Child Development, January/February 2000, Volume 71, Number 1, Pages 173–180

Effect Size, Practical Importance, and Social Policy for Children

Kathleen McCartney and Robert Rosenthal

Real decisions for real children are influenced by the papers developmentalists write, regardless of whether we ever intended our papers to be used in the policy arena. Yet most social scientists seldom analyze data in ways that are most useful to policymakers. The primary purpose of this paper is to share three ideas concerning how to evaluate the practical importance of a finding or set of findings. First, for research to be most useful not only in the policy arena but also more generally, significance tests need to be accompanied by effect size estimates. The practical importance of an effect size depends on the scientific context (i.e., measurement, design, and method) as well as the empirical literature context. Second, researchers need to use all existing data when weighing in on a policy debate; here, meta-analyses are particularly useful. Finally, researchers need to be careful about embracing null or small findings, because effects may well be small due to measurement problems alone, particularly early in the history of a research domain.

INTRODUCTION

Politicians, constituents, lobbyists, media analysts, and advocates all attempt to influence social policy for children (Hayes, 1982). Often, policymakers are influenced by compelling argument alone. Policymakers have also turned to social science research to guide their decision making about public expenditures for children's programs (Weiss & Bucuvalas, 1980). Descriptive research on incidence rates for problems such as poverty, malnutrition, and school failure provides a signal to policymakers about the need for intervention (Kamerman, 1996). Evaluation research informs policymakers about the benefits of programs (McCall, Green, Strauss, & Groark, 1998). Even basic research is sometimes brought to bear on policy decisions; for example, research on brain development has focused the attention of policymakers on the need for early care and education in the first three years of life (Bruer, 1998).

Developmentalists need to be cognizant of the fact that real decisions for real children are influenced by the papers we write, regardless of whether we ever intended our papers to be used in the policy arena. For this reason, it is incumbent upon us to consider how others use our data. Unfortunately, data sometimes seem to serve as stimuli for a projective test. Sometimes, differences in opinions are real and other times they reflect differences in political ideologies. One of the best examples of this can be found in the history of Head Start, where evaluation data have been interpreted to support a variety of positions (see Zigler & Muenchow, 1992).

Even setting politics aside, researchers accustomed to publishing in research journals seldom analyze data in ways that are most useful to policymakers. Indeed, the current ferment in the behavioral sciences about the use of statistics raises the broader question of whether some of our current practices in data analysis are of much use to anyone (Schmidt, 1996; Thompson, 1999). Most of our discipline's data-analytic work only addresses the question of whether there is an association between two variables, such as participation in a given program and a child outcome. In addition, our work needs to address the question of whether the association is of practical importance. The primary purpose of this paper is to share our ideas on how to evaluate the practical importance of a finding or set of findings.

RESISTANCE TO THE USE OF EFFECT SIZE ESTIMATES

Historically, social scientists have relied on significance testing as a data-analytic tool. In significance testing, a null hypothesis is specified (usually that there is no relation between two or more variables) and statistical probabilities are used either to reject or fail to reject a null hypothesis. The significance level, or *p*-level, provides an index of the probability that one has mistakenly rejected the null hypothesis. We owe null hypothesis testing to R. A. Fisher, and as Cohen (1990) has noted, the attractiveness of Fisher's ideas is hardly surprising—"they offered a deterministic scheme, mechanical and objective, independent of content, and led to clear-cut yes-no decisions" (p. 1307). Policymakers as well as researchers sometimes mistakenly confuse the significance of a finding with the importance of a finding. A significance test that results in a small *p* does not necessarily connote an important finding; it merely denotes that the null hy-

pothesis is unlikely to be true. Recently, the American Psychological Association Board of Scientific Affairs Task Force on Statistical Inference (1998) has urged researchers to offer "enhanced characterization" of inferential tests. By this, they mean that researchers should accompany their *p* values with an estimate of both the direction and size of an effect. An effect size estimate denotes the size of an effect in standard units. It provides a first step toward evaluating the practical importance of a finding.

