)URNAL OF PERSONALITY ASSESSMENT, 71(2), 195–211

ARTICLES

Error in Research and Assessment Data With an Erratum for Meyer (1993)

Gregory J. Meyer
Department of Psychology
University of Alaska Anchorage

After reviewing literature detailing the ubiquity of error in research and assessment data, I describe mistakes found with MMPI–2 and Rorschach scores in an earlier publication (Meyer, 1993). The mistakes emerged from several sources, including hand-scoring the assessment measures, manually retrieving scores from patient files, and manually entering information for statistical analyses. The proportion of erroneous scores in the original study (MMPI–2 = 5.88%, Rorschach = 1.56%) was relatively small and reliability coefficients between the original and corrected scales were uniformly high (i.e., > .940). Consequently, the findings detailed in the original publication were not greatly affected by the mistakes that were made. Nonetheless, because this will not always be the case, I discuss the discrepancies as a way to sensitize researchers, clinicians, and students to the presence of error and then suggest several strategies for minimizing its impact on assessment data and research.

Error is ubiquitous in human affairs. At times, it may be expressed as nothing more than forgetting to pick up an item from the store or carrying the wrong number while balancing a checkbook. At other times, the manifestations may be much more dramatic. The decision to launch the space shuttle Challenger despite a "thermally distressed" O-ring is a vivid reminder (Hammond, 1996). Although not broadcast live to millions of television viewers, the recent death of three experienced skydivers who perished at the South Pole under clear skies provides an equally dramatic illustration (Enfield, 1998). Two never even attempted to open their chutes and the chute for the third never had time to deploy fully. All hit the ground at terminal velocity, burrowing several feet into the crust of snow and ice.

Closer to home, there are a vast number of ways in which omission and commission errors can affect the final results of research; often dramatically. The most ex-

treme form of error in research would be an instance of deliberate fraud, in which data are altered or even completely fabricated. Fortunately, this type of error is notable for its rarity (cf. Kazdin, 1998). Other more common types of research error result from the genuine complexity of the enterprise in combination with a lack of sophistication on the part of investigators. Obviously, the goal of education is to train researchers to avoid these kinds of error and there are many excellent resources that provide thorough discussions of statistical, methodological, measurement, and experimenter-induced pitfalls, as well as strategies for avoiding them. Many of these works are relevant to research in general (e.g., Behrens, 1997; Campbell & Fiske, 1959; Cohen, 1988, 1990, 1994; Cook & Campbell, 1979; Kazdin, 1998; Nisbett & Wilson, 1977; Rosenthal & Rubin, 1978; Schmidt, 1992, 1996; Schmidt & Hunter, 1996; Shedler, Mayman, & Manis, 1993; Wampold, Davis, & Good, 1990; Wilcox, 1998; Zuckerman, Hodgins, Zuckerman, & Rosenthal, 1993), whereas others provide guidance that more specifically addresses personality assessment research (e.g., Butcher, Graham, Haynes, & Nelson, 1995; Cronbach & Meehl, 1955; Exner, 1995; Masling, 1960/1992; Meyer, 1996, 1997; Schinka & Greene, 1997; Wiggins, 1973/1988). A final type of research error emerges from unintentional lapses or mistakes during the collection, recording, transfer, analysis, and publication of data. These clerical errors or data mistakes are the primary focus of this article.

The research literature is replete with instances in which data mistakes have been identified after a study was published. For instance, a search of PsycINFO using the term *erratum* produced 1,305 documents published since 1985. An equivalent MEDLINE search produced 28,495 published errata during the same time period. These figures do not quantify the number of published retractions or corrections that entered the literature without being designated as an erratum. In addition, because errata typically correct obvious clerical mistakes, such as mislabeled table columns, transposed numbers, omitted citations, misspelled names, or misspecified p values, these figures grossly underestimate the extent to which more general errors pervade the literature, such as Type II errors from low statistical power, chance findings that result from sampling error, or misleading conclusions that result from measuring variables with just a single method of assessment.

In one of the more dramatic instances of clerical error, the 1950 census reported a remarkably high proportion of 14-year-olds who were widowed or divorced, as well as unexpectedly large populations of Native Americans living in certain regions of the country (Arndt, Tyrrell, Woolson, Flaum, & Andreasen, 1994). Although these anomalies were not discovered until well after the census data were published, they resulted from a single formatting glitch that caused the data columns from certain census centers to shift by one position, turning those who were 42-year-old "heads of household" into 14-year-old widows and turning their children into Native Americans.

Of course, all clerical errors in the published literature are not recognized or corrected. In an early report, Wolins (1962) described efforts to obtain raw data from 37

published studies. Only seven authors provided data for analysis. Of these, Wolins found three studies (43%) that had published results containing "gross errors" of analysis. More recently, Rossi (1987) examined 46 statistical tests reported in psychology journal articles. Using descriptive data that corresponded to each original *F* or *t* test, he found approximately 25% of the reported statistical values were wrong by a margin of 20% or more. In about 7% of the cases, the published statistics were off by more than 50% and about 13% of the values that had been reported as statistically significant actually were not. Several reviews of the journal literature have reported that up to 50% of published articles contain at least one statistical flaw (see Murray, 1988).

