Psychological Methods
2001, Vol. 6, No. 4, 413–429

1082-989X/01/$5.00 DOI: 10.1037//1082-989X.6.4.413

The Role of Method in Treatment Effectiveness Research: Evidence From Meta-Analysis

David B. Wilson
George Mason University

Mark W. Lipsey
Vanderbilt University

A synthesis of 319 meta-analyses of psychological, behavioral, and educational treatment research was conducted to assess the influence of study method on observed effect sizes relative to that of substantive features of the interventions. An index was used to estimate the proportion of effect size variance associated with various study features. Study methods accounted for nearly as much variability in study outcomes as characteristics of the interventions. Type of research design and operationalization of the dependent variable were the method features associated with the largest proportion of variance. The variance as a result of sampling error was about as large as that associated with the features of the interventions studied. These results underscore the difficulty of detecting treatment outcomes, the importance of cautiously interpreting findings from a single study, and the importance of meta-analysis in summarizing results across studies.

Systematic knowledge about the effectiveness of psychological and behavioral intervention depends almost exclusively on studies using experimental or quasi-experimental research designs. Among such designs, the well-executed randomized experiment is widely considered the gold standard because it is expected to produce an estimate of the mean treatment effect on a given dependent variable that deviates from the true value only by random error, which is kept small when statistical power is adequate. Unfortunately, when conducting treatment effectiveness research in real-world settings, ideal experimental design often cannot be attained: Randomization is incomplete, is undone by attrition, or is unethical or impractical; sample sizes are not sufficient to keep sampling error small relative to treatment effects; experimental control of conditions is lax or impossible; dependent variables are limited by low reliability or do not represent the outcome construct well; and so forth (Conrad, 1994; Cook & Shadish, 1994; Dennis, 1990; Kazdin, 1992; Lipsey & Cordray, 2000).

Each such departure from the ideal potentially degrades the treatment effect estimate. But by how much? In what direction? Under what circumstances? The statistical and epistemological theory that underlies experimental design supports the claim that ideal design will yield valid estimates but provides little basis for appraising the consequences of various departures from the ideal. Indeed, the nature and magnitude of those consequences are largely empirical matters, but they can be investigated directly only with the results of ideal designs in hand as a standard of comparison, a difficult condition to fulfill. An alternative way to conceptualize the empirical question is in signal-detection terms. Research in a typical intervention domain investigates the effects of different treatment variants on different outcome constructs for different respondent samples using different methods and procedures. Some of the variation in observed effects across these studies stems from differences in substantive aspects of the intervention being investigated (treatment, outcome construct, respondents); this variation represents the "signal" the researcher wishes to detect. The remaining variation sterns from differences in method or from randomly distributed sampling and measurement error; this variation represents the "noise" that potentially distorts or obscures the signal the researcher is attempting to detect. Over

David B. Wilson, Administration of Justice Program, George Mason University; Mark W. Lipsey, Department of Psychology and Human Development, Peabody College, Vanderbilt University.

This work was supported in part by National Institute of Mental Health Grant RO1-MH51701.

Correspondence concerning this article should be addressed to David B. Wilson, Administration of Justice Program, George Mason University, 10900 University Boulevard, MS 4F4, Manassas, Virginia 20110-2203.

the typical range of substantive and method differences in a body of research, it would be informative to know that the proportion of variance in the observed effect sizes associated with the signal was large relative to that associated with the noise. This would tell us whether estimates of treatment effects are robust in the face of method variation and random error of the sort typical in the treatment domain examined; that is, how threatening are departures from ideal design to the resulting conclusions about treatment effects?

Obtaining a good empirical estimate of this signal-to-noise ratio would require an experiment on experiments. In this experiment, there would be a factor for each design feature of interest (e.g., type of assignment to conditions, treatment variant, type of outcome measure), each varied over a range typical of actual practice. Researchers would then be randomly assigned to these various factor levels and would be required to conduct a study using the stipulated configuration of substantive and method features. Each observation within this factorial design would thus reflect the results of an entire outcome study conducted using the research methods specified by the grand experiment.