Old news, one might say? True. Many have argued that effect size estimates should be routinely reported not only for significant findings but also for nonsignificant ones. Some go further by noting that the primary product of any research inquiry should be an effect size, and not a *p* value (Cohen, 1990). Some go further still by arguing that effect size estimates with confidence intervals should replace significance tests (Schmidt, 1996). As early as 1982 an outgoing journal editor once compared *p* values to mosquitoes and lamented that "no amount of scratching, swatting or spraying will dislodge them" (Campbell, 1982, p. 698). Thompson (1999) notes that the field's overall resistance to reporting effect size probably lies in the fact that journal editors have merely encouraged such reporting rather than having required it.

There is resistance to reporting effect size estimates at the level of the individual as well. We believe that there may be two primary reasons for this. The first is that most researchers do not know enough about how to compute and report effect size estimates. The second is that neither experienced researchers nor experienced statisticians have a good intuitive feel for the practical meaning of common effect size estimates. As a result, most effects are labeled small or trivial, which is a discouraging conclusion; it is also a mistaken one. This state of affairs needs to be rectified not only with respect to policy analysis but also more generally.

HOW TO COMPUTE AND REPORT EFFECT SIZES

The computation of an effect size estimate is straightforward and easy. It typically involves pulling a number from significance test results or juggling several numbers readily available from test results. Table 1 offers definitions and descriptions for common effect size indicators used in social science research—these indicators will suffice for most developmental work (for a more complete review of work on effect size estimates, see Cooper & Hedges, 1994). It should be noted that the effect size estimates described in this paper are all based on 1 degree of freedom (*df*) tests (e.g., a comparison between two groups or a test for a predicted trend in three or more groups). There are few reasons ever to report an effect size estimate based on an *F* with 2 or more *df* in the numerator (i.e., a comparison among 3 or more groups) or a χ^2 with 2 or more *df*, because these effect size indexes do not tell "where the action is." Such omnibus analyses require further 1-*df* contrasts, such as a contrast between two of the groups or a linear trend (Rosenthal & Rosnow, 1985; Rosenthal, Rosnow, & Rubin, 1999).

Table 1 Two Families of Effect Size Indicators: Some Examples

Statistic	Definition	Description
r-Family		
r	Pearson product moment correlation	Direction and magnitude of effect
p*r*	Partial correlation for two variables with the effect(s) of one or more other variables removed	Direction and magnitude of effect
d-Family		
Cohen's *d*	$\frac{M_1 - M_2}{\sqrt{\frac{SS_{within}}{N}}}$	Standardized difference between means
Hedges' *g*	$\frac{M_1 - M_2}{\sqrt{\frac{SS_{within}}{N-1}}}$	Standardized difference between means
d'	$p_1 - p_2$	Difference between proportions

Note that there are two types, or families, of effect size estimates: *r* is typically used when the relation is assessed via a correlation, and *d* is typically used when the relation is assessed via comparison of group means (see Rosenthal, 1994). In practice, although *r*-type and *d*-type estimates can readily be converted to one another, meta-analysts tend to prefer the *r* type, because of its greater interpretability when contrasts among three or more groups are employed. When only two groups are compared, *r* and *d* are both directly interpretable.

r Family. The *r* family gets its name from the Pearson product moment correlation, a measure of association between two variables. In addition to the ordinary Pearson product moment correlation, which is used when both variables are in continuous forms, there are other Pearson correlations: ϕ (phi coefficient) when both variables are dichotomous, r_{pb} (point biserial coefficient) when one variable is continuous and one is dichotomous, and ρ (rho) when both variables are in ranked form. When using multiple regression equations to test hypotheses, researchers should report effect sizes for predictors. For each predictor, there is a regression coefficient, or *B*, that represents

the amount of change in the outcome variable associated with a one-unit change in the predictor variable. And for each regression coefficient, there is a corresponding t test that assesses the degree to which it differs from 0. A partial correlation, which serves as an effect size estimate, can be easily computed from the t statistic, in this case and more generally:

$$r = \sqrt{\frac{t^2}{t^2 + df}}$$

Other conversion formulas can be used to compute r as an effect size estimate from common test statistics, such as chi-square and F tests. This is often an early step in meta-analytic work.

$$r_\phi = \sqrt{\frac{\chi^2(1)}{N}}$$

$$r = \sqrt{\frac{F(1, -)}{F(1, -) + df_{\text{error}}}}$$

d Family. The d family gets its name from Cohen's d, which is the difference between two means divided by the pooled standard deviation for the sample. The index, d, is in standard units. If a treatment were associated with an effect size of $d = .33$, a researcher could then say that the treatment led to a one-third standard deviation increase in an outcome. There are numerous variations on d, most notably Hedges' g, which differs from d in that it uses $N - 1$, rather than N, as the within-group divisor for the sums of squares. Whether to use d or g is largely a question of preference. The last indicator in Table 1, d', is the difference between two proportions, an intuitive indicator of effect size.

