JOURNAL OF PERSONALITY ASSESSMENT, 1992, 59(2), 366-379

Statistical Power and the Rorschach: 1975–1991

Marvin W. Acklin
Departments of Psychiatry and Psychology
University of Hawaii
Honolulu, HI

Claude J. McDowell, II
Forest Institute of Professional Psychology
Honolulu, HI

Steffani Orndoff
Castle Medical Center
Kailua, HI

The Rorschach Inkblot Test has been the source of long-standing controversies as to its nature and its psychometric properties. Consistent with behavioral science research in general, the concept of statistical power has been entirely ignored by Rorschach researchers. The concept of *power* is introduced and discussed, and a power survey of the Rorschach literature published between 1975 and 1991 in the *Journal of Personality Assessment*, *Journal of Consulting and Clinical Psychology*, *Journal of Abnormal Psychology*, *Journal of Clinical Psychology*, *Journal of Personality*, *Psychological Bulletin*, *American Journal of Psychiatry*, and *Journal of Personality and Social Psychology* was undertaken. Power was calculated for 2,300 statistical tests in 158 journal articles. Power to detect small, medium, and large effect sizes was .13, .56, and .85, respectively. Similar to the findings in other power surveys conducted on behavioral science research, we concluded that Rorschach research is underpowered to detect the differences under investigation. This undoubtedly contributes to the inconsistency of research findings which has been a source of controversy and criticism over the decades. It appears that research conducted according to the Comprehensive System for the Rorschach is more powerful. Recommendations are offered for improving power and strengthening the design sensitivity of Rorschach research, including increasing sample sizes, use of parametric statistics, reduction of error variance, more accurate reporting of findings, and editorial policies reflecting concern about the magnitude of relationships beyond an exclusive focus on levels of statistical significance.

The Rorschach Inkblot Test has long remained the center of controversy, despite its widespread popularity in clinical settings (Lubin, Larsen, & Matarazzo, 1984). This is nowhere more evident than in the long, often acrimonious, series of reviews of the test published in the *Mental Measurements Yearbook* (MMYB; cf. Jensen, 1965). One long-standing controversy concerns whether the Rorschach is actually a "test" at all or whether it is more appropriately thought off as a clinical "technique" (Eron, 1965; Rabin, 1972; Zubin, Eron, & Schumer, 1965). In the fourth MMYB, Sargent (1953), a supporter of the test, emphatically stated, "the Rorschach test is a clinical technique, not a psychometric method" (p. 218). As a test, the Rorschach has been assailed by psychometrically minded psychologists as failing to meet many, if not most, of the standard criteria of test construction, including indices of internal consistency, interrater reliability, and validity (Dana, 1965; Jensen, 1965; McArthur, 1972). Furthermore, as early as 1949, the eminent psychometrician, Cronbach, expressed concerns about the quality of Rorschach research. He wrote in a quote used by Eysenck (1959) in a scathing MMYB review of the Rorschach that "Perhaps ninety percent of the conclusions so far published as a result of statistical studies are unsubstantiated—not necessarily false—but based on unsound analysis" (Cronbach, 1949, p. 425). Cronbach went on to state that "one cannot attack the test merely because most Rorschach hypotheses are still in the pre-research stage" (p. 426). Alternately, the critics of the test wonder that "years of negative research have not cooled the ardor of the Rorschach supporter" (Knutson, 1972, p. 440). The very nature of the Rorschach; the divergent systems of administration, scoring, and research; the nature of Rorschach scores and the shapes of score distributions obtained; and the type of statistics commonly used (typically distribution free or nonparametric) seem to favor the views of the Rorschach's critics.

Various studies over the years have attempted to answer these criticisms. Over the past 30 years, a steady stream of articles has addressed reliability and validity issues. These efforts culminated in the work of Exner. Exner's (1974, 1978, 1986; Exner & Weiner, 1982) Comprehensive System, developed since 1974, has lead to a general standardization of the test. It is generally assumed, though it has never been empirically demonstrated, that the Comprehensive System with its emphasis on standardization, increased reliability, and systematic efforts at validation has placed the Rorschach on a solid foundation as a psychometric instrument.