It would be especially informative if the experiment on experiments identified the specific method features with the greatest potential to distort estimates of treatment effects and the circumstances in which those distortions were most likely. Designing treatment effectiveness studies often involves making trade-offs with respect to ideal design, and such information would give useful guidance to researchers about which compromises were likely to introduce serious error.

Of course, the experiment on experiments is not practical, but the issue it would address can be examined, albeit less definitively, by analyzing the results of multiple studies within a treatment domain in relation to the naturally occurring method and substantive variation across those studies. Meta-analysis does just this (Cook et al., 1992; Cooper & Hedges, 1994; Hunter & Schmidt, 1990; Rosenthal, 1991). A meta-analysis of a particular treatment domain, therefore, can be viewed as a quasi-experimental alternative to the experiment on experiments. Typically, method and substantive features vary across studies within a specific treatment research context, and the differences in the mean effect sizes associated with the method and substantive dimensions indicate their relative contribution to that variation. Analysis of such data can provide an assessment of the potential biases associated with different method features within the respective treatment domain (for similar approaches, see Heinsman & Shadish, 1996; Shadish & Ragsdale, 1996).

The findings resulting from this procedure, however, would be limited to the particular treatment domain for which the meta-analysis was done. Greater generality would be possible if multiple treatment domains were examined. This can be accomplished through a synthesis of the substantive and method effect size breakouts across multiple meta-analyses.

The current study uses the body of meta-analyses identified by Lipsey and Wilson (1993) to construct a meta-analysis of meta-analyses analog to the experiment on experiments described previously. This body of 319 meta-analyses encompasses 16,525 separate studies of the effects of psychologically based treatments, predominately mental health and educational interventions but with considerable diversity within those categories. This diversity provides the basis for a broad examination of the role of method factors relative to substantive factors in accounting for the effects observed in studies of psychological intervention. Thus, the research question addressed by this study is, what is the role of method in treatment effect estimates? More specifically, what is the influence of method features relative to substantive intervention features on observed study outcomes?

Method

Identification and Retrieval of Meta-Analyses

A variety of search strategies were used to identify meta-analyses of psychologically based interventions reported between 1976, the year of Glass's pioneering work (Glass, 1976; Smith & Glass, 1977), and mid-1991. Potential meta-analyses were identified mainly through a computerized search of the following databases: Academic Index, Ageline, British Education Index, Child Abuse and Neglect, Criminal Justice Periodical Index, Dissertation Abstracts, ECER/EXCEP Child, ERIC, Family Resources, Mental Health Abstracts, NCJRS, PAIS International, Population Bibliography, PsycINFO, Public Opinion Online, Religion Index, Social Scisearch, Sociological Abstracts, and U.S. Political Science Documents. Search terms included variations on meta-analysis (e.g., meta-analysis, metaanalysis, meta-analytic) and variations on quantitative review (e.g., quantitative review, quantitative synthesis). Meta-analyses were also identified through the references in articles reporting or

discussing meta-analysis and through contact with other meta-analysts.

Selection Criteria

To be eligible for inclusion, a meta-analysis had to meet three criteria. First, it had to provide standardized mean difference effect sizes for treatment–control contrasts or statistics from which such effect sizes could be derived. Second, the studies meta-analyzed had to present research on the effects of treatments that manipulated psychological variables to produce psychological change. Third, those treatments had to represent types of interventions that are currently applied in practical domains (e.g., psychotherapy, parent effectiveness training, programs for juvenile delinquents, smoking cessation programs, pain management interventions, computer-based instruction, mastery learning). The search and selection procedure yielded 332 reports of 319 distinct meta-analyses: 181 for educational interventions, 123 for mental health interventions, and 15 for industrial/organizational interventions of a psychosocial nature. Lipsey and Wilson (1993) briefly described the intervention area covered by each of these meta-analyses, and a bibliography of those meta-analyses not listed in Lipsey and Wilson (1993) is provided in the reference list.

