Green. Nine-hundred and four outpatients seen for neuropsychological assessment in the context of a Canadian Workers' Compensation Board claim ($n = 376$), a medical disability claim ($n = 317$) or personal injury litigation ($n = 196$). Financial benefits for disability were potentially available to or were being received by the remaining 15 patients referred privately. The sample included head injured patients and neurological patients ($n = 550$), psychiatric patients ($n = 107$; major depression, anxiety disorders, bipolar mood disorders, and psychotic illness), and medical patients ($n = 246$; orthopedic injuries, chronic fatigue syndrome, chronic pain syndrome, fibromyalgia, and other various conditions). Included in the test battery were 42 neuropsychological dependent variables.

Dataset 3: Used by permission from Dr John Meyers, in Sioux City, IA. Seventeen-hundred and thirty four inpatients and outpatients who were primarily referred for neuropsychological assessment. Included the Meyers Short Battery were 26 neuropsychological dependent variables.

Dataset 4: Used by permission from Drs Russell Adams, Oklahoma City, Oklahoma, and David J. Williamson, Clearwater Florida. One-hundred fourteen patients, with 73 identified as brain injured and 42 identified as "pseudoneurological controls" (i.e., psychiatric patients). Each patient has a complete HRB along with other measures. Each patient's GNDS had also been calculated for the purposes of cross-validation of this global measure of severity of neurocognitive impairment (Sherer and Adams, 1993). Furthermore, each patient had been diagnosed as brain-injured or not without reference to psychometric data, using a variety of medical tests (e.g., CT, EEG, and MRI). Included in the test battery were 36 neuropsychological dependent variables.

Dataset 2 consisted of 904 outpatients seen for neuropsychological assessment, all involved in some sort of medical–legal disability claim. These patients were also from Edmonton, Alberta. This sample included head injured and neurological patients ($n = 550$), psychiatric patients ($n = 107$), and medical patients ($n = 246$). Five symptom validity measures and 43 neuropsychological tests were administered. Various analyses also have been published from subsets of this sample (e.g., Green et al., 2001; Green and Iverson, 2001a,b; Rohling et al., 2002 a,b).

Dataset 3 consisted of 1,734 mixed inpatients and outpatients who had been mostly referred to assess neurocognition and had been given the Meyers Short Battery (e.g., Pilgrim et al., 1999; and Volbrecht et al., 2000). The patients were all seen in Sioux City, IA. Again, reports on subsets of these patients have been published in the literature (e.g., Meyers et al., 2002a,b; Meyers and Volbrecht, 2003; Volbrecht et al., 2000).

Dataset 4 consisted of 114 patients independently identified as either brain-injured ($n = 73$) or psychiatric ($n = 41$), from the Oklahoma University Health Sciences Center, Oklahoma City, OK. Each patient had been administered a complete Halstead-Reitan Battery (HRB; Reitan and Wolfson, 1985, 1993), as well as several other measures. This dataset also has been previously analyzed and the results published (Sherer and Adams, 1993; Rohling et al., 2003c Vanderploeg et al., 1997).

CRITICISM 1: ERRORS IN THE CALCULATION OF THE STANDARD DEVIATIONS

The first criticism of the RIM centers on how we recommend that clinicians calculate *SD*s for the relevant summary statistics. After reviewing both critiques, we believe that Willson and Reynolds (2004) and Palmer et al. (2004) misunderstood our recommendations as to how a clinician is to go about generating these *SD*s. It is far simpler than was suggested in their articles. Both sets of critics impressively, and at some length, explained "errors" in our *SD* calculations and go on to present statistical procedures that they believe are required to accurately generate the *SD*s of interest. However, after careful reading, it is apparent that they have focused on *group data* generated from *a number of patients*, rather than on *groups of data* generated from a *single patient*. In essence, we believe that they focused on what most researchers focus on when examining data (i.e., interindividual *SD*s), compared to that which is of most interest to practicing clinical neuropsychologists (i.e., intraindividual *SD*s).

We found no error in their statistical reasoning or the methods they recommended if one were interested in calculating the interindividual *SD* for a sample of patients on a single composite score. However, this is *not* the *SD* that is relevant, nor is it the *SD* that we recommended generating when using the RIM. It is also *not* the *SD* which most clinicians would have access. Again, the *SD*s of importance in the RIM are the intraindividual *SD*s, or the amount of variability within a single patient's set of scores, *not* the amount of variability across a group of patients on a single composite score.

For illustrative purposes, consider the typical spreadsheet used to enter data from an empirical study. Columns are used to record the dependent variables for the participants in the study. Rows are used to record a single participant's set of scores on each dependent variable. If a sample of patients had been given a battery of tests (e.g., WAIS-III), then the columns of the spreadsheet are usually defined by the subtest scores and the rows of the spreadsheet are usually defined by the participants included in the study. The *SD* that both critiques focused on is that which is generated from a single column (i.e., dependent variable) across multiple rows (i.e., participants). The *SD*s used in the RIM are best thought of as generated from multiple columns (i.e., dependent variables) across a single row (i.e., a participant).

Palmer et al. (2004) used formulas to suggest that the *SD* of a composite score shrinks as one adds more measures. Furthermore, they stated that the *SD* for an overall test battery mean *could not* exceed 9.99 (see latter). However, what is obvious is that the *SD* of a battery of tests for a *single* patient is *independent* of the standard deviation of the measures within that battery across a sample of patients. Therefore, adding measures to a particular patient's test battery to create a composite score for all patients does not alter the amount of variability across measures displayed by any one patient, which can and does exceed 9.99.

To demonstrate, we conducted analyses on Dataset 1. Within this dataset, a subset consisting of 20–34-year-old patients was selected, as earlier versions of the Wechsler intelligence test used this normative group to generate scaled scores for each subtest (Wechsler, 1955, 1981). Within this age range, we analyzed a subset of 457 patients who had been administered the WAIS-R and 150 patients who had been administered the WAIS. Within each subset, individual patient's subtest scaled scores were summed within two global domains–verbal subtests (i.e., VIQ) and performance subtests (i.e., PIQ). Finally, the sum of all scaled scores was used to generate the Full Scale IQ (FSIQ). These three summary scores were recorded using age-corrections. To address the *SD* criticism, using Dataset 1, we summed each patient's scaled scores within the verbal and performance subtests,

Table 2. Mean and Standard Deviations for the Four Datasets Summary Scores

	Mean *SD*			
	Interindividual group		Interaindividual	
Dataset 1 (psychiatric Pts.)				
WAIS-R ($n = 457$)	43.2	7.2	6.8	2.0
WAIS ($n = 150$)	45.0	9.1	7.4	2.2
Dataset 2 (Green)	44.8	7.3	11.4	2.9
Dataset 3 (Meyers)	42.0	7.3	11.9	2.9
Dataset 4 (HRB-OHSU)	42.8	6.8	10.6	2.4

separately. We then generated an intraindividual *SD* for these verbal and performance summary scores. Again, this is more straightforward than suggested by our critics. The sample size for the verbal summary score was six subtests for each patient and for the performance summary score there were five subtests for each patient. Finally, when generating this "test battery's" overall mean (i.e., Overall Test Battery Mean [OTBM]), we used all 11 subtest scores included in the WAIS and WAIS-R and typically used to generate the FSIQ. In essence, the RIM requires the calculation of variability across the 11 scores of the OTBM, as well as the six scores of the verbal and five scores of the performance domains. These are the intraindividual *SD*s (in this case, three for each patient). Results are shown in Table 2.