Meta-analytic researchers regularly discover statistical outliers in a body of literature that almost certainly result from errors of data entry or analysis (e.g., Hunter & Schmidt, 1990; Rosenthal, 1991). Hunter and Schmidt described these types of faulty data as the most elusive source of error in meta-analytic research. They had this to say about it:

> Bad data can arise from any step in the scientific process. The raw data may be incorrectly recorded or incorrectly entered into the computer. The computer may correlate the wrong variable because the format was incorrectly specified or because a transformation formula was incorrectly written. The sign of the correlation may be wrong because the analyst reverse scored the variable, but the person who read the output did not know that, or because the reader thought the variable had been reverse scored when it had not. When computer output is put in tables, the correlation can be incorrectly copied; the sign may be lost or digits may be reversed, or, when the table is published, the typesetter may miscopy the correlation. (p. 67)

An important consideration is the extent to which clerical errors exist in scientific data. Hunter and Schmidt (1990, pp. 67–68) deferred to several statisticians in this regard. They cited Tukey (1960) as asserting that all real data sets contain this type of error. They also quoted Gulliksen (1986), who asserted he found mistakes in every data set he ever examined, including data he himself produced or data he examined for other investigators. The ubiquity of clerical error is supported by Rosenthal (1991), who examined 27 studies in which he was able to define relatively simple recording mistakes (i.e., when a researcher failed to record an event but a monitoring machine did or vice versa). Alternative sorts of error that may have been present in the data could not be identified and examined systematically. Despite the limited scope of his analysis, Rosenthal found 25 of the 27 data sets contained recording mistakes. Across these 25 studies, mistakes affected a fluctuating proportion of the observations, ranging from a low of .23% of the observations in one study to a high of 48.48% of the observations in another. Typically, about 1% of the recorded observations were faulty.

Vantongelen, Rotmensz, and van der Schueren (1989) examined the rate of clerical errors that occurred while collecting data for a multisite clinical trial.

Across 15 sites, about 3% of the data points were incorrect. This rate of error remained fairly constant over subsequent evaluations (Steward, Vantongelen, Verweij, Thomas, & van Oosterom, 1993; Verweij, Nielsen, Therasse, & van Oosterom, 1997), although it was reduced somewhat with constant monitoring. A slightly higher rate of recording error has been reported in a large multicenter data base on low birth weight infants (Horbar & Leahy, 1995). On average, about 4% of the recorded information was erroneous (e.g., the error rate for recording infant gender was 2.1%, for cesarean birth it was 2.5%).

Although not a function of transcription or recording mistakes, Meyer, Bates, and Gacono (1998) recently described another source of artifact affecting research data. These authors collected observer ratings from a group of mental health counselors and a group of college students, all of whom rated target participants they currently interacted with. The rating scale used in this study (Meyer, 1996) contains four implausible items designed to identify haphazardly generated ratings. Strictly speaking, each of these items only have one "correct" answer. For instance, the only correct response to an item asking whether the target person is "able to breathe on a regular basis" is "definitely true." Prior to completing the scale, all raters were informed it contained items to check whether the ratings were completed in a random or overly careless manner. Despite this caution, of 309 initial ratings, 123 (39.8%) did not provide the logically correct response to at least one of the four implausible items. Furthermore, 50 (16.2%) of the raters provided responses to at least one of these items that were clearly incorrect (e.g., responding "definitely true" or "mostly true" to an item asserting the person had not slept at all over the past 3 months). Obviously, if the data from faulty, haphazard raters had not been detected, it would have skewed the results of any subsequent statistical analyses.

The incidence of meaningless or erroneous survey data appears to be an underappreciated problem. For instance, Barnett (1997) followed a design similar to Meyer et al. (1998) and discovered 8.5% of his participants produced clearly impossible or "incorrect" answers to items imbedded in his questionnaires. An even larger proportion of participants produced implausible answers. For instance, only 68% of the participants in his study responded "strongly disagree" to the statement, "I have lied a lot on this questionnaire." Similarly, only 73% of the participants emphatically denied the statement, "I have absolutely no hair on my body."

Both of these studies relied on typical psychology research participants—students recruited from undergraduate courses who participated for course credit. Should these findings generalize to other settings, they would suggest that much of the psychological literature is confounded by faulty data points that are caused by participants who misread items, have momentary lapses while completing questionnaires, or even deliberately and intentionally mislead researchers (cf. Masling & Schwartz; 1979; Taylor & Shepperd, 1996).

Without looking hard, it is possible to identify many other instances in which participants provided researchers with meaningless or inaccurate data. When the

Minnesota Multiphasic Personality Inventory–2 (MMPI–2; Butcher, Dahlstrom, Graham, Tellegen, & Kaemmer, 1989) was developed, 10.3% of the initial normative sample had to be dropped because they produced unusable or random data. Using similar criteria, a slightly larger number of patients (14.1%) were dropped from the clinical samples that had been gathered when the adolescent version of the MMPI was developed (Williams & Butcher, 1989). During repeat interviews for an epidemiological registry to study heart attacks, about 3% of the patients reported two different birth dates, 13% reported different education levels, and up to 10% reported different events in their medical history (Lim & Dobbins, 1996). Even more remarkable disparities have been found when neurological histories and exams have been repeated on the same patients over a short interval (e.g., for history of prior stroke, kappa = .31; see Shinar et al., 1985; Sisk, Ziegler, & Zileli, 1970), although it is not clear whether the inconsistencies arose from the patients or from the neurologists examining them. Ellish, Weisman, Celentano, and Zenilman (1996) studied the reliability of information obtained from heterosexual partners who were independently interviewed in the same clinic on the same day. The extent of agreement between the partners was remarkably poor for factual information that should have been relatively obvious and easy to agree upon. For instance, the partners' agreement about whether they were married was only kappa = .90. When asked whether they had a child with each other, the agreement was only kappa = .87. When asked whether their partner was currently working, the agreement was kappa = .695. Obviously, the lack of complete consensus on such seemingly obvious information should leave researchers wondering whether any meaningful information could be obtained from such informants.