Coding of Meta-Analysis Reports

Two types of data were extracted from each meta-analysis. First, the total effect size variance around the grand mean effect size was coded. When a total variance or standard deviation was not reported directly, variance was estimated from other reported statistics (e.g., the standard error). Second, information related to the effect size variance associated with selected study features was coded when available. Most of the meta-analyses reported breakouts of the mean effect sizes for such variables as type of research design (e.g., random and nonrandom), type of treatment (e.g., behavioral self-management, cognitive–behavioral, and biofeedback/relaxation therapy), different outcomes (e.g., depression and anxiety), and different samples (e.g., males and females).

The breakout dimensions were divided into those involving the intervention (e.g., treatment types, respondent characteristics, outcome constructs), referred to as substantive features, and those involving the methods and procedures used to study the intervention (e.g., design types, method quality, operationalization of dependent variables), referred to as method features. The specific categories into which breakouts of substantive and method features were sorted are described in the Results section. Effect size data associated with each breakout of interest were coded, including the mean effect sizes, number of effect sizes, number of studies, standard deviation (or variance or standard error), number of respondents, and any correlations between a breakout variable and effect size.

Estimating Variance Components

To represent the proportion of effect size variance associated with the study features of interest, eta-squared was computed for each breakout on a relevant study feature in each meta-analysis (Winer, Brown, & Michels, 1991, pp. 123–126). For the respective breakdown groups, eta-squared is the ratio of the between-groups sum of squares to the total sum of squares (Hays, 1988). The between-groups sum of squares was estimated as the weighted sum of the squared deviations of the mean effect size for each category of the breakout from the grand mean effect size, as follows, where j is the number of categories, k_j is the number of effect sizes per category, SS is the sum of squares, and ES is the effect size:

$$SS_{\text{between}} = \Sigma k_j(\overline{ES}_j - \overline{ES})^2.$$

Estimation of the total sum of squares depended on the data available. When the variance or standard deviation around each mean effect in a breakout was reported, the total sum of squares was the sum of the SS between-groups and the SS within groups, with the latter computed as

$$SS_{\text{within}} = \Sigma k_j v_j,$$

where j is the number of categories, k_j is the number of effect sizes per category, and v_j is the variance of the effect sizes within each category. When variances for each breakout category were not provided, the total sum of squares was estimated from the total variance multiplied by the total number of effect sizes used in the breakout, which may not always have exactly equaled the k on which the original variance estimate was computed.

The resulting eta-squared values provide an estimate of the proportion of the total variance in observed effect sizes associated with a study feature in each meta-analysis, as appropriate to the signal-to-noise framework for this investigation, but they do not

carry information about the direction of the relationship involved. In many instances, it is informative to also know which category of a breakout variable is associated with larger or smaller effect sizes (e.g., whether the effect size variance associated with random vs. nonrandom assignment to conditions represents a tendency for randomized studies to yield larger or smaller effect sizes than nonrandomized ones). To describe further the direction and pattern of the relationship with study features for which this was meaningful, two additional indices were computed.

The simplest of these two indices was the arithmetic difference between the mean effect sizes for two-way comparisons of interest. For instance, we subtracted the mean effect size for nonrandomized comparison studies from the mean for randomized studies within each meta-analysis reporting this breakout. Assuming that randomized designs yield less bias, a positive difference (other things equal) indicates that nonrandom designs underestimate treatment effects and a negative difference indicates overestimation. This index not only describes the direction of the relationship between the study feature and effect size but indicates the magnitude of the associated difference directly.

The other directional index computed, when applicable, was the product–moment correlation coefficient (r). The correlation provides information in familiar form on both the direction and strength of the relationships by representing the linear relationship between ordered categories of a breakout variable and effect size. Because r^2 is the portion of the total variance of the dependent variable predictable from the least squares regression line, r can be defined as the square root of the ratio of the linear component of the between sum of squares to the total sum of squares (McNemar, 1966). The sum of squares for the linear component was computed using standard analysis of variance methods (e.g., Ferguson, 1966, pp. 343–344).

Statistical Analysis

The major forms of analysis for this project were description and comparison of the mean eta-squared values and, when appropriate, the mean difference and r indices for different categories of breakout variables. For such analyses, the values included should be statistically independent and weighted to reflect the precision with which they were estimated. In a typical meta-analysis, this is accomplished by using only a single effect size from each study in any given analysis and by weighting each by the inverse of its sampling variance (Hedges & Olkin, 1985). In the current instance, this approach was modified to accommodate the meta-analysis as the unit of analysis.