Since the RIM recommends the use of *T* scores, and that was the type of norm-referenced score used by both critiques, we chose to use these scores in our discussion. Because each patient's set of scores has their own unique variability, there are 457 intraindividual FSIQ *SD*s in the WAIS-R dataset and 150 intraindividual *SD*s in the WAIS dataset. As seen in Table 2, the OTBM interindividual mean across all patients was 43.2 with a *SD* of 7.2. Furthermore, the mean of the 457 intraindividual *SD*s was 6.8, ranging from 2.5 to 13.8. In fact, 7% of the WAIS-R subset had *SD*s greater than 9.99. Similarly, using the 150 patients in the WAIS subset, the OTBM interindividual mean was 45.0 with a *SD* of 9.1. However, the mean of the 150 intraindividual *SD*s was 7.4, ranging from 2.7 to 15.8. Finally, 10% of the WAIS subset had *SD*s greater than 9.99. Palmer et al. (2004) reported that the *SD*s of a summary measure, generated from a set of scores that each had normative *SD*s equal to 10, *could not* be any larger than 9.99 because of the intercorrelation of test scores amongst one another. As seen in the results generated from an actual clinical dataset, and using the correct method of calculating the intraindividual statistics, it is quite possible to generate scores that exceed the suggested ceiling. Notice that the *SD* to which both of the critiques referred is a single *SD* for the global measure and not several unique *SD*s for the patients included in the sample.

We conducted similar analyses on the three other datasets to show that these *impossible* results occur no matter what type of battery is administered by a clinician. Figure 1 shows the histograms of the group mean OTBM for each of the four datasets. Figure 2 illustrates histograms of the group mean intraindividual *SD*. As shown in Fig. 1, using Dataset 2, the mean OTBM was 44.8 with an interindividual *SD* of 7.3. However, the intraindividual *SD* from 858 patients shown in Fig. 2 had a mean of 11.4 (*SD* = 2.9), which ranged from 4.1 to 23.7. Furthermore, 65% of the sample had *SD*s greater than 9.99. Dataset 3 had a mean OTBM of 42.0, with an interindividual *SD* of 7.3. But, the mean intraindividual *SD* from these 1,731 patients was 11.9 (*SD* = 2.9), which ranged from 2.5 to 21.5. In addition, 56% of the sample had *SD*s greater than 9.99. Finally, using the HRB data from Dataset 4, we used Heaton et al.'s norms (Heaton et al., 1991) to calculate regression-based *T* scores. Again, the mean OTBM was 42.8, with an interindividual *SD* from 114 patients of 6.8. This, in fact, is similar to results published by Heaton et al. (2001) in their study of schizophrenia patients showing an OTBM (a.k.a. Global Neurological *T* Score) of 41.9 and an interindividual *SD* of 6.5. However, the mean intraindividual *SD* from these 114 patients was 10.6 (*SD* = 2.4) and ranged from 5.6 to 17.0. In addition, 61% of the sample had *SD*s greater than 9.99. Heaton et al. (2001) did not present intraindividual *SD*s in their article.

As suggested by both critiques, the OTBM's interindividual *SD* was indeed less than 10 in each of the clinical datasets we analyzed. In fact, Palmer et al.'s estimated *SD* for the OTBM (Palmer et al., 2004) of 6.4 was close to the actual *SD*s generated from datasets two through four (7.3, 7.3, and 6.8, respectively). However, the mean intraindividual *SD* was generally larger, equivalent across test batteries (11.4, 11.9, and 10.6, respectively), and had similar variability of the intraindividual *SD*s (2.9, 2.9, and 2.4, respectively). So, as the reader can see, not

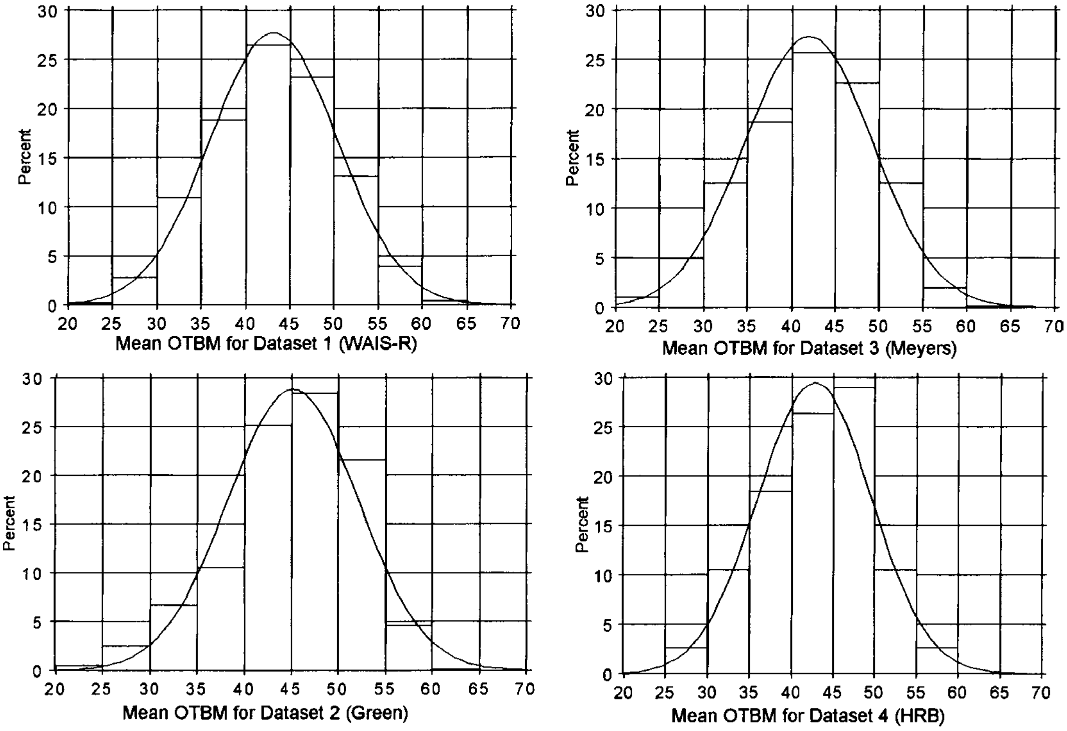

Fig. 1. Average Overall Test Battery Means (OTBM) for each of the four datasets analyzed.

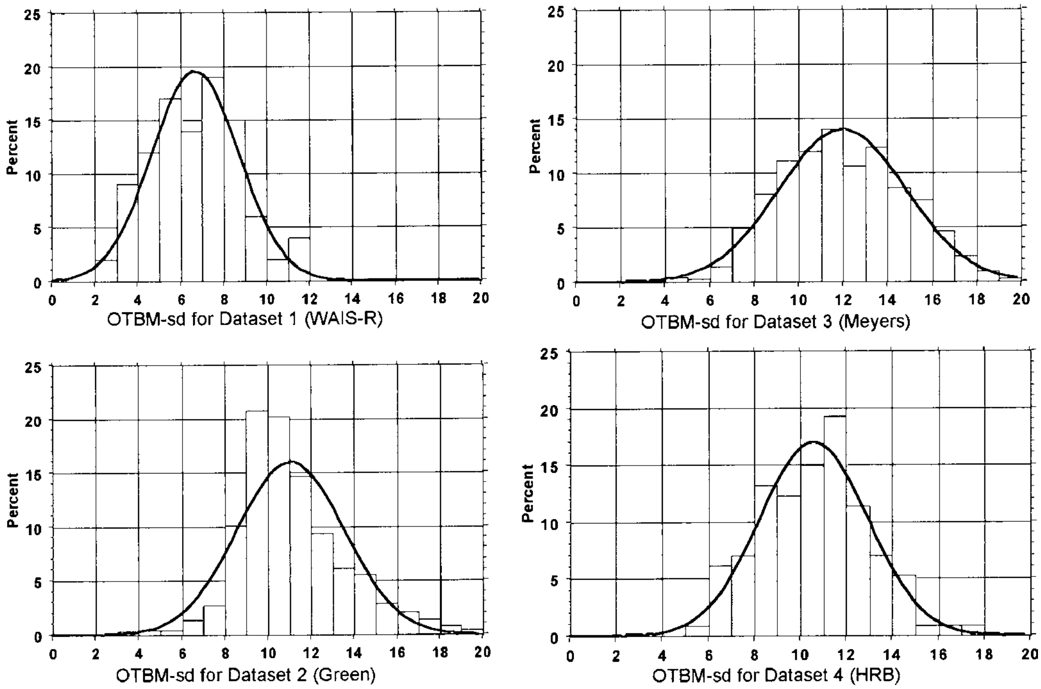

Fig. 2. Mean intraindividual SD for patients' OTBMs for each of the four datasets.

only is it possible for a *SD* of a composite score to be greater than 9.99, but, for the majority of patients (58% of all 2,706 patients) in the four datasets, this was the more likely outcome. If one looks at the patients described in the original RIM article (Miller and Rohling, 2001), the *SD* was 11.4 for Mr Strokes and 11.7 for Mr Sugar. This is just as one would expect, considering the results just presented for these large datasets. The assessment data for both of these patients' were published in our original article (Miller and Rohling, 2001). Interested readers could rerun our analyses with these data and would find that neither of our critics used the correct method for generating the *SD* recommended for the RIM. Again, it is not that their methods are wrong if one wants to calculate the *SD* for a sample of patients on a composite generated from multiple test scores. They just misunderstood what we recommend as the *SD* of interest. Clinicians are less interested in the interindividual *SD* for the single composite compared to the unique intraindividual *SD*s for a patient's set of scores. One need not know the correlation matrix or use complicated statistics to generate this intraindividual *SD*. Instead, any clinician who has administered a number of tests to a single patient can easily generate it—which, in reality, is the situation facing most clinical neuropsychologists and one of our underlying rationales for the development of the RIM.