Faulty research data may also emerge from participants who conscientiously attempt to provide accurate information but simply are incapable of doing so, either because of psychological defenses or because it is virtually impossible to provide correct information to the questions posed (cf. Nisbett & Wilson, 1977; Shedler et al., 1993). For example, a survey of one million high school students determined that 70% thought they were above average in leadership ability whereas only 2% thought they were below average (Gilovich, 1991). Obviously, no more than 49% can be above average and no less than 49% can be below average. When asked how well they get along with others, all students reported they were above average and 25% reported they were in the top 1% on this characteristic. Gilovich also reported that 94% of university professors stated they did their job better than their average colleague. Such faulty information can wreak havoc with any conclusions derived from the data.

Returning to the issue of clerical error, several recent studies have explored specific manifestations in assessment data. Arndt et al. (1994) reported on data gathered for a *Diagnostic and Statistical manual of Mental Disorders* (4th ed. [*DSM–IV*]; American Psychiatric Association, 1994) field trial examining psychotic disorders. They employed a seven-stage error-catching procedure designed to net as many faulty data points as possible. The study examined record forms ob-

tained from 688 structured interviews, each of which contained 429 variables. Comparisons between the initial scores entered into the computer at Stage 1 and the final "clean" data obtained at Stage 7 revealed that 2.4% of the initial data points were erroneous. Furthermore, 674 of the 688 initial record forms (97.96%) contained at least one mistake.

Allard, Butler, Faust, and Shea (1995) explored hand-scoring errors observed with the Personality Diagnostic Questionnaire–Revised (PDQ–R) and the Symptom Checklist–90–Revised (SCL–90–R). In a sample of 43 hand-scored versus computer-scored PDQ–R profiles, they found an average of 2.26 mistakes per hand-scored inventory. The number of mistakes ranged from 0 to 13 and at least one was found with 53% of the patients. For 19% of the patients, erroneous scores resulted in faulty diagnostic classifications. In a second sample of 35 SCL–90–R protocols, the authors found that hand-scoring produced an average of 2.43 mistakes per patient. Summary *T*-score values were affected for 83% of these patients.

Slate, Jones, and Murray (1991) examined 150 protocols from the Wechsler Adult Intelligence Scale–Revised. Across 10 categories of scoring or administration errors, they documented a total of 4,317 errors in the 150 protocols. Although most errors reflected a failure to record responses, a salient proportion were due to mental mistakes like faulty addition when computing subtest raw scores, faulty conversion of raw scores to scaled scores, or miscalculations of the patient's age. Similar rates of error on this Wechsler scale were discussed in Kaufman's (1990) review of this literature. For instance, Kaufman described one study in which 39 examiners knew they would be evaluated for scoring accuracy yet still produced wildly different scores for the same patients. The erroneous scores deviated by as much as a full standard deviation, such that 1 patient with an actual IQ score of 108 was assigned scores by the examiners that ranged from 98 to 116. Many of these errors were a function of incorrect score conversions or simple addition mistakes that emerged when generating raw score totals.

In some respects, this literature offered a bit of relief from the surprise and embarrassment I felt after discovering errors in my own data. In what follows, I describe the latter in more detail and report the impact these mistakes had on the findings reported previously. By presenting this information, my goal is not simply to provide an erratum for the published study. Rather, I also hope to alert others to the pervasive and insidious presence of error in personality assessment data. The article closes by offering several recommendations to help identify and minimize these problems.

ERRORS IN MEYER (1993)

While preparing a recent article (Meyer, in press), I reexamined some findings that had been published earlier (Meyer, 1993). Briefly, the original study considered two issues. First, it explored the extent to which Rorschach response frequency (*R*) was associated with scores on the six constellation indexes from the Comprehen-

sive System. Second, it considered the extent to which *R* moderated the validity of the constellation indexes, using MMPI–2 scales as criterion measures for many of the analyses. For all analyses, patients were drawn from three groups: (a) those with brief Rorschach records ($R = 14–17$), (b) average length records ($R = 21–24$), and (c) long records ($R > 28$).

Initially, I recomputed correlations between the Rorschach Depression Index (*DEPI*) and the seven MMPI–2 depression scales used in the original study, using the same 81 participants from the original analyses (Meyer, 1993).[1] Much to my surprise, I could not replicate the results in Table 4 of the published article. Instead, the average convergent correlation for the low-*R* group was now $r = -.254$ rather than $r = -.317$. For the average-*R* group, the mean correlation was now $r = -.206$ rather than $r = -.190$. It was $r = .340$ rather than $r = .380$ for the high-*R* group. In the full sample, the average construct convergence was –.020 rather than the published value of –.021.

These differences are not dramatic (nor statistically significant), having magnitudes of .063, –.016, .040, and .001, respectively. Although there was some regression toward a mean of 0.0 in the low- and high-*R* groups, the current results also would not have lead to different conclusions. Although the latter was somewhat reassuring, it was beside the point because there should have been absolutely no differences between the published findings and my reanalysis of the same variables. The results should have been identical.