The test has been subjected to several meta-analyses. Parker (1983) found that studies guided by theory, prior research, or both tended to support the Rorschach but found little support for the Rorschach among studies in which experimental hypotheses lacked a theoretical or empirical rationale. Other studies have found that conceptual, theory-based studies show greater support for the Rorschach than do undirected studies (Atkinson, 1986; Atkinson, Quarrington, Alp, & Cyr, 1986). Furthermore, the power of the statistics used

(i.e., the tests' probability of detecting an effect when one is actually present) was shown to influence the magnitude of observed differences (Parker, 1983).

Meta-analytic studies have compared the Rorschach with the Minnesota Multiphasic Personality Inventory (MMPI), which is "considered the standard of psychological assessment" (Kendall & Norton-Ford, 1982, p. 310), and the Wechsler Adult Intelligence Scale (Atkinson, 1986; Parker, Hansen, & Hunsley, 1988). These studies have found broadly comparable and respectable psychometric properties, especially for the Rorschach. One study concluded that the Rorschach and MMPI "have acceptable and roughly equivalent psychometric properties when used in appropriate circumstances (Parker et al., 1988, p. 372). Finally, consistent with earlier studies, the statistics used to report the results were found to influence the magnitude of the findings.

One may then conclude that the standardization of the test and recent, favorable meta-analyses have made the Rorschach "psychometrically respectable" and, consequently, that all is well with Rorschach research? As just mentioned, the assumption that the Comprehensive System for the Rorschach has resulted in a more psychometrically respectable test has yet to be demonstrated empirically. Another more serious, methodological issue has emerged—statistical power—which has serious implications for not only Rorschach research but for the whole behavioral science research enterprise.

A brief digression is necessary to introduce the concept of statistical power in psychological research. The mainstay of psychological research—hypothesis testing, statistical inference, and dichotomous significance-testing decisions, including the sanctified .05 level of significance—are the legacy of the preeminent statistician Fisher. In Fisher's (1935) scheme, the *null hypothesis* (i.e., the hypothesis of no effect, designated as Ho) is accepted or rejected on the basis of statistical inference. Concern with obtaining statistical significance has long been a central focus in American behavioral science (Bakan, 1966; Rosnow & Rosenthal, 1989). Starting as early as 1942, there has been consistent criticism of the employment of significance tests as ultimate objectives in experimental research (Chase & Tucker, 1976; Meehl, 1978). These criticisms have had little impact on research in general or in the training of students. Critics have encouraged researchers to examine the magnitude of relationships between variables, not merely the probability of their occurrence. Rosnow and Rosenthal (1989) wrote

> it is important to realize that the effect size tell us something very different from the *p* level. A result that is statistically significant is not necessarily practically significant as judged by the magnitude of the effect. Consequently, highly significant *p* values should not be interpreted as automatically reflecting large effects. (p. 1279)

Contrary to the beliefs of many students in psychology (and an alarming number of academic psychologists), the level of statistical significance (*p* values)

obtained says nothing about the magnitude or importance of group differences and nothing about the probability of the truth of the null hypothesis. Kish (1959) aptly pointed out that

> the function of statistical tests is merely to answer: Is the variation great enough for us to place some confidence in the result; or, contrarily, may the latter be merely a happenstance of the specific sample on which the test was made? The question is interesting, but surely it is secondary, auxiliary, to the main question: Does the result show a relationship which is of substantive interest because of its nature and magnitude? (p. 336)

In Fisher's (1935) scheme, hypothesis testing is asymmetrical. There is no alternate or research hypothesis (traditionally designated as H1). Positing the alternate hypothesis to the null hypothesis, an approach advocated by critics of Fisher (Neyman & Pearson, 1928, 1933), gradually made its way into research practice and teaching, creating a sort of hybrid approach to statistical inference.