There are two common reasons why patients have larger intraindividual *SD*s. First, the patient may have suffered a localized brain disorder (i.e., trauma or disease). Second, the patient may have not put forth consistent effort across subtests. Patients who fail symptom validity tests typically have larger *SD*s, with greater intraindividual variability across tests. They are also more likely to fail the RIM heterogeneity check. For example, when examining the two clinical datasets, each has a unique method of determining if a patient's obtained scores should be considered valid. Dr Green uses the Word Memory Test (WMT; Green et al., 2002; Green and Allen, 1995b; Hartman, 2002) and the Computerized Assessment of Response Bias (CARB; Green and Allen, 1995a; and Green and Iverson, 2001a). Dr Meyers uses a series of ability algorithms that have been shown to be sensitive to inadequate effort (Meyers and Volbrecht, 2003). In the Green dataset, comparing genuine patients to exaggerating patients, the genuine patients' mean intraindividual *SD* was 10.7 versus 13.2, respectively, $t(850) = 11.4$, $p < .0001$. In the Meyers dataset, the results were 10.2 versus 11.5, respectively, $t(454) = 4.82$, $p < .0001$. Finally, when examining the data from the HRB in the Oklahoma sample, the mean intraindividual *SD* for the pseudoneurological group was 9.7 versus 11.1 for the neurologically impaired group, $t(112) = 3.19$, $p = .0019$.

Despite our critics misunderstanding, we wish to thank them for presenting their concerns. Obviously, this point was not as clear as we intended it to be in our original paper (Miller and Rohling, 2001). Furthermore, in our initial recommendations, we suggested that clinicians use an *SD* of 10 to calculate effect sizes for each patient (Step 10B). This makes for a simple rule and approximates the intraindividual differences likely to exist in most genuine performing patients. However, to increase precision of a RIM analysis, we now believe that the use of the patient's actual intraindividual *SD* is the more appropriate procedure. We revise our original recommendation by suggesting that clinicians use the most precise *SD* that they can. Along these lines, for greater precision clinicians should use Hedge's method of calculating an effect size (*g*), generating a *pooled-SD* from the premorbid (EPGA), and postmorbid (OTBM) variables.

One additional point needs to be made. Willson and Reynolds (2004) claimed that we confused and inappropriately presented an estimate of the standard error of estimate, rather than the standard error of measurement. This is incorrect, but highlights our poor description of the formula presented in Step 11 to obtain domain confidence intervals (CIs). The formula to be used, and presented correctly in the RIM, remains unchanged: $S_{\text{mean of X}} = S_X/\sqrt{n}$ (Heiman, 2003, p. 272; Wechsler, 1997, pp. 53–56). This is more accurately defined as the formula for the estimated standard error of the mean. We thank Willson and Reynolds (2004) for identifying our awkward descriptor. Nevertheless, this remains the appropriate statistic to generate CI's and appears to be another example of our failure to convey our recommended use of intra rather than interindividual statistics.

Criticism 2: RIM Is Too Sensitive and Statistical Differences Are Not Clinically Meaningful

A second concern raised by our critics is that the RIM is overly sensitive to differences in scores that may be statistically significant, but of little clinical meaning. According to Palmer et al. (2003), over-responding to statistical significance obtained via the RIM would result in encouraging clinicians to label too many patients as cognitively impaired.

To test this assertion, we used the WAIS-R data from Dataset 1. To simulate the RIM recommendations and test the validity of our critics' concern, we looked at two different methods of determining statistically significant differences between a patient's VIQ and PIQ. The first method used the normative sample and followed the procedures recommended by Matarazzo et al., (1988), Sattler (2001), and that which is detailed in the WAIS-III Administration and Scoring Manual (Wechsler, 1997; Tulsky, Zhu, & Ledbetter, 1997). The *Manual* method relies on the standard error of the difference and requires that one know the correlation between the two composite scores. This is the method most clinicians use to determine if there is a significant difference between Verbal and Performance IQs. In contrast, the second method used the RIM procedures (i.e., unpaired *t* test) to determine if the obtained differences between VIQ and PIQ were statistically significant for an individual patient. To test this, we used a between-subjects *t* test, comparing the mean and standard deviation of the scaled scores for the verbal subtests ($n = 6$) and the mean and standard deviation of the scaled scores for the performance subtests ($n = 5$). Again, we highlight that, although we recommend a between-subject *t* test, we are actually conducting the *t* test between two types of data generated from a *single* individual. We examined all 457 patients in the 25–34-year-old age group. As expected, the *manual* method found 44% of the sample to have significant VIQ-PIQ splits. These results are similar to what was found for the WAIS-III, where 42% of the standardization sample had significant VIQ-PIQ splits (Wechsler, 1997). However, using the *t* test procedure of the RIM on each of the 457 patients, only 22% of the sample was found to have significant VIQ-PIQ differences. The contingency results comparing these two methods are presented in Table 3. Furthermore, the mean effect size needed to determine that there was a significant VIQ-PIQ split for an individual was *larger* for the RIM method than for the traditional method. Therefore, contrary to our critics' concern, the RIM method appears to be more conservative than the method recommended in the WAIS-III Manual.

It is important to note that these two methods of analyzing VIQ-PIQ differences are actually orthogonal to one another. Variability in the power of each of the two types of statistical tests will determine which method is more liberal and which is more conservative. In practice, however, we assert that the power to detect differences will usually be greater when using the interindividual method than when using the intraindividual method. Examining the results shown in Table 3, the methods overlapped 75% of the time, with 54% of patients showing nonsignificant VIQ-PIQ differences and 21% showing significant differences between VIQ and PIQ. However, the *manual* method identified an additional 23% of sample as having significant VIQ-PIQ differences that were not identified by the RIM method. Furthermore, the RIM method identified only an additional 1% of cases as having significant VIQ-PIQ differences that were not identified by the *manual* method.

Remember that the method recommended by the Manual, which uses the results of the interindividual differences, uses the group data to estimate the variability of an individual. The RIM method uses the actual variability of the individual who was assessed. Because of the additional statistical power of the large normative sample in the Manual method, statistically significant differences obtained in the traditional fashion may not be of much clinical significance. This is not the case for the RIM. These data support our contention that the RIM method, as described in the original manuscript, is generally more conservative than the traditional method of determining statistically significant differences using data from a normative sample. This, in turn, increases the likelihood that the RIM method generates clinically meaningful information.