Given these disparities, I expanded the reanalysis. In the original article, Table 7 reported correlations between the Rorschach Hypervigilance Index (*HVI*) and nine MMPI–2 scales, and Table 8 reported correlations between the Obsessive Style Index (*OBS*) and eight MMPI–2 scales. Once again, when the analyses were repeated, I could not exactly replicate the published findings. As with the *DEPI*, the differences were relatively small and statistically nonsignificant. For the *HVI*, the difference in average validity coefficients for the full sample and the low-, average-, and high-*R* groups were –.015, .006, –.062, and .032, respectively. For the *OBS* the differences were .018, –.040, .067, and .040, respectively. Although these reanalyzed data actually strengthened the initial pattern of findings and would have reinforced my original interpretation, this was again beside the point because the results should have been identical.

At first, the disparities were very puzzling because I thought the data sets were equivalent. However, I temporarily forgot that three factors differentiated

[1]Attesting to the insidious influence of error, I initially reanalyzed data from only 80 of the 81 original patients. It was not until after the analyses were complete that I realized the mistake. Although obvious in hindsight, the mistake occurred because I no longer have Minnesota Multiphasic Personality Inventory–2 (MMPI–2) data for 1 patient. Thus, even though a computer program selected all the original patients from the current data file, unbeknownst to me, this patient was excluded because his MMPI–2 scores did not exist.

the current data set from the original. First, the original publication relied primarily on hand-scored MMPI–2 and Rorschach protocols, whereas most of the same protocols have now been scored with computerized software. For the MMPI–2, initially 76.5% of the protocols had been scored and plotted by hand and now all have been scored using software from National Computer Systems. Although I never checked the accuracy of the secretaries who entered each of the 567 MMPI–2 items, this software prevents errors that may emerge when items are summed to calculate scale raw scores or when raw scores are transformed to gender corrected *T* scores. For the Rorschach, all protocols were initially hand-scored and all structural summary values were generated from hand calculations. Now, 47 of the 90 protocols (52%) have been entered into RIAP (Exner, Cohen, & McGuire, 1990), a computerized program that automatically checks for a large number of coding errors and also prevents mistakes that may occur when performing the addition, subtraction, multiplication, and division required to generate structural summary values.

A second differentiating factor is the way in which scores were gathered for the statistical analyses. In the original study, a research assistant manually obtained scores from each patient's file. This was a two-step process in which raw scores for 17 MMPI–2 scales and 50 Rorschach variables (i.e., *R* and all criteria for the six constellation indexes) were recorded on handwritten forms and then manually transferred to a computer file.[2] In the current data set, computer programs were written to electronically transfer scores from the storage files produced by the National Computer Systems and RIAP software into a statistical data file. Relying on software for this stage of the process prevents keystroke mistakes associated with the tedious manual entry of each MMPI–2 and Rorschach score.

Finally, the current data set still contains 43 Rorschach protocols that were manually scored and entered. However, the scores from these protocols contain less error than those in the original analysis because all scores have now been scrutinized by a fairly large error-checking program. The latter incorporates 80 checks on the accuracy of Rorschach data entry and structural summary calculations. These checks succeed in part because certain logical relations hold among Rorschach scores (e.g., $R = DQ+ + DQo + DQv + DQv/+ = W + D + Dd = FQ+ + FQo + FQu + FQ- + FQnone$) and because there is numerical redundancy between the scores from the top and bottom of a structural summary (e.g., $X+\% = [FQ+ + FQo] / R$; $M = Ma + Mp$). Thus, when a complete structural summary is entered into the computer by hand, the program can be used to identify calculation and entry errors.

The error-checking program evaluated each of the individual criteria for each of the constellation indexes, as well as all of the location scores, developmental

[2]On at least one occasion, most of the Rorschach scores that had been entered into the computer were checked against the scores on the handwritten forms to catch entry errors. I do not have a record indicating this was done with the Minnesota Multiphasic Personality Inventory–2 scores.

quality variables, determinants, form quality scores, pairs, and cognitive special scores. In addition, most of the contents, some Popular values, and several other special scores were evaluated for accuracy. Even though the program does not check each structural summary calculation and every data entry keystroke, it provides a reasonably thorough review of potential errors. In particular, by checking so many scores, the program readily identifies instances of systematic error, such as when a score is skipped or entered twice, both of which offset all subsequent entries by one column. Whenever discrepancies were identified by this program, they were resolved by checking the original structural summary. Finally, all of the Rorschach variables also were examined for statistical outliers or impossible values (e.g., $R = 0$; *SCZI* total scores > 6). Errors identified at this stage were again rectified by consulting the original structural summaries.

To understand fully the extent of error in the original report, it would be necessary to compare each score in the current statistical data file to each score in the original. Unfortunately, my research assistant had entered scores and calculated results using a mainframe computer housed at a separate university. As a result, I never possessed a computerized data file. However, I do have all of the original handwritten forms that were used to record file information. These data are one step removed from what was actually analyzed because it indicates what should have been entered, rather than what was entered with each keystroke.