Cohen (1990) stated a proposition fundamental to Neyman and Pearson's approach,

> The rejection of the null hypothesis when it is true was an error of the first kind [a Type I error] controlled by the alpha criterion, but the failure to reject it when the alternate hypothesis was true was also an error, an error of the second kind [a Type II error] which could be controlled to occur at a rate beta. (p. 130)

An alpha of .05 corresponds to a .95 probability of a correct statistical conclusion when the null hypothesis is true (1 − alpha). In this sort of decision calculus, a decision can and must be made about the relative seriousness of Type I or Type II errors. Traditional applied statistics have focused almost exclusively on controlling Type I errors (i.e., the probability of rejecting a true null hypothesis) with a focus of the level on significance (p values) and has entirely neglected Type II errors (i.e., the probability of accepting a false null hypothesis) and the power of tests. In general, it has been assumed that Type I errors are much more serious than Type II errors (McNemar, 1960).

Assuming the magnitude of the effect size (either predicted in the planning of a study or calculated after a study) and the setting of alpha, one can determine the necessary sample size to meet acceptable conditions for hypothesis acceptance or rejection. One is then in a position to determine the probability of correctly rejecting a false null hypothesis or accepting a true alternate hypothesis: the power of the test. In short, the power of a study is the probability of detecting a difference when one is really there.

The power of a statistical test is a function of the effect size (i.e., a measure of the magnitude of the differences among means expressed in standard deviation units or the amount of variance accounted for), error variance, alpha criterion,

sample size, and data analysis (i.e., the inherent power of the statistical tests to detect differences). Nonparametric tests (i.e., those that rank order or categorize information) generally have less inherent power than do parametric tests (i.e., those that use scores representing degrees of the variable along some continuum; Lipsey, 1990). Furthermore, directional tests have greater statistical power than nondirectional tests. The combination of these factors determines "design sensitivity" ("the ability to detect a real contrast or difference between experimental conditions on some characteristic of interest"; Lipsey, 1990, pp. 13–14). Design sensitivity is also affected by other factors, including use of alpha adjusted procedures (Sedlmeier & Gigerenzer, 1989), violations of independence, false assumptions concerning equality of variances, and false assumptions about measurement scales and the shape of distributions. Cohen (1988) suggested a beta equal to .20 as a reasonable value for general use, or more specifically, he suggested .80 as a desirable minimum for statistical power (1 − beta). That is, a study has acceptable power when it can detect an effect 8 times out of 10. The neglect of power has serious implications for the overall conduct and robustness of psychological research, for underpowered studies may not be sensitive enough to detect the differences that they purport to investigate.

Cohen's (1962) classic paper introduced the concept of power to the psychological literature. He surveyed the 1960 volume of the *Journal of Abnormal and Social Psychology* and concluded that, on average, research studies had about one chance in five or six (.18) for detecting small effects (expressed in standard deviation units or proportion of predicted variance accounted for), less than one chance in two (.48) of detecting medium effects, and about 8 out of 10 chances of detecting large effects (.83). Cohen (1962) expressed his concerns as follows: "The consequences of this state of affairs are fairly obvious. If many investigators are running high risks of failing to detect substantial population effects, much research is resulting in spuriously 'negative results' " (p. 153).

These concerns focus on published research, reflecting the widespread editorial bias against publishing nonsignificant results (Bakan, 1966). Cohen (1962) wrote, ". . . if anything, published studies are more powerful than those which do not reach publication, certainly not less powerful" (p. 152). Unsubmitted and unpublished "file drawer" studies (Rosenthal, 1979), representing perhaps thousands of investigations and whole lines of research, may have been undertaken and abandoned due to the lack of design sensitivity. Thus behavioral research as a whole may stand on shaky methodological foundations.

Cohen's (1962) findings and warnings have had little or no effect on subsequent psychological research (Rossi, 1990; Sedlmeier & Gigerenzer, 1989). Twenty-five power surveys conducted since 1962 have attempted to replicate Cohen's findings (Rossi, 1990). These studies have been conducted in several disciplines, including communications; speech pathology; occupational therapy; management; and educational, social, abnormal, and applied psychology. Quoting Rossi's (1990) excellent review, "the average statistical power for all

twenty-five power surveys (including Cohen's) was .26 for small effects, .64 for medium effects and .85 for large effects and was based on 40,000 statistical tests published in over 1,500 journal articles" (p. 64). In these studies, the findings are clear that power to detect small and medium differences is unacceptably low. For example, if a small effect were in fact present, the typical test would yield statistical significance only 26% of the time and yield nonsignificant results 74% of the time.