Independence of indices. Two sources of dependencies among the index values required attention. First, multiple values relating to a category of breakout variables often were generated from a single meta-analysis. These dependencies were handled by averaging the values within a meta-analysis related to the same category or breakout variable so that each meta-analysis contributed only a single value to a given analysis. This was done separately for each analysis because both broad and narrow groupings of breakout variables were examined.

A second source of dependencies was overlap in studies included in related meta-analyses, such as two meta-analyses on cognitive–behavioral therapy or several meta-analyses on computer-aided instruction. The degree of statistical dependency in these cases is a function of the proportion of studies common to any two meta-analyses. This source of dependency was addressed by selectively eliminating one of any pair of meta-analyses with 25% or more studies in common. When the bibliography of studies included in a meta-analysis was unavailable, a judgment was made about the likely degree of overlap based on the topic. When two or more meta-analyses overlapped, the one based on the largest number of studies was selected except when exclusion of that one allowed for inclusion of smaller meta-analyses with a greater combined size. Few pairs of meta-analyses with any overlap were included in any analysis, and of those, the amount of overlap was generally less than 10%.

The weighted-bootstrap mean and confidence interval. The central tendencies of the eta-squared values across the meta-analyses, and those for the other indices, were computed as bootstrap means weighted by the harmonic means of the number of effect sizes contributing to each level of the breakout. The bootstrap resampling approach (Efron, 1982; Lunneborg, 1985; Mooney & Duval, 1993; Stine, 1990) was selected because it provided a method for estimating confidence intervals around the mean for each index without requiring assumptions about its underlying distributional properties.

Results

The analysis focused on the portion of variability in effect sizes between studies within meta-analyses associated with various study features. An important initial question was, How much variability is there to

be explained? If the effect sizes for the studies in a particular treatment domain show little variation, indicating substantial agreement on the outcome, then there is little variation to be explained by study features, methodological or substantive. For the 250 meta-analyses that provided pertinent data, the average variance across effect sizes within a meta-analysis was .52, which translates to a standard deviation of .72. J. Cohen's (1988) well-known guidelines identify .20 as a "small" effect size and .80 as a "large" one. Moreover, the grand mean effect size in this collection of meta-analyses has a standard deviation of only .29 across all the diverse interventions represented in it. Relative to these ranges, it is clear that the effect size variability within the typical meta-analysis in this set is quite substantial.

A relatively small proportion of the total variance associated with a study feature can represent meaningful differences among the associated effect sizes. For example, 4% of a total variance of .50 (i.e., an η^2 of .04) associated with a two-category breakout on a study feature (e.g., random vs. nonrandom assignment) represents a difference of .28 between the subgroup effect size means. Relative to the grand mean effect size for the meta-analysis in this example (.50), a .28 difference between the means for two subgroups of studies is substantial.

As mentioned earlier, the effect size breakouts of interest in relation to the effect size variance were those in one of four broad categories representing the features of the treatment, respondents, measurement, and design. Each of these is discussed in turn and then summarized in an overall model for observed treatment effect sizes. An assumption of these analyses is that there is little covariation in the eta-squared between distinct breakout categories across meta-analyses. Study features are unlikely, however, to be truly independent. Of greatest concern is disproportionate covariation between breakout pairs because the driving research question is the relative effect of these breakout categories on effect size variability.

Treatment Features

Treatment features were differentiated into three subcategories: treatment types, treatment components (i.e., elements of treatments such as relaxation or empathetic reflection), and treatment dosage (i.e., intensity or duration). The categorization of pertinent breakouts as representing treatment types or treatment components distinguished general treatment approaches or protocols from their constituent elements or techniques. The heuristic used in making this distinction was that a treatment type was a relatively free-standing intervention, whereas a treatment component could be added to or subtracted from a treatment but would not generally stand alone as a complete intervention.

