JOURNAL OF PERSONALITY ASSESSMENT, *81*(1), 104–110
Copyright © 2003, Lawrence Erlbaum Associates, Inc.

Enhancing Accuracy in Observational Test Scoring: The Comprehensive System As a Case Example

Robert E. McGrath
School of Psychology
Fairleigh Dickinson University

Inaccuracies in administration and scoring can potentially compromise the validity of any standardized psychosocial measure. The threat is particularly pertinent to methods involving behavioral observation, a category that includes many intelligence tests, neuropsychological measures, personality assessment instruments, and diagnostic procedures. Despite evidence and conjecture that errors in testing procedure are common for at least some of these measures and that these errors are often severe enough to influence interpretation, the topic has received relatively little attention. In particular, the absence of any safeguard against inaccurate test use in clinical situations can put the respondent at risk and violates ethical standards for the use of tests. In this article, I review some issues surrounding accuracy in testing procedures, including a discussion of what is known about the problem, an evaluation of several approaches to improving testing practices, and a review of recommendations for the statistical evaluation of rater accuracy. In this article, I use the Rorschach Comprehensive System (Exner, 1993) to demonstrate the concepts discussed.

The failure to adhere to established standards for administration and scoring represents a potential threat to test validity that has received relatively little attention in the testing literature. For example, clinicians often tacitly assume that multiple-choice measures are relatively immune to this problem because interaction between the tester and respondent relevant to test interpretation is minimized during administration and because scoring is reduced to a series of algorithms. Even in such highly structured circumstances although, meaningful errors in hand scoring can occur. Allard and Faust (2000) found surprisingly high rates of errors across clinical settings in the scoring of the Minnesota Multiphasic Personality Inventory (Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989), Beck Depression Inventory (Beck, Steer, & Brown, 1996), and State–Trait Anxiety Inventory (Spielberger, 1983) despite their very structured scoring systems (see also Allard, Butler, Faust, & Shea, 1995; Simons, Goddard, & Patton, 2002). However, error rates were reduced in settings where scoring accuracy was emphasized.

The problem of administration and scoring accuracy is compounded when the assessment is based on observations of the respondent's semantic or motoric behavior. Many of the most popular psychological assessment instruments in use today are observational rating scales, including measures of personal style such as the Rorschach and Thematic Apperception Test (Murray, 1943); individually administered intelligence and psychoeducational tests; neuropsychological instruments; and diagnostic procedures. Such measures typically require human intervention in both the administration and scoring, increasing the potential for error when compared with multiple-choice measures.

In some cases, such as the Information and Arithmetic subscales from the Wechsler intelligence tests (Wechsler, 1997), the range of appropriate responses to the test stimuli is sufficiently bounded such that the error rate in administration and scoring should theoretically still be comparable with that of multiple-choice measures, assuming appropriate care is taken. In many cases though, the universe of responses associated with any particular scoring alternative is theoretically infinite. Furthermore, it is often the case that administrative protocol varies depending on the respondent's behavior, as when the tester is required to clarify the response if the response's scoring is equivocal. The complexities in administration and scoring created by these circumstances dramatically increase the potential for test misuse.

Errors in administration and scoring are not always trivial in their impact. This issue has been examined most extensively in relation to the scoring of the Wechsler Intelligence scales. For example, Slate, Jones, Murray, and Coulter (1993) found scoring errors in every one of 50 Wechsler Adult Intelligence Scale–Revised (Wechsler, 1981) protocols they reviewed, although the testers all had extensive ex-

perience with the instrument. In two cases, the errors would have resulted in withholding of services for which the respondent was eligible (see also Klassen & Kishor, 1996; Slate & Jones, 1990; Slate, Jones, Coulter, & Covert, 1992; Whitten, Slate, Jones, & Shine, 1994).