The implications for psychological research of this state of affairs should not be minimized. These include the proliferation of Type I errors, spurious overacceptance of the null hypothesis, the often-observed failure of replication studies, and the difficulties in interpreting negative findings. Research areas that are controversial because of equivocal findings (e.g., the Rorschach) may be victims of "artifactual controversy" (Rossi, 1982, 1983). Rossi wrote that "dependence on statistical tests to establish the existence of an effect may lead to artifactual controversy if the average power of the research designed to detect the effect is only about one-half." When average power for a research domain is around .50 (essentially a coin toss), a mixed pattern of significant and nonsignificant findings is likely.

We investigated the Rorschach literature published between 1975 and 1991. This time span constitutes the years encompassing the advent and emergence of the Comprehensive System. The study focuses on all of the Rorschach research published in the *Journal of Personality Assessment* (formerly the *Rorschach Research Exchange* and the *Journal of Projective Techniques*), the *Journal of Consulting and Clinical Psychology*, *Journal of Clinical Psychology*, *Journal of Abnormal Psychology*, *Psychological Bulletin*, *American Journal of Psychiatry*, and *Journal of Personality and Social Psychology*, which are major outlets for American Rorschach research. Our objective was to evaluate the average power of published Rorschach research to determine the extent to which artifactual controversy may plague the test's performance as a research tool. Our second objective was to determine the extent to which the Comprehensive System, with its standardization of administration and scoring, has presumably rectified the situation.

METHOD

All of the articles published in the referenced journals between 1975 and 1991 were examined ($N = 293$), and articles not reporting statistics were eliminated from the study. In addition, a number of studies were eliminated because of their inapplicability for power analysis (e.g., canonical and factor analyses, reviews, case studies, and commentaries) or because sufficient information for calculation of power coefficients was not reported. A total of 135 articles were eliminated.

Similar to previous power surveys, a distinction was made between major and peripheral statistical tests. Major tests dealt directly with Rorschach variables, whereas peripheral tests did not. Peripheral tests that were excluded from the survey included, for example, all of the correlation coefficients of a factor analysis, cluster analysis, and multidimensional scaling; interrater, internal consistency, and temporal consistency reliability coefficients; tests of statistical assumptions; and statistical tests not bearing directing on Rorschach variables. Power was determined for the following tests: *t* tests, Pearson's *r*, partial correlations, chi-square tests, and *F* test in the analysis of variance (ANOVA) and covariance (ANCOVA). Power coefficients were computed for main effects only in factorial ANOVAs. In contrast to other power surveys, but following Cohen (1962), power analyses were calculated for nonparametric techniques, because these are overrepresented in the Rorschach research literature. In these cases, power was determined for the analogous parametric test, for example, the *t* test for means was substituted for the Mann–Whitney *U* test and for the Wilcoxon matched-pairs signed-ranks test, *F* test for the Kruskal–Wallis *H* test and for the Friedman test, and Pearson's *r* for Spearman's rho. According to Cohen, the effect of this substitution is a slight overestimation of power based on the usual assumption that the conditions required by parametric tests were obtained. Power coefficients were calculated using a computer program (Borenstein & Cohen, 1988). When the computer program was not helpful, we referred to power tables from Cohen (1988) and Lipsey (1990).

Of the 293 studies examined, 135 articles were excluded because they were either inapplicable to power analysis (e.g., case reports or literature reviews) or failed to provide enough information (e.g., means, standard deviations, sample sizes, degrees of freedom, and critical values) to calculate power coefficients. Studies were initially examined as a whole group, and those with Comprehensive System methodologies were subsequently examined separately.

The customary procedure in power surveys is to identify the magnitude of effect sizes considered small, medium, and large in the domain of interest (see Cohen, 1988), then determine the average power of the studies to detect such effects (Lipsey, 1990, p. 21). Following Cohen (1962), power was determined by averaging across the statistical tests reported in each article so that each article contributed equally to the overall power assessment. Power estimates were based on Cohen's (1988) later revisions for small, medium, and large effect sizes. Mean power of the major tests was determined for the three levels of effect sizes for each study. Mean power values were then distributed, and their central tendency and variability was determined. We hypothesized that the Rorschach research using a Comprehensive System methodology would be more powerful than non-Comprehensive System research. An alpha criterion of .05 was assumed as was a directional, one-tailed test.