Squared indexes of effect sizes. Notice that we have not included any squared indexes of effect sizes, i.e., proportion of variance in one variable that is predicted or explained by another variable (or set of variables), such as r^2, η^2, ω^2, ε^2, and so on. Ozer (1985) has demonstrated that the sum of absolute deviations from the mean is a better measure of variation than the sum of squared deviations from the mean; further, he argued that using squared indexes for effect sizes leads to interpretations that "grossly underestimate" the magnitude of associations. In other words, using squared indexes makes even large effects seem small. Because many researchers have developed the habit of reporting squared indexes of effect size, we should note that reporting them will do no harm as long as we also report the more directly interpretable indexes from the r family or the d family.

Always compute an effect size estimate. It is important to mention that an effect size estimate can be computed regardless of whether "significance" is obtained. As Folger (1989) notes, an effect size estimate is more useful for making judgments than a binary choice between significance and nonsignificance. One reason this constitutes good practice is that readers may conclude that a nonsignificant finding has an effect size of 0; this demonstrates faulty logic. For example, suppose d is equal to .20 for an analysis in which the researcher fails to reject the null hypothesis. It is just as likely that the true d for the effect is .40 as it is that it is 0. Rosenthal and Rubin (1994) introduced a new statistic, the counternull value of an effect size (.40 in this example), to emphasize this point. Reporting confidence intervals for effect sizes helps readers evaluate how good the estimate actually is.

WHAT IS THE PRACTICAL MEANING OF AN EFFECT SIZE ESTIMATE?

Once an investigator computes an effect size, the question remains how to describe its importance. Cohen's (1969) well-cited conventions for psychological investigations continue to guide most developmentalists: an r of .10 is small, .30 moderate, and .50 large; a d of .20 is small, .50 moderate, and .80 large. These guides are sometimes useful, especially when conducting power analyses, which is why Cohen offered them; however, when they are applied blindly, results with material import could erroneously be dismissed as trivial. In actuality, researchers examining the association between two variables seldom obtain rs as large as .50 or ds as large as .80. This has led many researchers to the discouraging conclusion that all their findings from the lab and from the field are small. We may all be doing far better than we have been led to believe, however.

An assessment of the practical importance of an effect is a necessary second step following the computation of an effect size, especially with respect to policy analysis. There are no easy conventions for determining practical importance. Just as children are best understood in context, so are effect sizes. There are two main contexts for an effect size, a scientific and an empirical literature context.

Scientific context. Ideally, social science would inform social policy completely. By that we mean that policy decisions would be data based. Basic research on developmental processes and applied research on program evaluation would be brought to bear on policy debates. To the extent that policy should reflect a culture's values and morality (McCartney, Phillips, & Scarr, 1993), social science research on attitudes could be used to inform policy. The problem, however, is

that social science research is in its infancy. Clearly, the size of an effect is to some extent dependent on limitations in operations used to generate the data, specifically operations associated with measurement, design, and method.

Kagan (1998) equated trying to understand human processes via questionnaires to "trying to understand the galaxies with the naked eye, without the advantages of a telescope." Theory and conceptualization in social science far exceed measurement. Many of the constructs in our field cannot be measured with very high reliability. Measurement error biases effect size estimates downward toward zero and so the psychometric properties of measures provide one context for effect size estimates.

Measurement error takes place whenever anything is scaled. As O'Grady (1982) noted, even a seemingly straightforward variable like sex is scaled; it can be measured in numerous ways (chromosomal, gonadal, hormonal, external genital appearance, assigned, gender identity, or choice of a sex partner), each of which includes error. All operationalizations of constructs, from attitudes measured with questionnaires to behaviors measured in the field to manipulations defined in the laboratory, include error. Reflecting on psychology's measurement woes, O'Grady argued that all studies in our field can in some way be thought of as construct validity work. He further admonished that "much psychological research, at least for the near future, will probably produce small measures of explained variance because of measurement problems alone" (p. 770).