In the field of personality assessment, the testing accuracy of observational measures has been discussed most recently in relation to the Comprehensive System for the Rorschach (Exner, 1993). The issue was first raised by Wood, Nezworski, and Stejskal (1996) and discussed in more detail by Hunsley and Bailey (1999, 2001). Although acknowledging that what they referred to as field reliability is an issue for observational instruments in general, Hunsley and Bailey (1999) listed several reasons why they thought it was of particular importance for the Comprehensive System. First, even among observational measures, Comprehensive System coding is particularly complex. Second, a tradition of alternate scoring systems and idiosyncratic scoring may render users of the Rorschach more likely to deviate purposefully from standardized scoring. However, Hunsley and Bailey (1999, 2001) provided no evidence to suggest that idiosyncrasies in scoring remain widespread, and it is likely that inadvertent errors are the more serious problem.

It is clear that human involvement in the administration and scoring of psychological measures creates the potential for error regardless of the type of measure and that this threat can impact on the validity of both research designs and clinical evaluations. The purpose of this article is to focus greater attention on this issue and to discuss some approaches to protecting against inaccuracies in testing. For illustrative purposes, the Rorschach Comprehensive System is used as a case example. This choice is not intended to imply the problem is particularly acute for the Comprehensive System because there is no empirical evidence that such is the case. In fact, if awareness of the potential for testing inaccuracies serves to mitigate the problem, testing inaccuracies in field use may be less of a problem for the Comprehensive System at present than for other complex observational methods, thanks to the attention focused on the problem by critics as well as the publication of two references specifically intended to reduce the rate of testing inaccuracies (Exner, 2001; Viglione, 2002). The Comprehensive System is an appropriate choice primarily because the issue has been discussed in greater detail than is true of most personality assessment instruments.

To date, three studies have investigated the issue of accuracy in Rorschach scoring. The first was an unpublished investigation in which clinicians who had completed training in the Comprehensive System were asked to code responses for which the correct coding had been determined; results were disappointing (discussed by Wood et al., 1996). A second unpublished study by McKinzey and Campagna (2002) asked 30 experienced Rorschach users to score the same protocol. Although the study was methodologically flawed, one finding was unequivocally relevant to the issue of scoring accuracy. McKinzey and Campagna reported that 19 of 27 scorers (70%) made errors in the computation of summative scores based on their own coding of the responses.

The only published study of this issue was completed by Guarnaccia, Dill, Sabatino, and Southwick (2001) who asked users of the Rorschach to code a series of responses for which a correct coding had previously been established. They concluded that coding accuracy was generally unacceptable. The accuracy of coding Special Scores was particularly poor, a finding that may be expected because the identification of the behaviors meriting a Special Score can be difficult. However, the authors used an unusual approach to gauge accuracy in which credit was deducted for errors of omission. The results therefore cannot be compared to typical standards for adequate agreement rates.

Two more studies have investigated the reliability of scoring in field settings. In an as-yet unpublished study, Pogge et al. (2002) examined interrater reliability for a sample of Rorschach protocols completed by adolescent inpatients in which the first scoring was completed without an expectation that the results would later be reviewed for research purposes. Meyer et al. (2002, Sample 4) conducted a similar analysis with protocols completed by adult inpatients. In both cases, the results indicated field reliability of the protocols was more than adequate. However, neither of these studies established the level of consistency between field raters and correct scoring of the responses.

The lack of research concerning field accuracy in the scoring of standardized observational measures is particularly disturbing when one considers the potential impact of a psychological assessment on the respondent. Forensic assessors rely heavily on neuropsychological and other standardized observational measures to make recommendations on issues as weighty as child custody, disability claims, and incarceration status (Lees-Haley, 1992; Martin, Allan, & Allan, 2001). Even outside of forensic settings, the results of psychological assessments can have profound implications for the respondent. Although it represents normal operating procedure, the absence of safeguards to ensure accurate testing increases the potential for inadvertent violations of Standard 9.02, "Use of Assessments," from the *Ethical Principles for Psychologists and Code of Conduct* (American Psychological Association, 2002).