RESULTS

Table 1 presents frequencies and cumulative percentage distributions of the power to detect small, medium, and large effect sizes for 158 journal articles studied. Studies in which small effects were found had less than one in six chances of detecting significant results (.131). Studies with medium effect sizes had a slightly better than one in two chances of detecting significant differences (.561). Finally, studies with large effect sizes met Cohen's (1988) criteria for acceptable power (.85). The power coefficients indicate that Rorschach research is not so different than behavioral science research in general with respect to its power.

Table 2 presents frequencies and cumulative percentage distributions of the power to detect small, medium, and large effect sizes for Rorschach research, including Comprehensive System methodologies (n = 43). These findings indicate low power to detect differences in studies with small effect sizes, with slightly better than one chances in six of detecting significance (.166), slightly more than 6 chances out of 10 in detecting medium effects (.623), and acceptable levels of power in studies with large effect sizes (.890).

TABLE 1
Frequency and Cumulative Percentage Distributions of Power Coefficients by Effect Size Using Rorschach Research: 1975–1991

	Small Effects		*Medium Effects*		*Large Effects*	
Power	*Frequency*	*Cumulative Percentage*	*Frequency*	*Cumulative Percentage*	*Frequency*	*Cumulative Percentage*
.99	1	100	6	100	54	100
.95–.98	–	99	7	96	25	66
.90–.94	–	99	4	92	14	50
.80–.89	–	99	10	89	20	41
.70–.79	–	99	23	83	12	29
.60–.69	–	99	25	68	13	21
.50–.59	–	99	16	53	10	13
.40–.49	2	99	18	42	3	6
.30–.39	5	98	22	31	5	4
.20–.29	17	95	20	17	2	1
.10–.19	65	84	6	4	–	–
.05–.09	68	43	1	–	–	–
n	158		158		158	
M	.131		.561		.850	
Median	.105		.570		.945	
SD	.104		.238		.186	
Q1	.080		.348		.738	
Q3	.150		.745		.990	

Note. Q1 = 25th percentile. Q3 = 75th percentile

TABLE 2
Frequency and Cumulative Percentage Distributions of Power Coefficients by Effect Size Using Rorschach Research, Comprehensive System Only: 1975–1991

	Small Effects		*Medium Effects*		*Large Effects*	
Power	*Frequency*	*Cumulative Percentage*	*Frequency*	*Cumulative Percentage*	*Frequency*	*Cumulative Percentage*
.99	1	100	3	100	20	100
.95–.98	–	98	1	93	6	54
.90–.94	–	98	1	91	5	40
.80–.89	–	98	3	88	3	28
.70–.79	–	98	8	81	4	21
.60–.69	–	98	12	63	1	12
.50–.59	–	98	3	35	1	9
.40–.49	1	98	2	28	2	7
.30–.39	1	95	6	23	1	2
.20–.29	7	93	3	9	–	–
.10–.19	24	77	1	2	–	–
.05–.09	9	21	–	–	–	–
n	43		43		43	
M	.166		.623		.890	
Median	.120		.650		.980	
SD	.154		.224		.167	
Q1	.100		.470		.860	
Q3	.190		.770		.990	

Note. Q1 = 25th percentile. Q3 = 75th percentile.

Table 3 presents frequencies and cumulative percentage distributions of the power to detect small, medium, and large effect sizes for Rorschach research excluding Comprehensive System methodology (n = 115). These findings suggest that in non-Comprehensive System research studies with small effect sizes were significantly underpowered, with about one chance in seven of detecting significant differences (.117), slightly more than one chance in two of detecting medium effect sizes (.538), and barely acceptable power for large effect sizes (.835).

Although there is some variation, it appears that, overall, Rorschach research has approximately the same power (.13, .56, and .85 for small, medium, and large effect sizes, respectively) as research performed in other disciplines of behavioral science (.26, .64, and .85 for small, medium and large effect sizes, Rossi, 1990).