Finally, detection of clinically meaningful differences is a function of base rate, sensitivity, and specificity. To address these issues, we examined a subset of

Table 3. Contingency Table Between RIM and Manual Methods for Detecting Differences in Domain Scores

	Manual method		
RIM *t* test method	VIQ-PIQ split nonsignificant	VIQ-PIQ split significant	Marginal means for VIQ-PIQ splits
VIQ-PIQ split: nonsignificant	$n = 248$ (54%)	$n = 107$ (23%)	$n = 355$ (78%)
	$g = 0.38$ (.30)	$g = 0.80$ (.41)	$g = 0.50$ (.39)
	SS $M = 3.9$ (2.5)	SS $M = 13.2$ (3.7)	SS $M = 6.7$ (5.2)
VIQ-PIQ split: significant	$n = 6$ (1%)	$n = 96$ (21%)	$n = 102$ (22%)
	$g = 1.58$ (.82)	$g = 1.70$ (.86)	$g = 1.69$ (.85)
	SS $M = 6.7$ (.8)	SS $M = 19.0$ (6.5)	SS $M = 18.3$ (6.9)
Marginal means VIQ-PIQ split	$n = 254$ (56%)	$n = 203$ (44%)	$n = 457$ (100%)
	$g = 0.40$ (.37)	$g = 1.22$ (.80)	$g = 0.90$ (.71)
	SS $M = 4.0$ (2.5)	SS $M = 15.9$ (6.0)	SS $M = 9.3$ (7.4)

Note. g: effect size with pooled *SD*; Number in parentheses are standard deviations. SS: Standard score differences: Numbers in the parentheses are standard deviations.

TBI patients from Dataset 3 (Rohling et al., 2003b). Patients were assigned to one of six severity groups based on criteria generated by Dikmen et al. (1995). The sample contained 291 patients, the majority of which had been assigned to the mildest of severity group. There were clear differences based on severity of TBI. Specifically, for the six severity groups, the following percentages of patients were identified as suffering from statistically significant differences from premorbid functioning using a one sample *t* test as recommended by the RIM: 27, 38, 50, 71, 89, and 83%, respectively. The differences in detection rates correspond with the differences in the effect size related to the severity of injury. The magnitude of these effect sizes were estimated from Dikmen et al. (1995), which equaled −.02, −.22, −.45, −.68, −1.33, and −2.31, respectively. Only 47% of the entire sample of TBI patients were found to be suffering from significant neurocognitive impairment, as measured by the RIM one-sample *t* test, using the OTBM and the EPGA as an estimate of the population mean. The minimum effect size detectable with this 26-item OTBM was −.47. As expected, as the effect size of severity increases across the groups, the percent of patients within a group who are identified by the RIM as suffering from neurocognitive impairment also increases. Moreover, in Dataset 4, which uses the HRB, results of the RIM method overlapped with results from the GNDS of Reitan and Wolfson (1993) 83% of the time. When the two methods disagreed, it was twice as likely that the RIM method would have identified impairment when the GNDS did not as vice versa Rohling et al., 2003c.

These results highlight that the base rate of clinically meaningful differences, as detected by the RIM method, is conservative and not liberal, as suggested by our critics. This is a basic power issue for test batteries. Batteries that have too few measures are likely to generate more Type II errors (i.e., fail to reject the null when the alternative hypothesis of cognitive impairment is true) than Type I errors (i.e., reject the null and conclude that the alternative hypothesis of cognitive impairment is true).

CRITICISM 3: LOW INTERRATER RELIABILITY AND IDIOSYNCRATIC FINDINGS USING DIFFERENT BATTERIES

Willson and Reynolds (2004) noted, and we concur, that the RIM was recommended for clinicians using a flexible battery approach to neuropsychological assessment. The RIM provides a method of obtaining benefits traditionally associated with fixed batteries, while still maintaining a flexible approach. However, we certainly did not recommend that clinicians use "skewed" test batteries, which would result in noncomprehensive assessments of important neurocognitive domains. If such test batteries are used and then submitted to RIM procedures, significant differences among the OTBM, Domain Test Battery Mean (DTBM), and Instrument Test Battery Mean (ITBM) are likely to emerge.[5] For example, if clinicians only assesses the domains of memory-learning and verbal-comprehension skills and then generates the three summary test battery means, they will typically find them to be discrepant. This warns clinician that there might have been selective testing of the relevant cognitive domains and is a strength of the RIM method. Furthermore, the power of the test of any poorly assessed domain will be rather low, because the sample size for those domains will be low. Low statistical power for the assessment of a particular domain should be another clue that the clinician has not comprehensively assessed a domain of the patient. The broader question of the degree to which ethical practice would support a clinician's decision to minimally assess or even fail to assess a particular domain is a question not addressed by the RIM methodology.

A broader concern evoked by the reviewers is that clinicians will not agree as to which domains a particular test result belongs, because there is not a universally agreed upon factor structure. While there is some validity to this point, we believe that there is converging evidence that there *are* universal domains of cognition. For example, Tulsky et al. (2003) recently conducted a series of factor analyses on the standardization sample of the WAIS-III and WMS-III. They found that a six-factor solution best fit the data, similar to the six-factor solution we presented in our original RIM paper. The factors identified were (1) Verbal Comprehension, (2) Perceptual Organization, (3) Working Memory, (4) Processing Speed, (5) Auditory Memory, and (6) Visual Memory. The only differences between this domain structure and that which we presented is that our Memory and Learning factor has been split into a auditory and visual components and Tulsky et al. (2003) did not include an executive function factor.

We factor analyzed the WAIS-R data from Dataset 1 and came to a four-factor solution, which included the same factors found in the WAIS-III standardization sample with each subtest loading on the same factors that were generated with the WAIS-III (i.e., VCI, POI, WMI, and PSI). True, it would be best to try to generate such a universal domain structure from a meta-analysis of all available factor analysis studies on this issue. However, to date we do not have these results. If one wishes to

[5] When examining the three clinical datasets (i.e., numbers 2,3, and 4), these Test Battery Means correlated highly with one another, with a mean coefficient of .97 that ranged from .92 to .99. Furthermore, in two of the four datasets (i.e., numbers 2 and 4), these global summary indices were not significantly different from one another in magnitude.

administer a battery using the WAIS-III and the WMS-III, along with a few other tests, it is our impression that we can assume that the factor/domain structure we originally presented reasonably model the factor structure of the available data.

Many commonly used flexible battery tests (Lees-Haley et al., 1996; Sweet et al., 2000) have been included in a variety of factor analyses and have been empirically tied to particular domains (e.g., see Leonberger et al., 1992). Thus, for many, if not most routinely used tests, domain placement is self-evident. When placement of a particular test is inappropriate, the RIM provides evaluative information for this with the domain heterogeneity statistic. High heterogeneity suggests that the clinician has chosen to load a test on an inappropriate factor.

Finally, we would remind our critics that the OTBM is independent of domain placement, because all individual tests are used to calculate this summary score. Therefore, comparisons between the OTBM and the EPGA are unaffected by test score domain assignment.

CRITICISM 4: FACTOR LOADINGS ERROR INTRODUCED BY UNIT WEIGHTING

We spent a great deal of time considering various permutations of the factor analysis concern when designing the RIM. We concur with our critics that placing test scores on factors and then unit weighting these scores requires clinician expertise and judgment. We expect practicing clinical neuropsychologists to have some command of the literature. We understood the dangers of this assumption when we recommended that clinicians, particularly those already using flexible batteries (which by their nature are infused with clinical judgments about tests and measurements), use the RIM. Here are a few of our thoughts related to published factor analytic results and the application of these results to the RIM.

First, concerning factor analysis, the number of dependent variables assessed in a research sample will have an impact on the factor structure and loadings (Nunnally & Bernstein, 1994). The more variables examined and the less distinct the constructs assessed, the less stable will be the factor structure and the smaller will be the factor loadings. Second, factor loadings are influenced by sample demographics. The more homogeneous a sample (e.g., Alzheimer's patients), the smaller the number of factors that are generated, the larger will be the factor loadings, and the less generalizable the results will be to patients who are not well represented by the sample. Third, the less construct validity a variable has, the less stable will be the factor structure and the smaller will be the factor loadings. For example, Picture Arrangement (PA) of the WAIS-III loads on two or three domains in the factor analytic studies detailed in the WAIS-III & WMS-III Technical Manual (e.g., verbal comprehension, perceptual organization, and working memory). When PA is included in a factor analysis, it will have smaller factor loadings on any one domain and these loadings will tend to be less stable than will a subtest like Vocabulary. Vocabulary loads highly on only one factor because it is a "purer" measure of a hypothetical construct (i.e., verbal intelligence) than is PA. Finally, the closeness with which a distribution of scores adheres to the assumptions required for parametric statistics (i.e., normally distributed, equal variance, and independently sampled), the more stable will be the factor structure and the resulting loadings. When these assumptions are violated, as is often the case with tests like the Boston Naming Test (BNT; Kaplan et al., 1983), the factor structure becomes less stable and the loadings less reliable. All of these problems with factor analysis are well described by Nunnally and Bernstein (1994). Given these realities, variability in factor loadings is to be expected across studies and samples. This is actually one of the reasons why we recommended using unit weighting.