How different are the scores on the handwritten forms from the scores that were actually entered into the computer? To a close approximation, they are equivalent. To ascertain this, I reentered the scores from the handwritten forms (and double-checked each entry) and then calculated the convergent validity coefficients reported in Tables 4, 7, and 8 in Meyer (1993). Across the 28 correlations that dealt with the *DEPI,* 27 were identical and 1 was off by a slight margin, producing an average difference of .002. The 36 coefficients dealing with the *HVI* differed more regularly, although the discrepancies were small, producing an average difference of .014. The recomputed correlations for the *OBS* were identical in 26 of 32 instances. Slight transcription errors produced an average difference of –.006 across the 32 coefficients. Thus, my research assistant made several keystroke mistakes while entering scores from the handwritten forms. However, the extent of this error was quite small.[3] Consequently, although it is not strictly warranted, I treat the scores on the handwritten forms as if they were equivalent to the scores that produced the published findings to explore how computer-assisted scoring, computerized data entry, and a Rorschach error-checking program refined the data set.

Before proceeding, a potentially important consideration is whether the transcription errors that emerged during data entry were random or whether they fa-

[3]It is also important to emphasize that my research assistant was quite bright and had approached this task in a conscientious, meticulous, and well-organized fashion. As such, the errors should be viewed as a natural element of an otherwise sound data entry process.

vored the experimental hypotheses (cf. Rosenthal & Rubin, 1978). I believe my research assistant knew the guiding hypotheses, although at this point I cannot be positive. In either case, his data entry errors appeared to be random. In two instances, the differences could be interpreted as favoring my initial hypotheses (that Rorschach–MMPI convergent correlations would be strongest when Rorschach protocols were of an average length); however, in the seven other instances they could not. Furthermore, in the two instances that could be considered potential instances of expectancy bias, the magnitude of the differences created by the data entry errors were trivial—an average difference of .005 in one case and .002 in the other. Thus, the very slight data entry mistakes did not favor the experimental hypotheses.

In this context, it is also important to note that the overall results regarding Rorschach and MMPI convergent validity in the original study directly contradicted the experimental hypotheses. That is, my original hypotheses on this topic were not supported by the data. As such, none of the errors discussed in this article had favored my preconceived notions.

Given that the handwritten data forms led to results that very closely approximated the Rorschach and MMPI–2 findings as they were originally reported, the next step compared the handwritten forms to the current data file. For the MMPI–2 there were a total of 1,377 data entries (17 scales × 81 patients). Of these, there were a total of 81 discrepancies, leading to an error rate of 5.88%.[4] Most discrepancies consisted of 1- or 2-point differences on a continuous scale. These probably reflect counting mistakes that emerged when the MMPI–2s were first hand-scored and items were mentally summed to obtain total scores for each scale. In a smaller number of instances, the discrepancies were larger. These errors exceeded 4 or more points and extended to a maximum of 12 raw score points. Such errors could reflect more serious counting mistakes, number transpositions, or perhaps an instance when my research assistant intended to record a total score for one scale but erroneously recorded a score that actually belonged to an adjacent scale on the MMPI–2 profile sheet.

For the original analyses dealing with the Rorschach, the 90 patients produced 4,410 binary entries for the constellations and 90 entries for *R*. Of these, there were a total of 70 discrepancies, producing an error rate of 1.56%. The errors probably reflect minor problems with the initial Rorschach hand-scoring and hand-calculations, as well as transcription errors that developed when recording information from patient files. For instance, the most extreme Ror-

[4]As indicated in Footnote 1, I no longer posses full Minnesota Multiphasic Personality Inventory–2 (MMPI–2) data for one patient; all that is available are his 17 scores on the handwritten forms. I treated these scores as if they were entirely accurate for both data sets. Also, 19 of the original MMPI–2s had been scored by computer. They produced no discrepancies with the current data file. When the error rate was calculated using only the 61 MMPI–2s that were initially hand scored, it became 7.8%.

schach error occurred when the *R* values were transposed for 2 patients. One patient with 39 responses was assigned 57 responses, whereas the patient with 57 responses was assigned 39.

To quantify the impact of these data errors, reliability coefficients were generated by correlating scores from the initial coding forms with scores in the current data file. Reliability for the key Rorschach variables (*R, SCZI, DEPI, S–CON, CDI, HVI,* and *OBS*) ranged from .95 to .99, with a mean of .97. Reliability for the 17 MMPI–2 scales ranged from .94 to .9997, having a mean of .99.[5]

Another core set of analyses in the original article examined the hypothesis that *R* would be associated with the constellation indexes. This had been tested in two ways. Table 1 (Meyer, 1993) presented chi-square analyses testing the hypothesis that constellation criteria based on raw frequencies would be associated with *R* in a predictable pattern. The second analysis (Table 2 in the original) proposed that total scores on each of the constellations would be correlated with R. For both sets of analyses, the original results were virtually identical to those obtained using the current data file. For instance, the original correlations between *R* and the *S–CON, SCZI, DEPI, CDI, HVI,* and *OBS* were .25, .26, .30, –.37, .60, and .56, respectively. Using the current data file, the corresponding correlations are .29, .29, .29, –.34, .63, and .62, respectively (all $p < .05$). In general, the current, more error-free scores provided slightly stronger support for the original hypotheses than did the published results.[6]

To summarize, these analyses indicated several points. First, there was a small amount of seemingly random error in the original results (Meyer, 1993). However, reliability analyses indicated that the Rorschach and MMPI–2 scales used in those analyses were virtually identical to the scales in the current, more error-free data file. Furthermore, even though some specific findings are a bit stronger and others a bit weaker, the overall pattern of results reported by Meyer (1993) do not change when the more accurate data are utilized. Thus, the findings were correct as reported. For the participants in these analyses, *R* was correlated with total scores on the constellation indexes and most of the individual criteria based on raw frequen-