The responsibility to guard against incorrect test use also applies to research settings, particularly when the results have potential clinical implications. Accuracy in administration and scoring is a prerequisite for the norming and validation of standardized observational measures, although many test manuals for observational measures do not even address whether there were safeguards of procedural accuracy incorporated into the developmental research.

The remainder of this article is devoted to two topics having to do with improving accuracy in testing procedures. The first is an evaluation of approaches to improving the accuracy of testing in both clinical and research settings, with

some recommendations for practice. The second is a review of statistical methods for evaluating accuracy when compared to a standard, or correct, scoring. Given the degree to which the Rorschach Comprehensive System has been the focus of this issue in the area of personality assessment, I use it to demonstrate the issues involved. It is important to bear in mind though that these issues apply equally well to most assessment instruments based on observations.

ENSURING TESTING ACCURACY

There are several approaches possible for ensuring accuracy in testing procedures. In research settings, common practice demands the demonstration of an adequate level of interrater reliability for observational measures. Interrater reliability is a useful technique for identifying certain impediments to accurate scoring. Inadequately trained testers reduce interrater reliability. So would *rater drift,* the tendency of raters to modify scoring rules over time (Smith, 1986). There is in fact clear evidence that the Rorschach is capable of reliable scoring even in field settings (Meyer et al., 2002; Pogge et al., 2002).

However, adequate interrater reliability is not sufficient to ensure accuracy in scoring because it is possible for "local" groups (members of the same research team) to share certain misunderstandings about coding rules. This would particularly be the case if all members of the team shared training experiences, as is often the case. The existence of shared local inaccuracies can reduce testing accuracy without affecting interrater reliability.

One must wonder whether such local variations are in fact widespread in testing. Unfortunately, there is no research currently available on this topic. A pattern of research findings indicating that the accuracy of clinicians is poor, whereas their reliability is good, would provide indirect evidence of local drift. Geographic variation in the distribution of scores might also be attributable to local practice variations.

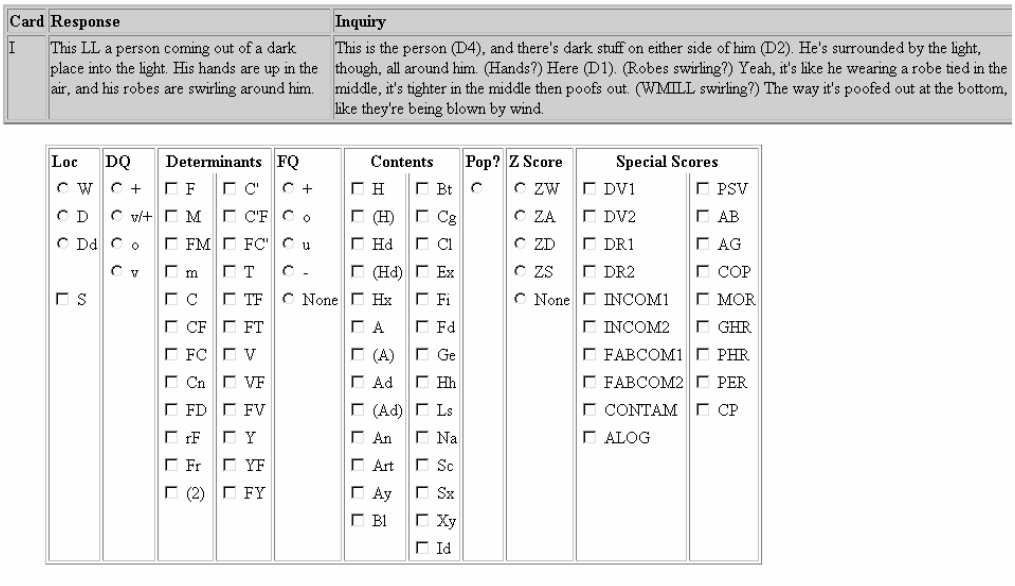
There is research indicating that such local variations are evident in medical decision making. Wennberg and Gittelsohn (1982) demonstrated local variations in surgical decision making. Such variations were particularly extreme for procedures in which the criteria used for making the decision are not definitive. For example, the circumstances under which a tonsillectomy is advisable is a topic of debate among physicians, and across regions there was a six-fold difference in the relative frequency of the procedure. In contrast, Wennberg and Gittelsohn found relatively little variability in the rate of surgical correction for inguinal hernia, a condition that is easily identified and for which surgery is accepted as the treatment of choice. Given these findings, it would be reasonable to hypothesize that the potential for local variations in administration and scoring increases as the complexity of the test increases.