We also chose to use unit weights because of the literature indicating that this is better than relying on clinicians' judgment. For example, research on beta weights has shown that unit weighting is superior to subjective weighting in human decision-making (Dawes, 1979; Dawes and Corrigan, 1974; Meehl, 1997; Wedding & Faust, 1989). Diagnostic accuracy can be improved only slightly (approximately 3–5%) by generating more appropriate beta weights (e.g., weighting scores by the normative sample sizes of each dependent variable or factor loadings from prior research). When such beta weights are not available, or when the time and effort required to incorporate them outweights the limited increase in diagnostic accuracy, one should not avoid using unit weighting because we know that these weights are not as good as they could be. Failing to use statistical/actuarial procedures because a clinician does not have the most appropriate beta weights will result in lower diagnostic accuracy, because s/he then must rely on less reliable and valid subjective judgment.

Dataset 1 allows us to estimate the differences in results obtained if a clinician were to use unit weighting, as we recommend, rather than ideal beta weighting advocated for by the critics. We conducted four multiple regression analyses using the 457 patients' WAIS-R data. We split the sample in half so that we could assess the effect of shrinkage on the accuracy of prediction generated from a single sample. Then, we used patients' scores on the verbal subtests and regressed them onto the PIQ.

This generated ideal weights for this sample. We used these weights to predict PIQ scores in the second half of the sample. These were then correlated against actual PIQ scores in the second half of the sample. We also generated weights using the second half of the sample, and used these weights to predict PIQ scores in the first half of the sample. Finally, we repeated this procedure, except we used the performance subtest scores to predict VIQ, splitting the sample in half and generating the same statistics as before. The purpose of these procedures was to see how much variance in factor weights are sample specific, and the amount of shrinkage one can expect when weights are cross-validated. This shrinkage error was then compared to the error introduced by using "unit weighting" rather than "ideal weighting." Results indicate that 98% of the variance accounted for using the ideal weights, after adjustments are made for expected shrinkage, is accounted for by using unit weights. These results support the use of unit weighting as a substitute for ideal weighting.

Nonetheless, we believe that it is important to remind our critics that the RIM methodology leaves it to the clinician to decide which domain structure s/he wants to use for a particular test. For the example patients in our original article (Miller and Rohling, 2001), it was our opinion that a 6-factor cognitive solution best modeled the patients' data. We believe that the results of Tulsky et al. (2003) support such a domain structure for these patients. However, we agree that other patients may be better modeled by a four-factor solution or some other solution. We were not attempting to limit a clinician's ability to make these decisions. Rather, we have suggested a factor structure that a clinician might want to consider, which we believe is reasonably valid for most cases. We elaborated our discussion of individual users' decisions regarding domain structure and test placement in our original paper (See Miller and Rohling, 2001; RIM Step 4, pp. 13–14).

Our critics also wondered about what the RIM recommends with regard to whether a test score should be loaded on multiple factors. The RIM procedure, as currently stated, does not allow test scores to load on multiple factors. Instead, we encourage clinicians to use the best exemplars of particular domains. Clearly, the accuracy of RIM results is enhanced if clinicians use purer measures of the relevant constructs. Yet, as Willson and Reynolds (2003) appropriately point out, good behavioral observations during assessment may, at times, lead a clinician to assign a particular test score to a domain that is atypical. Again, the flexibility of our method allows for this type of clinical judgment. As a caution, however, we strongly recommend that whatever domain structure a clinician decides upon and/or whatever placement of test scores to specific domains is chosen, that the clinician examine the validity of his or her decisions. This can often be done both statistically and by consulting the literature. The RIM heterogeneity statistic is helpful here. In addition, in Step 16 of the RIM, the clinician is instructed to sort the patient's test scores in ascending order, so that they can be inspected for outliers, inconsistencies, and unusual patterns of performance. Finally, we want to explicitly recommend great prudence in "floating" tests around to different factors based on clinical observations. Many psychologists have written about the dangers of post-hoc judgments and decisions (Meehl, 1997). As of yet, we have no data as to the interrater reliability of these types of judgments when various clinicians use the RIM. This will be an important area of future research.

CRITICISM 5: MULTIPLE MEASURES RESULTS IN WORSE ESTIMATES OF DOMAINS

The fifth concern of our critics is that including multiple measures will actually result in worse estimates of a domain. This criticism is valid under two conditions. First, if the clinician knows at the outset which of the multiple measures is the best predictor of a specific construct. Second, if the number of additional measures being combined with the best measure is too small, the composite score will not be able to overcome the error introduced by the addition of measures with low reliability. The degree to which these two conditions operate in real clinical situations are empirical questions. We used data from the 457 patients who had been administered the WAIS-R to address our critics' concern directly. First, to simulate the situation of multiple measures of a single construct we decided to use the subtest scores of the WAIS-R as multiple estimates of the construct of intelligence (i.e., FSIQ). Because we have access to the entire correlational matrix for this example, we know which subtest is the best predictor of the summary measure (i.e., Vocabulary). We also know which subtest is the worst predictor of the summary measure (i.e., Object Assembly).

Furthermore, since there would be bias in using WAIS-R FSIQs that were too far from the mean (i.e., 100), we restricted our sample to those patients whose FSIQ fell between 97 and 103, or .20 standard deviations from the population mean. With this information, we constructed the simulation with the results presented in Table 4 and illustrated in Fig. 3. As can be seen, when the initial and *best* estimator of the construct (i.e., intelligence) was combined with the poorest estimator, the mean estimate still improved for the sample. However, the standard error of the mean difference actually increased. Therefore, the confidence interval around the prediction increased. Furthermore, only 1% of the sample's estimated FSIQ

Table 4. Percent Improvement With Each Estimation Iteration, With No Improvement Equaling 50%

Diff. between 1 predictor and additional predictors	% Improved	Mean absolute value of diff	SEM of improved
1 Predictor (Voc)	—	4.99	.47
2 Predictors (1 + OA)	51	4.42	.49
3 Predictors (2 + BD)	70	3.58	.45
4 Predictors (3 + Sim)	70	3.01	.36
5 Predictors (4 + DSp)	81	2.42	.25
6 Predictors (5 + Info)	79	2.53	.26
7 Predictors (6 + PA)	81	2.11	.21
8 Predictors (7 + Comp)	88	2.06	.18
9 Predictors (8 + DSy)	93	1.31	.14
10 Predictors (9 + Arith)	91	1.36	.12
11 Predictors (10 + PC)	93	1.27	.06

improved when the second estimator was added to the first. The next obvious question then was, "How many more estimators must be added to the composite before the combination of estimators brought the confidence interval below its initial value?" To determine this, we added measures at random, resulting in the order presented in Table 4. Using this WAIS-R data, the answer to our question was, *just one more* estimator was needed before the confidence interval shrank below its initial value. Furthermore, for the most part, the mean estimation continued to improve as each additional estimator was added to the average. This, of course, is a worst-case example of this type of situation; that is the clinician knows the best predictor, a priori, and adds the worst predictor of the construct as the second measure. Even under these conditions, integrating just one additional measure with the initial two measures provided a better mean estimate and confidence interval than did the first measure alone. This makes sense if we think of measures as participants in a research design; that is, adding participants to a sample will likely produce a better estimate of the sample mean (i.e., cognitive construct).

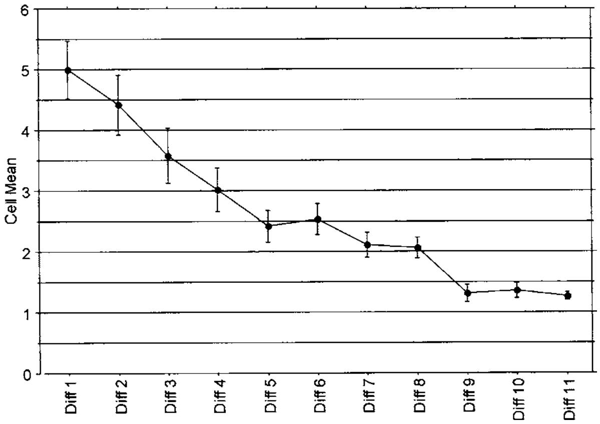

Fig. 3. Estimate of FSIQ based on means of subtest scores, presented as *T* scores.