---

[5]It may seem odd that the Rorschach scales have slightly lower reliabilities than the Minnesota Multiphasic Personality Inventory–2 (MMPI–2) scales, given that there was less absolute error in the Rorschach data. However, the Rorschach scales have a smaller possible range than the MMPI–2 scales. For instance, the maximum score for a Rorschach constellation is 12 (for the *S–CON*), whereas the maximum score for one of the 17 MMPI–2 scales is 78 (for Scale *8*). Because of these differences, errors have a stronger impact on the Rorschach scales than the MMPI–2 scales.

[6]It should be kept in mind that these correlations were obtained from three discontinuous groups selected on the basis of *R*. Using all patients regardless of *R* produces smaller correlations. In an unselected sample of 443 patients, the corresponding correlations are .16, .24, .19, –.24, .48, and .49, respectively (all $p < .05$). Despite these reductions, in the same sample of 443, *R* is still correlated with 31 of the 33 individual constellation criteria that are based on raw frequencies. The average correlation for these 33 criteria is .25, whereas the average correlation for the criteria expected to be independent of *R* is .05.

cies. In addition, Rorschach and MMPI–2 convergent validity was moderated by differences in *R*. Rorschach and MMPI–2 scales generally produced negative convergent validity coefficients in brief records and positive validity coefficients in lengthy records.

At the risk of complicating issues further, it is important to recognize that, even though the original findings were accurate as reported (i.e., unchanged by the small degree of error in the original data), the findings related to Rorschach and MMPI–2 convergent validity ultimately proved to be artifactual when the analyses were conducted using an expanded sample of adult patients (see Meyer, in press) or a sample of adolescent patients (Krishnamurthy, Archer, & House, 1996). Thus, two types of error need to be considered separately. In this article, I have focused on scoring and data entry errors. These clerical errors were relatively small and did not affect the conclusions reported in the original publication. However, another type of error also affected some of the original results. The original publication relied on small patient samples in each of the *R* groups. With small samples, there is an increased risk that the data will produce an artificial pattern of findings that will not replicate on new samples. This process functions because of statistical sampling error, which is independent of scoring error or data entry error. As detailed in Meyer (in press), this second type of error did affect the original results concerning Rorschach and MMPI–2 convergent validity.[7] Specifically, the small patient samples produced data that led me to draw erroneous conclusions about the role that *R* played in moderating the correlation between Rorschach and MMPI–2 scales. Using substantially larger patient samples or an independent patient sample, it has been found that *R* alone does not moderate convergent validity. Meyer (in press) provided a more complete discussion of these issues, along with data on a more complex set of factors that genuinely do seem to moderate Rorschach and MMPI–2 convergent validity.

CONCLUSIONS

All human activities are bedeviled by error, including research. Nonetheless, I was surprised when, after a lag of several years, I discovered it had infiltrated my own published data. Although the errors ultimately turned out to be relatively trivial, this will not always be the case. Arndt et al. (1994) discovered that unrecognized error in their *DSM–IV* field trial would have resulted in seemingly poor interrater reliability for some relatively simple variables. For instance, raters should be able to easily decide whether the person they are interviewing is a psychiatric inpatient. However, it was only after Arndt et al. closely scrutinized

[7]As Footnote 6 indicates, the original findings related to *R* and the Rorschach constellation indexes do generalize well and, thus, were not a function of sampling error.

their data and removed errors that the true reliability of this variable was evident, jumping from a kappa of approximately .44 to a kappa of .98. Other interrater reliability coefficients in their data set increased in magnitude by .30 or .40 after the data were cleaned. The research by Allard et al. (1995) is also instructive. Even though they found a relatively small error rate in their PDQ–R data, 19% of the patients in their sample were either classified as having a personality disorder when that was not correct or vice versa.

To counter the influence of error, researchers and clinicians who engage in personality assessment would be wise to anticipate its presence. Because personality assessment is founded on numbers, there is always an opportunity for error to infiltrate the data through miscalculations, transpositions, inattentiveness, or keystroke slips. Several common sense proactive steps can be taken to counter these prospects. First, whenever possible, assessment data should be scored using computerized algorithms instead of relying on manual procedures. Second, if forced to rely on hand scoring, whenever possible at least a portion of the scores should be double-checked for accuracy. Given how frequently clerical errors are made during test scoring (e.g., Allard et al., 1995; Kaufman, 1990), this should achieve two ends: (a) It should allow for the immediate correction of errors and (b) it should sensitize examiners to the frequency of mistakes, hastening conversion to some form of computerized scoring.

Third, when importing assessment data for use in research, it would be best to avoid "touching" the data if at all possible. Instead, it is best to write (or enlist someone to write) a computer program that will directly import the information from computerized scoring files into whatever data file is suitable for statistical analyses. Fourth, it is essential to double-check and triple-check a personal computer program to ensure it is producing the proper output. The only thing worse than the random errors associated with manual data entry is the systematic error created by a faulty computer program. As Arndt et al. (1994) reminded us, a faulty program can turn 42-year-old household heads into 14-year-old widows.