A second approach to guaranteeing accuracy would involve rescoring by individuals who did not share training experiences with the first scorer. There are unfortunately practical and ethical obstacles to this solution. One is the confidentiality issue created by sharing test data. Another is the practical problem involved in getting other assessors to agree to such an arrangement and the delay that would result while data are transferred and the second scoring is completed. Finally, the evaluation of interrater reliability does not provide an adequate basis for addressing inaccuracies in administration. Such inaccuracies may not be evident from the raw data. Even if they are, identification of the problem does not occur until after the test is administered, when the damage is already done.

A third approach would involve the development of a credentialing system for competency in test administration and scoring. The following represents one possible scenario for this system. Once a year an examination would be developed and posted online. The examination would include questions having to do with administration, although the proposed approach is more effective as a method of ensuring scoring accuracy than it is as a method of ensuring administration accuracy. In addition, 20 to 30 responses distributed across the cards would be presented for scoring. Responses could be chosen to approximate the distribution of codes in the general population as indicated by Comprehensive System normative data (Exner, 1993). Figure 1 provides an example of what an individual Web page might look like. Responses should reflect a range of difficulty in scoring. A fairly simple script could be used to grade the results. Passing could be set at 80% correct. A review of the Comprehensive System scoring criteria indicates there are 60 forced-choice coding decisions (see Table 1); therefore, passing a 20-response examination would require correct decisions for 960 of the coding decisions. A certificate could be generated online for participants who meet the criterion. There would be some costs associated with the development of this system, mainly for consensus scoring of the protocols by acknowledged experts in the Comprehensive System. These costs could easily be supported by participants, particularly if continuing education credits were awarded for successfully completing the examination.

This provides only a brief overview of the process, and implementation would require consideration of several other issues. Certain scoring decisions have no impact on interpretation in the Comprehensive System. For example, the choice between a TF and FT has no effect on the Structural Summary. This raises the question of whether the failure to make such discriminations correctly are relevant to competence. The existence of scoring decisions without interpretive significance is probably a relatively rare phenomenon in the universe of observational measures, however.

Another issue to be considered is the appropriate criterion for competence. The question encompasses the selection of a statistical approach to estimating competence as well as the identification of an optimal cut score. A variety of procedures are available for scoring competency tests. Options in-



Card	Response	Inquiry
I	This LL a person coming out of a dark place into the light. His hands are up in the air, and his robes are swirling around him.	This is the person (D4), and there's dark stuff on either side of him (D2). He's surrounded by the light, though, all around him. (Hands?) Here (D1). (Robes swirling?) Yeah, it's like he wearing a robe tied in the middle, it's tighter in the middle then poofs out. (WMILL swirling?) The way it's poofed out at the bottom, like they're being blown by wind.