We remind the reader that the conditions above do not typically exist for most of the domains neuropsychologists commonly assess. Instead, clinicians are not likely to know which of several estimators is the best estimator of a specific construct (i.e., executive functioning may be estimated by results from the Wisconsin Card Sorting Test and/or the Category Test). Therefore, any two or more combinations of estimators is likely to be a better estimate of a patient's true score than any single estimator. Furthermore, this is true whether we are considering the OTBM, EPGA, or any one of the mean cognitive domain summary scores.

CRITICISM 6: PROBLEMS WITH USING EPGA TO ASSESS IMPAIRMENT IN SPECIFIC COGNITIVE DOMAINS

This is the criticism that concerns us the most and, in our opinion, has the most legitimacy. Our critics have brought up an important consideration of statistical regression as it pertains to prediction of another variable. Specifically, both Palmer et al. and Willson and Reynolds expressed concern about the relevance of some of the estimators we recommended be included in the EPGA, how these estimators were transformed to a common metric, and whether this general factor is an accurate estimator of premorbid ability for specific cognitive domains.

We appreciate their concern, understanding that when predicting one variable from another the association between the two variables influences the accuracy of the prediction. For example, Sattler (2001) points out that students' grades' correlate only .43 with their WISC-III FSIQ. Therefore, to predict a specific student's FSIQ using only grades would introduces significant error and will result in a large standard error of the estimate. Clearly, it is better to take into consideration the relationship between variables to better determine the validity of the prediction. When the association between variables is not taken into account, it is functionally equivalent to assuming that the variables are perfectly correlated, which is never the case.

However, as reported in our original paper, our conceptualization of the EPGA is not as a predictor of FSIQ. The EPGA is designed to represent an individual's premorbid *general ability*, a substantial portion of which includes the construct of intelligence, but is not fully encompassed by it (Ardila, 1999; Ardila et al., 1998). Just as the OTBM, DTBM, and ITBM include more constructs than FSIQ (e.g., executive functioning, memory and learning), so too should a premorbid estimate of this construct include more than just FSIQ. Therefore, we purposefully did not recommend that a clinician consider regression to the mean when generating the EPGA from variables such as class rank. If a clinician were to do this, it would narrow the scope of the EPGA and lead to inappropriate conclusions when the premorbid and post-morbid means were compared. This issue was discussed in the WAIS-III and WMS-III Technical Manual when examining the comparison between the "simple-difference method" and the "predicted-difference method." It states

> One noteworthy limitation of the predicted-difference method is that when correlations between measures are low, the range of predicted scores is restricted (i.e., because of regression to the mean). Berk (1984) summarized the disadvantages of the predicted method, pointing out the limitations of imperfect correlations (discrepancies are due to prediction error as well as true differences) . . . The proper use of the simple-difference method requires the examiner to determine first if the difference is statistically significant, and if it is, to determine how frequently a difference of its size occurred in the standardization sample (Berk, 1984). These two steps should be familiar because they are the same ones used to interpret the difference between WISC-III VIQ and PIQ scores (Wechsler, 1991).

We expect to conduct such statistical analyses using several large samples to determine the base rate of these differences. We hope that others will do this as well. Such analyses will generate estimates of the frequency of increasing magnitudes of difference between the EPGA and any specific domain mean within the normal population.

A more important consideration in responding to this criticism is to determine if the RIM results in systematic error in diagnosis. As explained earlier, the statistical procedures we recommended already err on the side of being conservative. For example, assume that a patient's EPGA was calculated, without using a regression approach, to be a *T* score of 40. Furthermore, assume that the patient's post-morbid OTBM was 33. This results in a simple difference of 7 points. If we were to apply some sort of regression equation to generate an EPGA, the score would almost certainly move closer to the sample mean of 50. This might result in an EPGA equal to 43. Consequently, the difference between the EPGA and OTBM will have increased from 7 to 10 points. Therefore, the clinician would be more likely to conclude that the difference was statistically significant. However, a different outcome occurs when the clinician examines the upper tail of the sample distribution. Consider a patient whose EPGA was calculated to be 60 and whose OTBM was calculated to be 53. A regression-based EPGA might result in an estimate of 57. The difference between these two in this case has gone from 7 to 4 points. Thus, in these two examples, the difference between the EPGA and OTBM increased and at the lower tail, whereas it decreased at the upper tail. In the head trauma population, for example, the majority of cases occur at the lower tail of the premorbid distribution (Annegers et al., 1980; Gordon, Mann, & Willer, 1993). Therefore, more often than not, by not using a regression approach we have been statistically more conservative when examining the majority of TBI patients.

Furthermore, the magnitude of error introduced by our methods should be considered. The amount a predicted score regresses to the mean decreases the closer the predictor is to the sample mean. In our opinion, the magnitude of the error introduced becomes significant only when a patient's true premorbid ability was at least one standard deviation away from the sample mean. Considering the systemic error noted above, together with this factor, we estimate that the RIM might introduce significant error for 5% of a sample of TBI patients (i.e., premorbidly very high functioning patients who may have experienced a rather mild TBI).

Considering the alternatives to the RIM recommendations, the literature overwhelming indicates that clinicians are not capable of making such statistical corrections on their own. Any error introduced by the RIM is almost certainly smaller than that which would be introduced by the typical low reliability of clinician judgment. Until better methods of predicting premorbid functioning across a variety of cognitive domains (e.g., executive functioning, memory and learning, attention, and processing speed) have been developed, we believe that it is prudent to continue to use the RIM's current recommendations.

CRITICISM 7: NORMS USED TO GENERATE *T* SCORES MAY COME FROM DISSIMILAR NORMATIVE SAMPLES

Indeed, we are certain that many clinicians who use the RIM will use norms to generate *T* scores that are not equated for many relevant demographic variables (e.g., age, education, gender, and handedness). However, this is not a problem created by following the RIM's recommendations. Rather it is a preexisting problem for clinical neuropsychology. The concern noted by our critics is that when a clinician uses dissimilar normative samples, there will be greater variability in the distribution of scores (i.e., the *SDs* for the OTBM, DTBM, ITBM), and all other domain means will increase from that which might be generated if a clinician used only instruments that were conormed. Appealing to the concept of power analysis, such increased "noise" in the system would then require that more measures be included in a test battery for significant results to be revealed. If additional measures are not added, then the power of the current battery will be low. The clinician will have to consider that the probability of making a Type II error might be unacceptably high (i.e., the clinician will fail to detect a mild head injury if they use too few non-conormed measures in the test battery).

However, a method of determining if inappropriate norms have been used in the RIM analysis is to examine the heterogeneity statistic for the test battery means and for each of the domain means. When inappropriate norms are used to transform raw scores to T scores, the resulting "ballooning" of the variance will cause the heterogeneity statistic to be statistically significant. This will warn the clinician that there might be a problem with the norms being used.

The degree to which noise is introduced into the RIM system, because clinicians choose a battery of non-conormed measures, is an empirical question. To test the influence of the use of different normative groups of limited comparability, we ran analyses on Dataset 2. We chose this dataset because the OTBMs of each patient were generated from 42 dependent variables (i.e., a large test battery was administered). Individuals' data could thus be split into two sets of 21 test variables each. In this way, two independent OTBMs could be created from the same patient. In this example, the 42 variables were split into two based intentionally on the critics' concerns. That is, the dependent variables were separated such that no normative sample was included in both of the OTBM calculations. We then correlated the results of the two calculations, which resulted in a coefficient of .81 and accounted for 66% of the variance. The slope of the regression line was .82 ($SE = .027$) and the intercept was 9.2 ($SE = 1.20$). The mean for OTBM-1 was 45.0 ($SD = 7.3$) and the mean for OTBM-2 was 43.6 ($S = 7.2$). Furthermore, a paired t test found the OTBMs to be significantly different from one another, $t(501) = 7.12$, $p < .0001$; however; the effect size for this difference was small ($g = .20$). These results simulate the worse case scenario (i.e., what might happen if a particular patient were to be assessed by Clinician A using one set of norms to generate T scores, and Clinician B who used an entirely different, and unrelated, set of norms to generate an OTBM for the same patient).