This is a crucial point to remember in the policy arena. For example, consider the policy implications of research on child care. Recent reviewers, relying on Cohen's conventions, agree that effect sizes for child-care quality are small (e.g., Lamb, 1998). As such, Scarr (1999) concludes that "child care quality within the range of American child care centers does not have important impact on the development of children from ordinary homes" and argues that states should examine the impact of their regulations "on making child care affordable, available, and of sufficient quality to support good child development" (Scarr, 1998, p. 105). Scarr's conclusions might be somewhat more warranted if the psychometric properties associated with measures of constructs like child-care quality were already maximal.

Design can also influence effect size. There are two reasons that experimentation is always desirable: causality can be established and effect sizes are more accurate. The type of experiment performed will influence the size of an effect. Generally, experiments in the laboratory involve weak manipulations because they involve mere simulations, whereas experiments in the field often involve interventions that are likely to produce larger effects. Experimentation is often not possible, however, either for ethical or practical reasons, when examining the social ecology of children. Researchers cannot randomly assign children to the kinds of contexts and experiences that interest policymakers: poverty, child care, family violence, peer rejection, adolescent pregnancy, and so on. Effect sizes from nonrandomized and quasi-experimental designs are likely to be biased and researchers must consider probable sources of bias, such as possible third variables that moderate an association between the two variables under investigation (see Campbell & Stanley, 1963).

Methodological choices will influence effect sizes as well (see O'Grady, 1982). Decisions that minimize an error term will increase an effect size, for example the use of a within-subject versus a between-subject design. Decisions about the predictor variable also matter: the greater the range of scores on a predictor variable in a nonrandomized study (or the greater the spread among levels of a fixed-effect factor in an experimental study), the greater the effect size is likely to be.

The bottom line is that we need to consider issues concerning measurement, design, and method in evaluations of findings for policy formulations as well as more generally. Not surprisingly, better measurement, design, and method all tend to produce larger effects.

Empirical literature context. Results from a given study need to be evaluated in the context of the existing empirical literature. One of the best tools that researchers have to summarize a body of literature is meta-analysis, which is an analysis of analyses on a common research question (for a useful introduction to meta-analysis, see Rosenthal, 1991). Clearly, cumulative research knowledge from a set of studies is more reliable than findings from selected studies or a single study. Policymakers have embraced meta-analysis. By offering an average effect size from a set of studies, meta-analysis eliminates the problem of what to do when studies show opposing findings (Hunter & Schmidt, 1996).

As part of the presentation of results from meta-analyses, studies are typically weighted by sample size, so that larger studies will have more influence than smaller ones. Additionally, contrasts can be conducted to assess heterogeneity in quality associated with differences in methodology (e.g., nonrandomized versus randomized designs), measurement (e.g., self-report versus other report), and participants (e.g., working class versus middle class). To conduct a developmental meta-analysis, the average age of par-

ticipants can be correlated with effect sizes across studies, as McCartney, Harris, and Bernieri (1990) demonstrated in their twin meta-analysis.

Sometimes it is useful to compare effect sizes across domains. For example, Rosenthal (1994) has used the results of a randomized double-blind experiment on the effects of aspirin in reducing heart attacks (Steering Committee of the Physicians' Health Study Research Group, 1988) as a dramatic comparison for behavioral effects such as psychotherapy. The effect size in the aspirin study was small by any conventional standard, $r = .03$. The effect size for psychotherapy treatment had been considered small by some, $r = .32$, but perhaps not so small when compared with the aspirin study. Yet, is .03 small? The Steering Committee of the Physicians' Health Study Research Group did not think so. In fact, they terminated the study prematurely for ethical reasons, that is, to advise the control group on the benefits of aspirin to prevent heart attacks. The aspirin treatment effectively demonstrates a situation in which a small effect was thought by the investigators to have great importance. No doubt this research team was influenced by the fact that the cost of the treatment was also small (i.e., pennies a day) and the benefit was great (i.e., life versus death). Certainly, we would not suggest that all effect sizes of .03 are noteworthy; rather, we hope to demonstrate that any effect associated with life versus death, even one with a small effect size, has practical importance, especially when it is associated with an inexpensive treatment. No criterion can be developed to separate small, useless effects from small, useful ones; researchers need to evaluate effect sizes using logic and argument.