Loc	DQ	Determinants		FQ	Contents		Pop?	Z Score	Special Scores	
○ W	○ +	☐ F	☐ C'	○ +	☐ H	☐ Bt	○	○ ZW	☐ DV1	☐ PSV
○ D	○ v/+	☐ M	☐ C'F	○ o	☐ (H)	☐ Cg		○ ZA	☐ DV2	☐ AB
○ Dd	○ o	☐ FM	☐ FC'	○ u	☐ Hd	☐ Cl		○ ZD	☐ DR1	☐ AG
	○ v	☐ m	☐ T	○ -	☐ (Hd)	☐ Ex		○ ZS	☐ DR2	☐ COP
☐ S		☐ C	☐ TF	○ None	☐ Hx	☐ Fi		○ None	☐ INCOM1	☐ MOR
		☐ CF	☐ FT		☐ A	☐ Fd			☐ INCOM2	☐ GHR
		☐ FC	☐ V		☐ (A)	☐ Ge			☐ FABCOM1	☐ PHR
		☐ Cn	☐ VF		☐ Ad	☐ Hh			☐ FABCOM2	☐ PER
		☐ FD	☐ FV		☐ (Ad)	☐ Ls			☐ CONTAM	☐ CP
		☐ rF	☐ Y		☐ An	☐ Na			☐ ALOG	
		☐ Fr	☐ YF		☐ Art	☐ Sc				
		☐ (2)	☐ FY		☐ Ay	☐ Sx				
					☐ Bl	☐ Xy				
						☐ Id				

FIGURE 1 Sample Web page for evaluating competency in Comprehensive System coding.

TABLE 1
Forced-Choice Coding Decisions

Loc: W, D, Dd	(Hd)	Sc
S	Hx	Sx
DQ: +, o, v, v/+	A	Xy
M	(A)	Id
FM	Ad	Pop
M	(Ad)	Z score: ZW, ZA, ZD, ZS
ap: a, p, a–p	An	DV: Lv1, Lv2
C: FC, CF, C, Cn	Art	INCOM: Lv1, Lv2
C': FC', C'F, C'	Ay	DR: Lv1, Lv2
T: FT, TF, T	Bl	FABCOM: Lv1, Lv2
V: FV, VF, V	Bt	ALOG
Y: FY, YF, Y	Cg	CONTAM
r: Fr, rF	Cl	AB
FD	Ex	AG
F	Food	COP
FQ: +, o, u, –, none	Fi	CP
Pair	Ge	Human representation: GHR, PHR
H	Hh	MOR
(H)	Ls	PER
Hd	Na	PSV

Note. In all cases except Loc and DQ, "absent" was also a scoring option. All symbols are taken from Exner (1993).

clude differential weighting of errors depending on their criticality for interpretation, for example, failure to identify a Human movement response would be weighted more heavily than failure to identify a Food response. It is also worth considering the use of the reliability of ratings as an index of competency rather than percent correct by setting the criterion for competence at some value for the mean kappa coefficient. Although percent correct is more consistent with general practice in competency testing, it is a problematic measure of accuracy (Wood et al., 1996). Although some details remain, the proposal as presented here encompasses the key elements of a credentialing system.

Krishnamurthy (2001) raised several concerns about credentialing as a method of dealing with potential errors in test use. Although the program would initially be voluntary, Krishnamurthy pointed out that over time, managed care organizations may come to consider it a requirement. One might suspect the same would occur in forensic settings. The transition from voluntary to mandatory participation has in fact been a general pattern with proficiency testing. For example, when the Clinical Laboratory Improvement Act was first passed in 1967, it applied only to laboratories engaging in interstate commerce. When revised in 1988, it was extended to all clinical laboratories in the country. Similarly, accreditation of forensic laboratories began in 1978 as a voluntary program, but by 1981 had become mandated by the American Society of Crime Laboratory Directors (Peterson & Markham, 1995).

One may argue in response that mandatory credentialing for those who claim a special proficiency in a test instrument is not a particularly bad thing. Given the types of decisions made on the basis of psychological tests, requiring credentialing of practitioners who engage in testing should result in an improvement in the quality of service provided to the public. Given the implementation of a convenient, reasonably priced credentialing procedure as would be possible online, the benefits of mandatory credentialing should outweigh any costs.