Fifth, if forced to enter research data by hand (which is typically the case), steps should be taken to minimize the inevitable data entry errors. When possible, all the data should be independently entered on two separate occasions and a computer program should be used to compare each data point in both files (though see Day, Fayers, & Harvey, 1998). Double entry is an expensive process, but it is also a step that catches virtually all of the errors that are a function of data keying. If double entry is not an option, redundancy should be built into the data entry process so that mathematically equivalent scores are entered at two unique places in the data file. Taking this step allows the researcher to compare two sets of values that should be equal, thereby catching many instances of systematic entry error (e.g., any shifts in column alignment). Built-in redundancy also helps to catch errors that remain from the initial scoring because it allows one to generate an error-checking program that searches for logical inconsistencies. As discussed earlier for the Ror-

schach, a complete structural summary contains considerable redundancy, which allows one to make at least 80 logical checks on the integrity of the data. Similar programs could be devised for comparing MMPI–2 raw scores and their corresponding *T* scores. As a final step, it is critical to review all the variables in a data set for statistical outliers or for values that reside outside the range of possible scores (cf. Behrens, 1997).

An additional consideration that seems underappreciated in research more generally is the quality of information derived from participants who provide self-ratings or observer ratings. As discovered in several of the studies mentioned earlier, participants often treat the research tasks in a careless or quasi-random manner. Given this, it is always optimal to include validity items within a survey instrument that only have one legitimately correct answer. Deleting data from participants who provide impossible responses to these items is another sensible way to minimize the impact of error on research findings. Although such validity items can be easily incorporated into questionnaires and rating scales, contending with participants who unwittingly provide consistent but faulty information (e.g., Mabe & West, 1982; Nisbett & Wilson, 1977; Shedler et al., 1993) is a more vexing problem that is not easily resolved except by measuring each construct with multiple assessment methods.

These suggestions do not exhaust the ways in which determined, creative researchers and clinicians could minimize the error associated with personality assessment data. In fact, Smith, Budzeika, Edwards, Johnson, and Bearse (1986) provided additional recommendations for researchers who wish to avoid data errors, Freedland and Carney (1992) provided an excellent discussion of complications that arise with computerized data, and Cohen (1990, 1994) provided sage advice for avoiding some of the more general judgment lapses that affect research. Overall, an appreciation of how often mistakes occur with data, in conjunction with some common sense precautions to minimize their impact, should help avoid the uncomfortable discovery of errors in clinical practice and in research.

ACKNOWLEDGMENTS

I thank Bill Kinder, Joe Masling, and several anonymous reviewers for their helpful comments related to these issues.

REFERENCES

Allard, G., Butler, J., Faust, D., & Shea, M. T. (1995). Errors in hand scoring objective personality tests: The case of the Personality Diagnostic Questionnaire–Revised (PDQ–R). *Professional Psychology: Research and Practice, 26,* 304–308.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Arndt, S., Tyrrell, G., Woolson, R. F., Flaum, M., & Andreasen, N. C. (1994). Effects of errors in a multicenter medical study: Preventing misinterpreted data. *Journal of Psychiatric Research, 28,* 447–459.

Barnett, P. (1997). *Psychosocial balance and interpersonal style.* Unpublished master's thesis, University of Alaska, Anchorage.

Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. *Psychological Methods, 2,* 131–160.

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

Butcher, J. N., Graham, J. R., Haynes, S. N., & Nelson, L. D. (Eds.). (1995). Methodological issues in psychological assessment research [Special issue]. *Psychological Assessment, 7*(3).

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin, 56,* 81–105.

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

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45,* 1304–1312.

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

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Boston: Houghton Mifflin.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.

Day, S., Fayers, P., & Harvey, D. (1998). Double data entry: What value, what price? *Controlled Clinical Trials, 19,* 15–24.

Ellish, N. J., Weisman, C. S., Celentano, D., & Zenilman, J. M. (1996). Reliability of partner reports of sexual history in a heterosexual population at a sexually transmitted diseases clinic. *Sexually Transmitted Diseases, 23,* 446–452.

Enfield, S. (1998, February). A Pole too far: Three skydivers die in Antarctica, leaving the world to ask, "Why?" *Outside,* 22.

Exner, J. E., Jr. (Ed.). (1995). *Issues and methods in Rorschach research.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Exner, J. E., Jr., Cohen, J. B., & McGuire, H. (1990). *RIAP Version 2: Rorschach Interpretation Assistance Program.* Asheville, NC: Rorschach Workshops.

Freedland, K. E., & Carney, R. M. (1992). Data management and accountability in behavioral and biomedical research. *American Psychologist, 47,* 640–645.

Gilovich, T. (1991). *How we know what isn't so: The fallibility of human reason in everyday life.* New York: Free Press.

Gulliksen, H. (1986). The increasing importance of mathematics in psychological research (Part 3). *The Score, 9,* 1–5.

Hammond, K. R. (1996). *Human judgment and social policy: Irreducible uncertainty, inevitable error, unavoidable injustice.* New York: Oxford University Press.

Horbar, J. D., & Leahy, K. A. (1995). An assessment of data quality in the Vermont-Oxford Trials Network database. *Controlled Clinical Trials, 16,* 51–61.

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

Kaufman, A. S. (1990). *Assessing adolescent and adult intelligence.* Boston: Allyn & Bacon.

Kazdin, A. E. (1998). *Research design in clinical psychology* (3rd ed.). Boston: Allyn & Bacon.