Following Rosenthal's example, the NICHD Early Child Care Research Network (1999) compared effect sizes for child-care quality with those for home quality. At the heart of this comparison was the notion that the home quality effect sizes would provide a kind of upper bound for an environmental effect, not only because of the obvious importance of the home environment but also because the home environment reflects both genetic and environmental influence. In the cognitive domain, the child-care effects were small by Cohen's guidelines (e.g., school readiness, $r = .14$); the home quality effects were about twice the size. The research team argued that the child-care effects have practical importance given that they are as large as half the size of home effects, which are probably inflated given that they reflect both environmental and genetic effects.

Making effect sizes accessible: the Binomial Effect Size Display. Policymakers vary greatly in their statistical knowledge. For those with less knowledge, Rosenthal and Rubin (1982) have developed a procedure to display the practical importance of r to public audiences, called the Binomial Effect Size Display (BESD). The BESD is not an effect size; rather, it is a means to display an effect size. To present an effect size as a BESD, research findings are cast into dichotomous results. Consider a generic example where a researcher is comparing success and failure rates between treatment and control groups. The correlation is shown to be a simple difference in outcome rates between the treatment and control groups in a standard table which always adds up to column totals of 100 and row totals of 100. The treatment condition success rate is computed as $50 + (r \times 100)/2$, and the control condition success rate is computed as $50 - (r \times 100)/2$.

Any correlation can be cast into a BESD, even one that involves continuous variables (Rovine & von Eye, 1997); here, the table might be labeled something like high/low groups by above average/below average outcome scores. A BESD for such a study, with an $r = .14$, as in the NICHD school readiness finding, appears in Figure 1. One can readily see the difference in success rates (i.e., those in the above average group) for the high and low treatment groups. The BESD is a good way to explain research results to policymakers and the public generally because increases in success rate can be understood by everyone. Increases in success rates corresponding to various values of r^2 and r can be found in Table 2.

COSTS AND BENEFITS

Policymakers often assess practical importance in dollars rather than in terms of formal effect size estimates. Understandably, policymakers are impressed by cost:benefit analyses that demonstrate a high return on investment. The High/Scope Longitudinal studies have successfully demonstrated return on the investment in their high-quality, active-learning preschool program for children at high risk of school failure (Schweinhart, 1999). A follow-up study of partic-

		Outcome		
		Above-Average Outcome Scores	Below-Average Outcome Scores	Total
Predictor	High Group	57	43	100
	Low Group	43	57	100
	Total	100	100	200

Figure 1 Binomial effect size display for $r = .14$.

Table 2 Increases in Success Rate Corresponding to Various Values of r^2 and r

r^2	r	Success Rate Increased		Differences in Success Rates (r)
		From	To	
.00	.02	.49	.51	.02
.00	.04	.48	.52	.04
.00	.06	.47	.53	.06
.01	.08	.46	.54	.08
.01	.10	.45	.55	.10
.01	.12	.44	.56	.12
.03	.16	.42	.58	.16
.04	.20	.40	.60	.20
.06	.24	.38	.62	.24
.09	.30	.35	.65	.30
.16	.40	.30	.70	.40
.25	.50	.25	.75	.50
.36	.60	.20	.80	.60
.49	.70	.15	.85	.70
.64	.80	.10	.90	.80
.81	.90	.05	.95	.90
1.00	1.00	.00	1.00	1.00

Note: Reprinted with permission from *Essentials of Behavioral Research: Methods and Data Analysis* (p. 210), by R. Rosenthal & R. L. Rosnow, 1984, New York: McGraw-Hill.

ipants at 27 years of age showed a $7.16 return to the public per dollar invested in the program. The savings are attributed to the fact that the experimental group required fewer services associated with the criminal justice, mental health, and welfare systems. Similarly, Dodge (personal communication, April 9, 1998) and his colleagues have planned a cost:benefit analysis to assess the usefulness of Fast Track, an extensive intervention for children at risk for conduct disorder. The intervention itself is expensive, approximately $40,000 per child over its ten-year course. The cost of services for children with untreated conduct disorder is expensive as well, however, especially for those who adopt a life of crime—chronic criminals cost society 1.3 million dollars each across their careers. Dodge argues that Fast Track will prove cost effective if the rate of chronic criminality is reduced by a mere 3%. In order for cost:benefit analyses like these to be persuasive, policymakers, and the citizens who influence them, must be persuaded to invest in prevention efforts that may not be apparent for a generation or more.