A second concern raised by Krishnamurthy (2001) is that introducing proficiency testing for the Comprehensive System potentially stigmatizes the Rorschach as requiring special precautions. This hypothesis can be countered in two ways. First, critics of the Rorschach are just as likely to object to a certifying system because it potentially creates an aura of validity for the method. Second, it is important to keep in mind that the Rorschach is only being used as a case

example of a procedure that should be applied to all major observational systems. In fact, Campbell (2000) recently suggested the need for widespread training in the use of Hare Psychopathy Checklist (Hare, 1991).

Once the credentialing system is in place, each research team studying the Comprehensive System could include at least one rater who had passed the exam within the last year. That individual would in turn become the "gold standard" for the training and calibration of other members of the team. Studies of interrater reliability may then technically be more akin to studies of scoring accuracy.

STATISTICAL EVALUATION OF SCORING ACCURACY

Research teams interested in the validation or norming of the Rorschach could limit administration and scoring to individuals who have received credentialing in the past year. This may not always be practical when sample sizes are sufficiently large. Alternatively, one member of the team who is credentialed can serve as the standard against which the accuracy of raters may be judged (in this context, the term *rater* is used to refer to individuals who are not credentialed as a competent Comprehensive System scorer to distinguish them from *standards*). This suggestion raises the question of how best to judge the reliability of raters with a standard or correct scoring, what could be termed *rater-standard reliability* rather than interrater reliability.

Despite the important distinction, the statistics generally considered appropriate for rater-standard reliability are the same as those used for evaluating interrater reliability, although they are used in slightly different ways when there is more than one rater. Light (1971) concluded that the best statistic for estimating reliability with a standard is chance-corrected agreement, or kappa, when the variable is categorical (as is true of Rorschach response codes). By extension, the intraclass correlation coefficient would serve the same purpose when the variable is quantitative (as is true of Rorschach summative scores). In cases in which there is one standard and multiple raters, however, overall reliability is based on comparisons between the standard and each of the raters; comparisons between the raters themselves need not be considered (Berry & Mielke, 1997; Light, 1971). Given that the mean reliability of the raters with the standard would likely be higher than the mean reliability among the raters, it may well be that overall reliability estimates would be higher than is the case when the raters are all treated equally.

Berry and Mielke (1997) provided a general formula for computing reliability between a standard and raters that is applicable regardless of the number of raters and ratings and the level of measurement represented by the ratings. If there are m raters making r separate ratings of n objects, then the reliability of raters with the standard (symbolized as s) may be estimated by:

$$1-\frac{\frac{1}{n}\sum_{i-1}^{m}\sum_{j=1}^{n}\left[\sum_{k=1}^{r}(x_{sjk}-x_{ijk})^2\right]^{1/2}}{\frac{1}{n^2}\sum_{i-1}^{m}\sum_{j=1}^{n}\left[\sum_{k=1}^{r}(x_{sjk}-x_{ijk})^2\right]^{1/2}}.$$

This formula is applicable to both dimensional and categorical ratings.[1] A summary of guidelines for estimating reliability with and without a standard may be found in Table 2.

FINAL THOUGHTS

The critics of the Rorschach should be acknowledged for focusing attention on the issue of testing accuracy as it relates to personality assessment. However, in restricting their analysis to the Rorschach, they missed the more important general issue. All tests are susceptible to error, and the more complex the test, the greater its susceptibility. Unfortunately, there is no evidence concerning the degree to which the practice of psychological assessment is flawed by inaccuracies in test administration and scoring. Indirect evidence does suggest that such errors occur regularly, even in high-stakes cases. Research with the Wechsler Intelligence tests suggests that incorrect administration is widespread and in a small subset of cases can have a significant impact on decision making. McKinzey and Campagna (2002) presented a Rorschach protocol from an evaluation of a death-row inmate so poorly administered that 10% of individuals asked to score the protocol spontaneously declared it incapable of scoring, and those who attempted scoring did not even agree on the number of responses provided.

McKinzey and Campagna (2002) made the error of blaming these problems on the test rather than the tester. Any activity that requires human involvement is susceptible to human error (Garb, 1998). The risk of testing inaccuracies is inevitable, and there is probably no way to eliminate the problem completely short of removing the clinician from the testing process. However, there are ways to minimize the risk.