Some readers may think that a coefficient of .81 is sufficiently low to raise concern. However, several factors increase the likelihood of this correlation being smaller than one would find in reality. First, the test battery used has been cut in half, thus reducing the reliability of the original OTBM. Estimating the test–retest reliability coefficient that would have been generated with two OTBMs of 42 variables each would increase the reliability estimate from .82 to .88 using the Spearman–Brown correction. Second, because we deliberately split the data so that the two OTBMs had no overlap in normative samples, our results truly represents a worst-case condition. In practice, most flexible battery clinicians administer several instruments (e.g., WAIS-III), which results in the OTBMs being generated from "conormed" variables. Finally, even when clinicians use different norms, they are often administering the same instruments (e.g., AVLT or RCFT). Because of the nature of our simulation, no instrument used to calculate OTBM-1 was included in our calculation of OTBM-2. This almost certainly increases the disparity between the two OTBMs. It is likely that these three conditions worked together to shrink the relationship between OTBM-1 and OTBM-2. We expect, in most cases with real patients assessed by two different clinicians, that the correlation between the two OTBMs would likely be greater than.90. Substantiating this point, in Heaton et al.'s study (Heaton et al., 2001) of patients suffering from schizophrenia, these authors obtained a test–retest reliability coefficient of .97 for their Global Neurological T-Score. Of course, they most likely were comparing the results of two nearly identical test batteries, rather than our worst-case scenario.

CRITICISM 8: USING THE RIM WILL RESULT IN CLINICIAN'S BEING OVERCONFIDENT

Both critiques suggest that clinicians may inappropriately instill in the RIM greater confidence in its results than it is due, and that this is worse than errors brought about by clinicians' judgment alone. However, this is based primarily on their inaccurate assumptions regarding the use of intraindividual *SD*s, which we have already addressed at some length at the outset of this paper. We hope that our critics will be less concerned about over-confidence now that we have clarified the use of intra-rather than interindividual *SD*s. The degree to which the RIM would cause clinicians to become overconfident in their judgments is also an unanswered empirical question. However, a review of the literature would suggest that, in the absence of empirical methods such as the RIM, clinicians are overconfident of their judgments (e.g., Dawes, Faust, & Meehl, 1989; Meehl, 1997). We would point out that our own experience runs contrary to the concerns of our critics. We have found RIM results to have a tempering effect on our initial clinical judgments! One of the main reasons to use the RIM is that it moves the clinician away from a hard and fast dichotomous judgment (i.e., brain injured or not; malingering or not). Instead, the RIM neuropsychologist can temper his or her findings with statistical probabilities. The RIM neuropsychologist also will be able to give estimates of the power needed to detect differences and will have multiple ways of comparing current performance to estimated premorbid abilities. It is our belief that this will help the clinician to provide a more realistic appraisal of the validity of their findings.

Last, the reviewers seem to be concerned that if we add some mathematical rigor to the diagnostic process, clinician's will be fooled by "pseudoscience." However, clinical abilities remain essential within the RIM. The RIM relies on the clinician's ability to establish rapport with the patient, administering tests in a standardized fashion, and score the tests appropriately. The RIM then helps the clinician avoid many common cognitive errors to which all humans are prone, such as over-focusing on one piece of information, and/or confirmatory bias. In our original paper, we gave an example of how the RIM might detect brain injury that the clinician would otherwise miss. What we did not show, but could have, is that the RIM also works in the opposite direction. That is, at times the analyses will not substantiate a brain injury that the clinician expected to find. In addition, the RIM highlights that it is more difficult to detect brain injury in premorbidly low functioning individuals—thus; a more extensive battery may be required in these situations. Future research is being conducted to demonstrate how the RIM, versus standard clinical practice, detects cognitive impairment across the entire range of premorbid functioning. This direction of research is expected to improve clinical neuropsychological practice to a more thorough integration of science and practice.

CONCLUSION

At the risk of creating a straw man, if the critics' arguments were to be adopted wholesale, one could conclude

1. tests that are not conormed cannot be administered together and interpreted in a meaningful fashion. They imply that the 85% or so of neuropsychologists who use a flexible battery approach are generating uninterpretable results.
2. neuropsychologists can only generate global assessments of neuropsychological functioning, as no available method has generated agreed upon cognitive domains.
3. estimates of premorbid functioning are not valid. Therefore, neuropsychologists can never compare current functioning to some preinjury baseline.
4. clinicians are so "wowed" by mathematics that they will lose all ability to consider how the findings they have obtained might not be valid. There is even some suggestion that if we incorporate decision-making theory and statistics into clinical practice, clinical skills may deteriorate to the level of obsolescence.

Obviously, these are serious concerns that need continued empirical attention within our profession. At the outset, we thank our critics for their very careful review of our work and for their challenge to us to address a variety of concerns with our methodology—from statistical, to status of knowledge in the field, to the role of neuropsychologists as clinicians who, we believe can generate and interpret statistical data about individual patients. Our desire is to make the RIM a user-friendly method to help clinical neuropsychologists truly embody the scientist–practitioner model. We firmly believe that ethical practice in this field requires excellent clinical skills (e.g., in order to generate meaningful patient data). In addition, competent practitioners will follow the literature closely enough to choose reliable instruments to measure the domains of importance to them and their patients. In fact, the RIM is simply trying to make what clinicians already do in their head, and that which is written in their reports, more systematic with improved accuracy in their "calculations." We hope that our responses to the critiques adequately address the concerns generated. Clearly, ours is an evolving field. We believe the RIM has the flexibility to grow with advances in measurement, while currently being an important addition to the practice of all neuropsychologists who employ a flexible battery approach. We are eager to conduct more research to test the validity of these beliefs and our methodology.

REFERENCES

Annegers, J. F., Grabow, J. D., Kurland, L. T., and Laws, E. R., Jr. (1980). The incidence, causes, and secular trends of head trauma in Olmstead County, Minnesota, 1935–1974. *Neurology* **30:** 912–919.

Ardila, A. (1999). A neuropsychological approach to intelligence. *Neuropsychol. Rev.* **9:** 117–136.

Ardila, A., Galeano, L. M., and Rosselli, M. (1998). Toward a model of neuropsychological activity. *Neuropsychol. Rev.* **8:** 171–190.

Berk, R. A. (1984). *Screening and Diagnosis of Children With Learning Disabilities*, Charles C. Thomas, Springfield, IL.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *Am. Psychol.* **34:** 571–582.

Dawes, R. M., and Corrigan, B. (1974). Linear models in decision making. *Psychol. Bull.* **81:** 95–106.

Dawes, R. M., Faust, D., and Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science* **243:** 1668–1674.

Dikmen, S. S., Machamer, J. E., Winn, H. R., and Temkin, N. R. (1995). Neuropsychological outcome at 1-year post head injury. *Neuropsychol.* **9:** 80–90.

Gordon, W. A., Mann, N., and Willer, B. (1993). Demographic and social chatacteristics of the traumatic brain injury model system database. *J. Head Traum. Rehab.* **8:** 26–33.

Green, P., and Allen, L. M. (1995a). *Computer Administered Response Bias*, CogniSyst, Durham, NC.

Green, P., and Allen, L. M. (1995b). *Word Memory Test*, CogniSyst, Durham, NC.

Green, P., and Iverson, G. L. (2001a). Validation of the computerized assessment of response bias in litigating patients with head injuries. *Clin. Neuropsychol.* **15:** 492–497.

Green, P., and Iverson, G. L. (2001b). Effects of injury severity and cognitive exaggeration on olfactory deficits in head injury compensation claims. *Neurorehabilitation*, **16:** 237–243.

Green, P., Lees-Haley, P. R., and Allen L. M. (2002). The Word Memory Test and the validity of neuropsychological test scores. *Journal of Forensic Neuropsychology.* **2:** 95–122.