Because estimating costs and benefits is often not straightforward, it is important to note that methodological issues abound here as well. These analyses always involve assumptions about true costs and benefits, assumptions that often are debatable. Finally, however, it is important to note that a psychologist's job is not primarily to determine what is cheap, but to determine what is true. We have no special expertise in determining how members of a society should or should not spend funds. We can, however, choose to work with economists to do policy-relevant analyses. Developmentalists who advocate for children's programs will probably experience increasing pressure to do so.

CONCLUSION

Issues surrounding the uses and misuses of data to guide social policy go beyond the field of child development per se; they are, of course, science-wide issues. We offer three suggestions here. First, in order for our work to be most useful not only in the policy arena but also more generally, significance tests need to be accompanied by effect size estimates. The move from significance testing to effect size estimation is straightforward; the move from effect size estimation to the assessment of the practical importance of findings is not. Considerations of the scientific context (i.e., measurement, design, and method) as well as the empirical literature context should be factored into assessments of practical importance.

Second, we need to use all existing data when weighing in on a policy debate. Meta-analyses are particularly informative, especially those with contrasts to assess heterogeneity in methodological quality as a moderator variable. Sometimes, our knowledge base will not be good enough to inform policy debates. In fact, some might argue that this is often the case. Nevertheless, as Edward Zigler has noted again and again, policies for children will be made with or without input from researchers. While our field develops, we should feel free to use the existing knowledge base and render reasonable opinions to policymakers; certainly, we are free to do so as citizens.

Our third point is that effects are often likely to be small as a result of methodological limitations in social science research, especially with respect to measurement. Therefore, researchers and policymakers alike should be careful about embracing null or small findings. As Smith (1999) observed, many politicians are pleased to accept such judgments as a justification for their reluctance to allocate funds for children. Given that the stakes are so high, we should be wary of accepting the null hypothesis when it might very well be false—as it almost always is (Cohen, 1994).

Developmentalists who have entered into policy debates are typically humbled by the fact that the rules of evidence differ for policymakers and researchers; whereas policymakers can be swayed by a compel-

ling story, researchers rely—at least in principle—on data to draw conclusions (March, 1979; Phillips, in press). More and more, however, our data are becoming part of policymakers' stories. When we allocate funds to programs that do not work, we waste precious resources. When we fail to allocate funds to programs that can and do work, we allow children to be at risk, and the greater the risk, the greater our concern should be. It is no easy feat to balance the need to spend resources wisely with the need to promote the healthy development of children, yet this is our task.

In a brilliant essay, Jacob Cohen (1990) reflected on the statistical lessons he has learned and offered the following advice: "Finally, I have learned that there is no royal road to statistical induction, that the informed judgment of the investigator is the crucial element in the interpretation of data, and that things take time" (p. 1304). Let us use our best judgment when we bring research to bear on policy questions—and, when we do, let us take the time to evaluate effect sizes in context.

ACKNOWLEDGMENTS

The authors are grateful to Kristen Bub, Margaret Burchinal, Eric Dearing, Bill Hagen, and Deborah Phillips for their helpful comments on a previous version of the paper. The work of the first author was supported by a grant from the National Institute of Child Health and Human Development (HD 25451).

ADDRESSES AND AFFILIATIONS

Corresponding author: Kathleen McCartney, Department of Psychology, University of New Hampshire, Durham, NH 03824; e-mail: kathleen.mccartney@unh.edu. Robert Rosenthal is at the University of California at Riverside, Riverside, CA.

REFERENCES

American Psychological Association Board of Scientific Affairs. (1996, December). *Task Force on Statistical Inference initial report.* Washington, DC: Author.

Bruer, J. T. (1998). Brain science, brain fiction. *Educational Leadership,* 56, 14–18.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Campbell, N. (1982). Editorial: Some remarks from the outgoing editor. *Journal of Applied Psychology, 67*, 691–700.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences (*Rev. ed.). New York: Academic Press.

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.

Cooper, H., & Hedges, L. V. (Eds.). (1994). *The handbook of research synthesis.* New York: Russell Sage Foundation.