Although the extent to which tests are used incorrectly is unclear, what is clear is that incorrect test use occurs, with potentially serious consequences for the respondent (Wakefield & Underwager, 1993). No standardized observa-

[1]Berry and Mielke (1997) indicated how to obtain a FORTRAN–77 subroutine that generates the various statistics discussed in their article. I have also written a stand-alone program that generates most of the same statistics. Copies of the latter are available by contacting Robert E. McGrath.

TABLE 2
Evaluating the Reliability of Scoring

	Categorical Variables	*Quantitative Variables*
2 raters	Kappa	ICC
1 rater, 1 standard	Kappa	ICC
> 2 raters	Kappa for multiple raters, or mean kappa for all rater pairs	ICC
> 1 rater, 1 standard	Berry & Mielke (1997) statistic, or mean kappa between standard and raters	Berry & Mielke (1997) statistic, or mean ICC between standard and raters

Note. ICC = intraclass correlation coefficient.

tional measure is immune to potential misuse. If we as psychologists are to participate in the process of making important decisions about the people we test, we have an ethical obligation to ensure that test administration and scoring are conducted fairly and accurately. To date, we have not adequately addressed this obligation.

ACKNOWLEDGMENTS

Portions of this article were previously presented at the March 2001 midwinter Meeting of the Society for Personality Assessment, Philadelphia. I am grateful to Steven D. Hickman, Radhika Krishnamurthy, and the members of the Rorschach listserve for the various roles they played in the genesis of this article.

REFERENCES

Allard, G., Butler, J., Faust, D., & Shea, M. T. (1995). Errors in hand scoring objective personality tests. *Professional Psychology: Research & Practice, 26,* 304–308.

Allard, G., & Faust, D. (2000). Errors in scoring objective personality tests. *Assessment, 7,* 119–129.

American Psychological Association. (2002). Ethical principles of psychologists and code of conduct. *American Psychologist, 57,* 1060–1073.

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *The Beck Depression Inventory–Second edition.* San Antonio, TX: Psychological Corporation.

Berry, K. J., & Mielke, P. W., Jr. (1997). Measuring the joint agreement between multiple raters and a standard. *Educational and Psychological Measurement, 57,* 527–530.

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

Campbell, T. W. (2000). Sexual predator evaluations and phrenology: Considering issues of evidentiary reliability. *Behavioral Sciences and the Law, 18,* 111–130.

Clinical Laboratory Improvement Act, Pub. L. No. 90–174. (1967).

Clinical Laboratory Improvement Act–Revised, Pub. L. No. 100–578. (1988).

Exner, J. E., Jr. (1993). *The Rorschach: A Comprehensive System: Vol. 1. Basic foundations* (3rd ed.). New York: Wiley.

Exner, J. E., Jr. (2001). *A Rorschach workbook for the Comprehensive System* (5th ed.). Asheville, NC: Rorschach Workshops.

Garb, H. (1998). *Studying the clinician: Judgment research and psychological assessment.* Washington, DC: American Psychological Association.

Guarnaccia, V., Dill, C. A., Sabatino, S., & Southwick, S. (2001). Scoring accuracy using the Comprehensive System for the Rorschach. *Journal of Personality Assessment, 77,* 464–474.

Hare, R. D. (1991). *The Hare Psychopathy Checklist–Revised.* Toronto, Ontario, Canada: Multi-Health Systems.

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

Hunsley, J., & Bailey, J. M. (2001). Whither the Rorschach? An analysis of the evidence. *Psychological Assessment, 13,* 472–485.

Klassen, R. M., & Kishor, N. (1996). A comparative analysis of practitioners' errors on WISC–R and WISC–III. *Canadian Journal of School Psychology, 12,* 35–43.

Krishnamurthy, R. (2001, March). Practical issues and logistics of having a gold standard. In S. D. Hickman (Chair), *Exploring the implications of a Rorschach coding "gold standard."* Symposium conducted at the midwinter meeting of the Society for Personality Assessment, Philadelphia.