To test the hypothesis that Rorschach research conducted according to the Comprehensive System yielded higher power, one-tailed t tests of the mean power estimates between Comprehensive System and non-Comprehensive System research were conducted. Visual analysis reveals that non-Comprehensive System research, with power coefficients of .117, .538, and

TABLE 3
Frequency and Cumulative Percentage Distributions of Power Coefficients by Effect Size Using Rorschach Research, Comprehensive System Excluded: 1975–1991

	Small Effects		*Medium Effects*		*Large Effects*	
Power	*Frequency*	*Cumulative Percentage*	*Frequency*	*Cumulative Percentage*	*Frequency*	*Cumulative Percentage*
.99	–	100	3	100	34	100
.95–.98	–	100	6	97	19	70
.90–.94	–	100	3	92	9	54
.80–.89	–	100	7	90	17	46
.70–.79	–	100	15	84	8	31
.60–.69	–	100	13	70	12	24
.50–.59	–	100	13	59	9	14
.40–.49	1	100	16	48	1	6
.30–.39	3	99	16	34	4	5
.20–.29	10	97	17	20	2	2
.10–.19	42	88	5	5	–	–
.05–.09	59	51	1	1	–	–
n	115		115		115	
M	.117		.538		.835	
Median	.090		.510		.920	
SD	.072		.240		.191	
Q1	.070		.330		.720	
Q3	.140		.720		.990	

Note. Q1 = 25 percentile. Q3 = 75th percentile.

.835 for small, medium, and large effect sizes, respectively, were smaller than Comprehensive System research, with power coefficients of .166, .623, and .890 for small, medium, and large effect sizes, respectively. Findings supported the hypothesis. Differences for small effect sizes were statistically significant, $t(156) = 2.68$, $p = .004$; as were differences between Comprehensive System and non-Comprehensive System research for medium effect sizes, $t(81) = 2.09$, $p = .02$; and differences for large effects sizes, $t(156) = 1.69$, $p = .047$.

DISCUSSION

Given the critical reviews of the Rorschach as a research tool, it is somewhat of a surprise to discover that Rorschach research yields findings similar to those observed in other areas of behavioral science. As Tables 1, 2, and 3 reveal, research in which there is a small effect size has no better than a 18% chance of attaining statistical significance, assuming that the effect is actually present. Medium effect sizes have about one or two chances out of three to attain significance. Large effect sizes have much better chances of detecting significant

findings. However, as discussed previously, most research yields effect sizes in the low to medium range. Furthermore, it appears that impact of studies based on the Comprehensive System have indeed yielded significant increases in the power of Rorschach research. This is apparently the first empirical assessment of the notion that the standardization of the test according to the Comprehensive System improves the quality of Rorschach research.

Clearly these findings have significance for the health of Rorschach research and likely reflect the artifactual controversy that has plagued the field. When medium effect sizes are obtained and power is at or near .50, an inconsistent pattern of findings emerges. To the academic researcher whose focus on the Rorschach is primarily through the lens of design methodology and statistics, the Rorschach fails to achieve the degree of respectability commonly expected in a test. These findings may confirm the reason for the uncomplimentary views found, for example, in the *MMYBs*.

What are the reasons for the poor showing? Undoubtedly, sample sizes are a primary culprit. Small sample sizes (the mode for this study was 40) seem to be the norm in Rorschach research, especially in studies in which power was low. However, there are several other features of Rorschach research that have to be taken into account. The first factor is error variance and its attenuating impact on power. Until the advent of the Comprehensive System for the Rorschach, reliability for certain test variables was quite low. Stringent attention to reducing error variance may help, including, for example, the use of kappa statistics for interrater reliability over percentage of agreement. Second, researchers should eschew whenever possible the use of nonparametric statistics, including the power sapping practice of dichotomizing variables (Cohen, 1973). As Zubin et al. (1965) commented in the late 1950s and 1960s, content scaling rather than determinants tend to yield the most respectable findings. Increasing sample size, reducing error, and relying as much as possible on parametric statistics may go far to increase power and design sensitivity in Rorschach research.