Folger, R. (1989). Significance testing and the duplicity of binary decisions. *Psychological Bulletin, 106*, 155–160.

Hayes, C. D. (1982). Making policies for children: A study of the federal process. Washington, DC: National Research Council.

Hunter, J. E., & Schmidt, F. L. (1996). Cumulative research knowledge and social policy formation: The critical role of meta-analysis. *Psychology, Public Policy, and Law, 2*, 324–347.

Kagan, J. (1998, September 13). A parent's influence is peerless. *Boston Globe,* p. E3.

Kamerman, S. B. (1996). The new politics of child and family policies. *Social Work, 41*, 453–465.

Lamb, M. E. (1998). Nonparental child care: context, quality, correlates, and consequences. In I. E. Sigel & K. A. Renninger (Eds.), W. Damon (Series Ed.), *Handbook of child psychology: Vol. 4* (pp. 73–134). New York: Wiley.

March, J. G. (1979). Science, politics, and Mrs. Gruenberg. In *The National Research Council in 1979* (pp. 27–36). Washington, DC: National Academy of Sciences.

McCall, R. B. J., Green, B. L., Strauss, M. S., & Groark, C. (1998). Issues in community-based research and program evaluation. In I. E. Sigel & K. A. Renninger (Eds.), W. Damon (Series Ed.), *Handbook of child psychology: Vol. 4* (pp. 955–997). New York: Wiley.

McCartney, K., Harris, M. J., & Bernieri, F. (1990). Growing up and growing apart: A developmental meta-analysis of twin studies. *Psychological Bulletin, 107*, 226–237.

McCartney, K., Phillips D., & Scarr, S. (1993). On using research as a tool. *American Psychologist, 48*, 691–692.

NICHD Early Child Care Research Network. (1999, April). *Effect sizes from the NICHD Study of Early Child Care.* Paper presented at the Biennial Meetings of the Society for Research in Child Development, Albuquerque, NM.

O'Grady, K. E. (1982). Measures of explained variance: Cautions and limitations. *Psychological Bulletin, 92*, 766–777.

Ozer, D. J. (1985). Correlation and the coefficient of determination. *Psychological Bulletin, 97*, 307–315.

Phillips, D. (in press). Social policy and community psychology. In J. Rappaport & E. Seidman (Eds.), *Handbook of community psychology.* New York: Plenum.

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

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis.* New York: Russell Sage Foundation.

Rosenthal, R., & Rosnow, R. L. (1985). *Contrast analysis: Focused comparisons in the analysis of variance.* New York: Cambridge University Press.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (1999). *Contrasts and effect sizes in behavioral research: A correlational approach.* New York: Cambridge University Press.

Rosenthal, R., & Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology, 74*, 166–169.

Rosenthal, R., & Rubin, D. B. (1994). The counternull value of an effect size: A new statistic. *Psychological Review, 5*, 329–334.

Rovine, M. J., & von Eye, A. (1997). A 14th way to look at a correlation coefficient: Correlation as proportion of matches. *American Statistician, 51*, 420–425.

Scarr, S. (1998). American child care today. *American Psychologist, 53*, 95–108.

Scarr, S. (1999, April). *What can we tell working parents?* Paper presented at the biennial meetings of the Society for Research in Child Development, Albuquerque, NM.

Schweinhart, L. J. (1999, April). *Generalizing from high/scope longitudinal studies.* Paper presented at the biennial meetings of the Society for Research in Child Development, Albuquerque, NM.

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

Smith, S. (1999, April). Discussant on symposium, Selection Issues in Child Care: Estimating the effects of child care on child outcomes, presented at the Biennial Meetings of the Society for Research in Child Development, Albuquerque, NM.

Steering Committee of the Physicians' Health Study Research Group. (1988). Preliminary report: Findings from the aspirin component of the ongoing physicians' health study. *New England Journal of Medicine, 318*, 262–264.

Thompson, B. (1999). Why "encouraging" effect size reporting is not working: The etiology of researcher resistance to change in practices. *The Journal of Psychology, 133*, 133–140.

Weiss, C., & Bucuvalas, M. J. (1980). *Social science research and decision making*. New York: Columbia University Press.

Zigler, E., & Muenchow, S. (1992). *Head Start: The inside story of America's most successful educational experiment.* New York: Basic Books.