Lees-Haley, P. R. (1992). Psychodiagnostic test usage by forensic psychologists. *American Journal of Forensic Psychology, 10,* 25–30.

Light, R. J. (1971). Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin, 76,* 365–377.

Martin, M., Allan, A., & Allan, M. M. (2001). The use of psychological tests by Australian psychologists who do assessments for the courts. *Australian Journal of Psychology, 53,* 77–82.

McKinzey, R. K., & Campagna, V. (2002, April 27). *The Rorschach, Exner's Comprehensive System, interscorer agreement, and death.* Retrieved November 22, 2002, from http://home.earthlink.net/~rkmck/vault/Rorinterscor/McKCam02.pdf

Meyer, G. J., Hilsenroth, M. J., Baxter, D., Exner, J. E., Jr., Fowler, J. C., Piers, C. C., et al. (2002). An examination of interrater reliability for scoring the Rorschach Comprehensive System in eight data sets. *Journal of Personality Assessment, 78,* 219–274.

Murray, H. A. (1943). *Thematic Apperception Test: Manual.* Cambridge, MA: Harvard University Press.

Peterson, J. L., & Markham, P. N. (1995). Crime laboratory proficiency testing results, 1978–1991. I: Identification and classification of physical evidence. *Journal of Forensic Sciences, 40,* 994–1008.

Pogge, D. L., McGrath, R. E., Stokes, J. M., Cragnolino, A., Zaccario, M., Hayman, J., et al. (2002). *Comprehensive System scoring reliability in an adolescent inpatient sample.* Manuscript submitted for publication.

Simons, R., Goddard, R., & Patton, W. (2002). Hand-scoring error rates in psychological testing. *Assessment, 9,* 292–300.

Slate, J. R., & Jones, C. H. (1990). Examiner errors on the WAIS–R: A source of concern. *Journal of Psychology, 124,* 343–345.

Slate, J. R., Jones, C. H., Coulter, C., & Covert, T. L. (1992). Practitioners' administration and scoring of the WISC–R: Evidence that we do err. *Journal of School Psychology, 30,* 77–82.

Slate, J. R., Jones, C. H., Murray, R. A., & Coulter, C. (1993). Evidence that practitioners err in administering and scoring the WAIS–R. *Measurement and Evaluation in Counseling and Development, 25,* 156–161.

Smith, G. A. (1986). Observer drift: A drifting definition. *Behavior Analyst, 9,* 127–128.

Speilberger, C. D. (1983). *Manual for the State–Trait Anxiety Inventory (STAI).* Palo Alto, CA: Consulting Psychologists Press.

Viglione, D. J. (2002). *Rorschach coding solutions: A reference guide for the Comprehensive System.* San Diego, CA: Author. Available at http://www.geocities.com/donaldviglione

Wakefield, H., & Underwager, R. (1993). Misuse of psychological tests in forensic settings: Some horrible examples. *American Journal of Forensic Psychology, 11,* 55–75.

Wechsler, D. (1981). *Manual for the Wechsler Adult Intelligence Scale–Revised (WAIS–R).* San Antonio, TX: Psychological Corporation.

Wechsler, D. (1997). *WAIS–III administration and scoring manual*. San Antonio, TX: Psychological Corporation.

Wennberg, J., & Gittelsohn, A. (1982, April). Variations in medical care among small areas. *Scientific American, 246*(4)*,* 120–134.

Whitten, J., Slate, J. R., Jones, C. H., & Shine, A. E. (1994). Examiner errors in administering and scoring the WPPSI–R. *Journal of Psychoeducational Assessment, 12,* 49–54.

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

Robert McGrath
School of Psychology
T–WH1–01
Fairleigh Dickinson University
Teaneck NJ 07666
E-mail: mcgrath@fdu.edu

Received April 5, 2002
Revised February 18, 2003