A surprising discovery of our investigation was the difficulty encountered in analyzing studies due to inconsistent reporting of findings. Our situation was quite similar to that observed by Katzer and Sodt (1973) who wrote "the most frustrating aspect of this study was our inability to quickly understand the statistical procedures employed in each journal article. Usually, this was due to missing information" (p. 261). This laxity on the part of investigators and editors strikes at the very heart of scientific inquiry, with its operational point of view and focus on the provision of sufficient evidence to allow independence of interpretation and replication. In the current era of meta-analysis, test statistics (including alpha levels, means, standard deviations, sample sizes, tests which are nonsignificant, and, preferably, effect sizes) are crucial parameters for reporting. It is unlikely that change will be forthcoming in this area without vigorous editorial enforcement.

However, the quest for greater power does not end here. In effect, concern with power and design sensitivity reflects the quest for more thoughtful research. To this extent, our investigation stands in a long tradition of Rorschach research critiques. Several recommendations, nevertheless, are indicated. First, based on theory and previous research, investigators should estimate the probable effect size of the study, ascertain the power of the design, and thereby determine whether the study can actually detect what is being studied. This involves asking if the effect size is likely to be small, medium, or large. If effect size cannot be ascertained in an a priori fashion, then small or medium effect size predictions should be the basis on which the sensitivity of the study is determined. Second, effect size should be viewed as something inherent in the well-planned study and that will be enhanced by sensitive design methodology, not as something discovered independent of the methods of investigation. Good theory, as numerous commentators have pointed out, is indispensable. Finally, consideration of the relative importance of Type I versus Type II errors should be considered, as well as a relaxation of alpha, though it is unlikely that the alpha equals .05 convention will change.

Several caveats are in order. First, the findings most likely are overestimates of statistical power because parametric tests were substituted for the large number of nonparametric tests. Second, this survey of the literature may not be comprehensive. Not included were studies published outside of the referenced journals. It is likely, however, that the journals covered in this survey represent the majority of Rorschach research published in the U.S. The neglect of power seems to be endemic in behavioral science research in general. Consequently, it is unlikely that statistical power in these studies will be significantly greater than what has been observed here. As knowledge and concern over statistical power becomes more pervasive (Goldstein, 1989), it is likely that the technology for assessing power, both predictively and postdictively, will become more accessible and implemented.

In many ways, we find ourselves in the uncomfortable, but familiar position of Cronbach in 1949, who stated that criticisms of the test were premature. The developments in Rorschach research stimulated by the Comprehensive System are an improvement and reason for psychometric optimism. It remains to current and future generations of Rorschach researchers to improve upon the past and design and carry out Rorschach research that clearly brings out the best that a test has to offer.

ACKNOWLEDGMENTS

Special thanks to Charles A. Peterson and Joseph Rossi for their insightful comments.

REFERENCES

Atkinson, L. (1986). The comparative validities of the Rorschach and MMPI: A meta-analysis. *Canadian Psychology, 27,* 238–247.

Atkinson, L., Quarrington, B., Alp, I. E., & Cyr, J. J. (1986). Rorschach validity: An empirical approach to the literature. *Journal of Clinical Psychology, 42*(2), 360–362.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin,* 66, 423–437.

Chase, L. J., & Tucker, R. K. (1976). Statistical power: Derivation, development, and data-analytic implications. *The Psychological Record, 26,* 473–486.

Borenstein, M., & Cohen, J. (1988). *Statistical power analysis: A computer program.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Cohen, J. (1962). The statistical power of abnormal–social psychological research: A review. *Journal of Abnormal and Social Psychology, 65,* 145–153.

Cohen, J. (1973). The cost of dichotomizing. *Applied Psychological Measurement, 7,* 249–253.

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

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

Cronbach, L. J. (1949). Statistical methods applied to Rorschach scores: A review. *Psychological Bulletin, 46,* 393–429.

Dana, R. (1965). The Rorschach. In O. Buros (Ed.), *The sixth mental measurements yearbook* (pp. 492–495). Highland Park, NJ: Gryphon.

Eron, L. (1965). The Rorschach. In O. Buros (Ed.), *The sixth mental measurements yearbook* (pp. 495–501). Highland Park, NJ: Gryphon.

Exner, J. E. (1974). *The Rorschach: A comprehensive system. Volume 1.* New York: Wiley.