Green, P., Rohling, M. L., Lees-Haley, P. R., and Allen L. M. (2001). Effort has a greater effect on test scores than severe brain injury in compensation claimants. *Brain Injury* **15:** 1045–1060.

Hartman, D. (2002). The unexamined lie is a lie worth fibbing. Neuropsychological malingering and the Word memory Test. *Arch. Clin. Neuropsychol.* **17:** 709–714.

Heaton, R. K., Gladsjo, J. A., Palmer, B. W., Kick, J., Marcotte, T. D., and Jeste, D. V. (2001). Stability and course of neuropsychological deficits in schizophrenia. *Arch. Gen. Psychiatr.* **58:** 24–32.

Heaton, R. K., Grant, I., and Matthews, C. G. (1991). *Comprehensive Norms for an Expanded Halstead-Reitan Battery: Demographic Corrections, Research Findings, and Clinical Applications.* Psychological Assessment Resources, Odessa, FL.

Heiman, G. W. (2003). *Understanding Research Methods and Statistics: An Integrated Introduction for Psychology* (3rd edn.), Houghton Mifflin, Boston, MA.

Iverson, G. L., and Green, P. (2002). Measuring improvement or decline on the WAIS-R in inpatient psychiatry. *Psychol. Rep.* **89:** 457–462.

Iverson, G. L., Turner, R. A., and Green, P. (1999). Predictive validity of WAIS-R VIQ-PIQ splits in persons with major depression. *J. Clin. Psychol.* **55:** 519–524.

Iverson, G. L., Woodward, T. S., and Green, P. (2001). Base rate of WAIS-R VIQ-PIQ differences in 1593 psychiatric inpatients. *J. Clin. Psychol.* **57:** 1579–1587.

Kaplan, E. F., Goodglass, H., and Weintraub, S. (1983). *The Boston Naming Test* (2nd edn.), Lea and Febiger, Philadelphia.

Lees-Haley, P. R., Smith, H. H., Williams, C. W., and Dunn, J. T. (1996). Forensic neuropsychological test usage: An empirical survey. *Arch. Clin. Neuropsychol.* **11:** 45–51.

Leonberger, F. T., Nicks, S. D., Larrabee, G. J., and Goldfader, P. R (1992). Factor structure of the Wechsler Memory Scale-Revised within a comprehensive neuropsychological battery. *Neuropsychol.* **6:** 239–249.

Matarazzo, J. D., Daniel, M. H., Prifitera, A., and Herman, D. O. (1988). Inter-subtest scatter in the WAIS-R standardization sample. *J. Clin. Psychol.* **44:** 940–950.

Meehl, P. (1997). Credentialed persons, credentialed knowledge. *Clin. Psychol.: Sci. Pract.* **4:** 91–98.

Meyers, J. E., Millis, S. R., and Volkert, K. (2002a). A validity index for the MMPI-2. *Arch. Clin. Neuropsychol.* **17:** 157–169.

Meyers, J. E., Roberts, R. J., Bayless, J. D., Volkert, K. T., and Evitts, P. E. (2002b). Dichotic listening: Expanded norms and clinical application. *Arch. Clin. Neuropsychol.* **17:** 79–90.

Meyers, J. E., and Volbrecht, M. (2003). A validation of multiple malingering detection methods in a large clinical sample. *Arch. Clin. Neuropsychol.* **18:** 261–276.

Miller, L. S., and Rohling, M. L. (2001). A statistical interpretive method for neuropsychological test data. *Neuropsychol. Rev.* **11:** 143–169.

Nunnally, J. C., and Bernstein, I. H. (1994). *Psychometric Theory* (3rd edn.), McGraw-Hill, New York.

Palmer, B. W., Appelbaum, M. I., and Heaton, R. K. (2003). Rohling's interpretive method and inherent limitations on the flexibility of flexible batteries. *Neuropsychol. Rev.* 13.

Pilgrim, B., Meyers, J. E., Bayless, J., and Whetstone, M. (1999). Validity of the Ward Seven-Subtest WAIS-III Short Form in a neuropsychological population. *Appl. Neuropsychol.* **6:** 243–246.

Reitan, R. M., and Wolfson, D. (1985). *The Halstead-Reitan Neuropsychological Test battery: Theory and Clinical Interpretation*, Neuropsychology Press, Tucson, AZ.

Reitan, R. M., and Wolfson, D. (1993). *The Halstead-Reitan Neuropsychological Test Battery: Theory and Clinical Interpretation* (2nd edn.), Neuropsychological Press, Tucson, AZ.

Rohling, M. L., Allen, L. M., and Green, P. (2002a). Who is exaggerating cognitive impairment and who is not? *CNS Spectr.* **7:** 387–395.

Rohling, M. L., Green, P., Allen, L. M., and Iverson, G. L. (2002b). Depressive symptoms and neurocognition in patients passing symptom validity tests. *Arch. Clin. Neuropsychol.* **17:** 205–222.

Rohling, M. L., Langhinrichsen-Rohling, J., and Miller, L. S. (2003a). Actuarial assessment of malingering: The RIM Process. In: Franklin, R. D. (ed.), *Prediction in Forensic and Neuropsychology: Sound Statistical Procedures*, Erlbaum, Associates, Mahwah, NJ.

Rohling, M. L., Meyers, J. E., and Millis, S. (2003b). Neuropsychological impairment following TBI: A dose response analysis. *The Clinical Neuropsychologist*, **17:** 289–302.

Rohling, M. L., Williamson, D. J., Miller, L. S., and Adams, R. (2003c). Using the Halstead Reitan Battery to diagnose brain damage: A comparison of the predictive power of traditional techniques to Rohling's Interpretive Method. *The Clinical Neuropsychologist*, **17:** 531–544.

Sattler, J. M. (2001). *Assessment of Children: Cognitive Applications* (4th edn), Jerome M. Sattler, San Diego, CA.

Sherer, M., and Adams, R. L. (1993). Cross-validation of Reitan and Wolfson's neuropsychological deficit scales. *Arch. Clin. Neuropsychol.* **8:** 429–435.

Sweet, J., Moberg, P., and Suchy, Y. (2000). Ten-Year follow-up survey of clinical neuropsychologists: Part I. Practices and beliefs. *Clin. Neuropsychol.* **14:** 38–55.

Tulsky, D. S., Ivnik, R., Price, L., and Wilkins, C. (2003). Assessment of cognitive functioning with the WAIS-III and WMS-III: Development of a six-factor model. In: Tulsky, D. S., Sakolfske, D. H., Chelvne, G. J., Heaton, R. K., Ivnik, R. J., Bornstein, R., Prifitera, A., and Ledbetter, M. F. (eds.), *Clinical Interpretation of the WAIS-III and WMS-III: Practical Resources for the Mental Health Professional*; Academic Press, San Diego, CA.

Tulsky, D., Zhu, J., and Ledbetter, M. F. (1997). *Wechsler Adult Intelligence Scale-Third Edition and Wechsler Memory Scale-Third Edition Technical Manual*, The Psychological Corporation, San Antonio, TX.

Vanderploeg, R. D., Axelrod, B. N., Sherer, M., Scott, J., and Adams, R. L. (1997). The importance of demographic adjustments on neuropsychological test performance: A response to Reitan and Wolfson (1995). *Clin. Neuropsychol.* **11:** 210–217.

Volbrecht, M. E., Meyers, J. E., and Kaster-Bundgaard, J. (2000). Neuropsychological outcome of head injury using a short battery. *Arch. Clin. Neuropsychol.* **15:** 251–265.

Wechsler, D. (1955). *Manual for Wechsler Adult Intelligence Scale*, Psychological Corporation, New York.

Wechsler, D. (1981). *Manual for the Weschsler Adult Intelligence Scale—Revised*, Psychological Corporation, New York.

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale-Third Edition*, The Psychological Corporation, San Antonio TX.

Wedding, D., and Faust, D. (1989). Clinical judgment and decision making in neuropsychology. *Arch. Clin. Neuropsychol.* **4:** 233–265.

Willson, V. L., and Reynolds, C. R. (2003). A critique of Miller and Rohling's statistical interpretive method for neuropsychological test data. *Neuropsychol. Rev.* **13**.