Krishnamurthy, R., Archer, R. P., & House, J. J. (1996). The MMPI–A and Rorschach: A failure to establish convergent validity. *Assessment, 3,* 179–191.

Lim, L. L., & Dobbins, T. (1996). Reproducibility of data collected by patient interview. *Australian and New Zealand Journal of Public Health, 20,* 517–520.

Mabe, P. A., III, & West, S. G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. *Journal of Applied Psychology, 67,* 434–452.

Masling, J. (1992). The influence of situational and interpersonal variables in projective testing. *Journal of Personality Assessment, 59,* 616–640. (Reprinted from *Psychological Bulletin, 1960, 57,* 65–85.)

Masling, J., & Schwartz, M. (1979). A critique of research in psychoanalytic theory. *Genetic Psychology Monographs, 100,* 257–307.

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

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

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

Meyer, G. J. (in press). The convergent validity of MMPI and Rorschach scales: An extension using profile scores to define response/character styles on both methods and a re-examination of simple Rorschach response frequency. *Journal of Personality Assessment.*

Meyer, G. J., Bates, M., & Gacono, C. (1998). *The Rorschach Rating Scale: Item adequacy, scale development, and relationships with the five factor model of personality.* Manuscript under review.

Murray, G. D. (1988). The task of a statistical referee. *British Journal of Surgery, 75,* 664–667.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84,* 231–259.

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

Rosenthal R., & Rubin, D. B. (1978). Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences, 3,* 377–415.

Rossi, J. S. (1987). How often are our statistics wrong? A statistics class exercise. *Teaching of Psychology, 14,* 98–101.

Schinka, J. A., & Greene, R. L. (Eds.). (1997). *Emerging issues and methods in personality assessment.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Schmidt, F. L. (1992). What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *American Psychologist, 47,* 1173–1181.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1,* 115–129.

Schmidt, F. L., & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1,* 199–223.

Shedler, J., Mayman, M., & Manis, M. (1993). The *illusion* of mental health. *American Psychologist, 48,* 1117–1131.

Shinar, D., Gross, C. R., Mohr, J. P., Caplan, L. R., Price, T. R., Wolf, P. A., Hier, D. B., Kase, C. S., Fishman, I. G., Wolf, C. L., & Kunitz, S. C. (1985). Interobserver variability in the assessment of neurologic history and examination in the Stroke Data Bank. *Archives of Neurology, 42,* 557–565.

Sisk, C., Ziegler, D. K., & Zileli, T. (1970). Discrepancies in recorded results from duplicate neurological history and examination in patients studied for prognosis in cerebrovascular disease. *Stroke, 1,* 14–18.

Slate, J. R., Jones, C. H., & Murray, R. A. (1991). Teaching administration and scoring of the Wechsler Adult Intelligence Scale–Revised: An empirical evaluation of practice administrations. *Professional Psychology: Research and Practice, 22,* 375–379.

Smith, P. C., Budzeika, K. A., Edwards, N. A., Johnson, S. M., & Bearse, L. N. (1986). Guidelines for clean data: Detection of common mistakes. *Journal of Applied Psychology, 71,* 457–460.

Steward, W. P., Vantongelen, K., Verweij, J., Thomas, D., & van Oosterom, A. T. (1993). Chemotherapy administration and data collection in an EORTC collaborative group: Can we trust the results? *European Journal of Cancer, 29A,* 943–947.

Taylor, K. M., & Shepperd, J. A. (1996). Probing suspicion among participants in deception research. *American Psychologist, 51,* 886–887.

Tukey, J. W. (1960). A survey of sampling from contaminated distributions. In I. Olkin, J. G. Ghurye, W. Hoeffding, W. G. Madoo, & H. Mann (Eds.), *Contributions to probability and statistics.* Stanford, CA: Stanford University Press.

Vantongelen, K., Rotmensz, N., & van der Schueren, E. (1989). Quality control of validity of data collected in clinical trials: EORTC Study Group on Data Management (SGDM). *European Journal of Cancer and Clinical Oncology, 25,* 1241–1247.

Verweij, J., Nielsen, O. S., Therasse, P., & van Oosterom, A. T. (1997). The use of systematic therapy checklist improves the quality of data acquisition and recording in multicentre trials: A study of the EORTC Soft Tissue and Bone Sarcoma Group. *European Journal of Cancer, 33,* 1045–1049.

Wampold, B. E., Davis, B., & Good, R. H., III. (1990). Hypothesis validity in clinical research. *Journal of Consulting and Clinical Psychology, 58,* 360–367.

Wiggins, J. S. (1988). *Personality and prediction: Principles of personality assessment.* Malabar, FL: Krieger. (Original work published 1973)

Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist, 53,* 300–314.

Williams, C. L., & Butcher, J. N. (1989). An MMPI study of adolescents: I. Empirical validity of the standard scales. *Psychological Assessment, 1,* 251–259.

Wolins, L. (1962). Responsibility for raw data. *American Psychologist, 17,* 657–658.

Zuckerman, M., Hodgins, H. S., Zuckerman, A., & Rosenthal, R. (1993). Contemporary issues in the analysis of data: A survey of 551 psychologists. *Psychological Science, 4,* 49–53.

Gregory J. Meyer
Department of Psychology
University of Alaska Anchorage
3211 Providence Drive
Anchorage, AK 99508

Received April 29, 1998
Revised July 22, 1998