Exner, J. E. (1978). *The Rorschach: A comprehensive system. Volume 2. Current research and advanced interpretation.* New York: Wiley.

Exner, J. E. (1986). *The Rorschach: A comprehensive system. Volume 1* (2nd ed.). New York: Wiley.

Exner, J. E., & Weiner, I. B. (1982). *The Rorschach: A comprehensive system. Volume 3. Assessment of children and adolescents.* New York: Wiley.

Eysenck, H. J. (1959). The Rorschach. In O. Buros (Ed.), *The fifth mental measurements yearbook* (pp. 276–278). Highland Park, NJ: Gryphon.

Fisher, R. (1935). *The design of experiments.* London: Oliver & Boyd.

Goldstein, R. (1989). Power and sample size via MS/PC-DOS computers. *American Statistician, 43,* 253–260.

Jensen, A. (1965). The Rorschach. In O. Buros (Ed.), *The sixth mental measurements yearbook* (pp. 501–509). Highland Park, NJ: Gryphon.

Katzer, J., & Sodt, J. (1973). An analysis of the use of statistical testing in communication research. *Journal of Communication, 23,* 251–265.

Kendall, P. C., & Norton-Ford, J. D. (1982). *Clinical psychology.* New York: Wiley.

Kish, L. (1959). Some statistical problems in research design. *American Sociological Review, 24,* 328–338.

Knutson, J. (1972). The Rorschach. In O. Buros (Ed.), *The seventh mental measurements yearbook* (pp. 435–440). Highland Park, NJ: Gryphon.

Lipsey, M. W. (1990). *Design sensitivity: statistical power for experimental research.* Newbury Park, CA: Sage.

Lubin, B., Larsen, R. M., & Matarazzo, J. D. (1984). Patterns of psychological test usage in the United States: 1935–1982. *American Psychologist, 39,* 451–454.

McArthur, C. (1972). The Rorschach. In O. Buros (Ed.), *The seventh mental measurements yearbook* (pp. 440–443). Highland Park, NJ: Gryphon.

McNemar, Q. (1960). At random: sense and nonsense. *American Psychologist, 15,* 295–300.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*, 806–834.

Neyman, J., & Pearson, E. (1928). On the use and interpretation of certain test criteria for the purposes of statistical inference. *Biometrika, 20A*, 175–240, 263–294.

Neyman, J., & Pearson, E. (1933). On the problem of the most efficient tests of statistical hypotheses. *Transactions of the Royal Society of London Series A, 231*, 289–337.

Parker, K. (1983). A meta-analysis of the reliability and validity of the Rorschach. *Journal of Personality Assessment, 47*, 227–231.

Parker, K. C., Hanson, R. K., & Hunsley, J. (1988). MMPI, Rorschach, and WAIS: A meta-analytic comparison of reliability, stability, and validity. *Psychological Bulletin, 103*(3), 367–373.

Rabin, A. (1972). The Rorschach. In O. Buros (Ed.), *The seventh mental measurements yearbook* (pp. 443–446). Highland Park, NJ: Gryphon.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 86*, 638–641.

Rosnow, R., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44*, 10 1276–1284.

Rossi, J. (1982, April). *Meta-analysis, power analysis, and artifactual controversy: The case of spontaneous recovery of verbal associations.* Paper presented at the meeting of the Eastern Psychological Association, Baltimore.

Rossi, J. (1983, April). *Inadequate statistical power: A source of artifactual controversy in behavioral research.* Paper presented at the meeting of the Eastern Psychological Association, Philadelphia.

Rossi, J. (1990). Statistical power of psychological research: What have we gained in 20 years? *Journal of Consulting and Clinical Psychology, 58*(5), 646–656.

Sargent, H. (1953). The Rorschach. In O. Buros (Ed.), *The fourth mental measurements yearbook* (pp. 213–218). Highland Park, NJ: Gryphon.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Journal of Consulting and Clinical Psychology, 105*(2), 309–316.

Zubin, J., Eron, L. D., & Schumer, F. (1965). *An experimental approach to projective techniques.* New York: Wiley.

Marvin W. Acklin
1564 Keolu Drive
Kailua, HI 96734

Received January 10, 1992