The mean eta-squared values for effect size breakouts within each of these categories are shown in Table 1. Different treatment types were associated with the largest proportion of effect size variability, with treatment components and treatment intensity or duration, in turn, associated with roughly half as much. This indicates that, within many of the treatment domains represented, as would be expected, different treatment configurations show differential effects. For all the treatment characteristics, however, the range in the mean eta-squared values across meta-analyses was large, thus indicating greater differentials in some intervention areas than in others.

The mean eta-squared value for differences in treatment intensity or duration (usually defined as the number of weeks of treatment) indicates that this feature is associated with roughly 5% of the effect size variance. Recall that eta-squared is a nondirectional and nonlinear index. Different patterns of mean effects across categories of a study feature could each produce an equivalent eta-squared. Study features such as treatment intensity have an ordinal nature, and as such it is meaningful also to assess whether there is a linear relationship between the study feature and effect size. This was assessed by the linear correlations between treatment dosage and effect size (Table 2). The mean correlation was slightly negative, although not significantly different from zero, and has a very wide range across meta-analyses. Examination of the meta-analyses yielding the largest negative and positive correlations did not reveal any obvious characteristics of the respective treatment domains that would explain this finding. It seems likely, however, that dose is confounded with other study features that offset its expected relationship to effect size. One candidate is the diagnostic severity of the respondent population: A study involving a seriously impaired client group may have longer average treatment duration and poorer outcomes.

Respondent Features

The breakouts of respondent groups commonly reported in the meta-analyses were divided into those reflecting age, gender, ethnicity, socioeconomic status, diagnosis (psychological meta-analyses), and

Table 1
Mean Eta-Squared Values for Selected Study Features

Study feature	M^a	95% CI[b]	Range	Median	N^c
Treatment features					
Treatment type	.08	.06–.10	.00–.50	.08	116
Treatment component	.04	.03–.05	.00–.43	.03	64
Intensity or duration	.05	.03–.07	.00–.49	.05	81
Respondent features					
Age	.04	.02–.06	.00–.10	.03	90
Gender	.02	.01–.03	.00–.73	.01	34
Ethnicity	—	—	.01–.25	.09	7
Socioeconomic status	.05	.01–.09	.00–.14	.06	15
Diagnosis	.06	.03–.09	.00–.30	.04	26
Ability group	.05	.03–.07	.00–.63	.01	38
Measurement features					
Construct	.07	.05–.09	.00–.87	.06	107
Operationalization	.08	.02–.14	.00–.29	.05	11
Source of information	.05	.03–.07	.00–.21	.04	13
Researcher-developed measure	.02	.01–.03	.00–.48	.02	27
Design features					
Comparison group type	.05	.02–.08	.00–.52	.04	33
No treatment vs. placebo	.04	.02–.06	.00–.75	.03	23
No treatment vs. alternative treatment	.12	.00–.24	.00–.50	.14	16
Design type	.04	.02–.06	.00–.62	.02	93
Random vs. nonrandom	.02	.01–.03	.00–.59	.01	76
Comparison vs. pre–post	.06	.00–.12	.00–.62	.02	41
Methodological quality	.03	.02–.04	.00–.23	.02	65
Sample size	.04	.03–.05	.00–.68	.04	69

Note. Dashes indicate insufficient sample size for bootstrap procedure.
[a] Bootstrap mean, weighted by the harmonic mean of the number of studies contributing effect sizes to each level of the breakout.
[b] Confidence interval based on standard deviation of bootstrap distribution.
[c] Number of independent meta-analyses contributing to each mean eta-squared.

ability (educational meta-analyses). The mean eta-squared for these various categories of respondent breakouts ranged from .02 for gender to .06 for diagnosis (see Table 1). A large eta-squared median value for ethnicity (.09) was also observed but must be interpreted with caution given the small number of meta-analyses on which it is based (7; insufficient to compute a bootstrap mean).

It was possible to estimate the mean linear correlation for the breakouts of age and gender with effect size (see Table 2). Neither showed any clear directional relationship with effect size, and the means were not significantly different from zero. This finding indicates that, across the treatment domains examined, the mean effect sizes for males and females and for older and younger respondents were roughly comparable on average, although they varied widely across intervention areas. The majority of the gender and age breakouts were from educational meta-analyses. The relationship of respondent features to program effects is likely to be domain specific; as such, the prior finding has limited generalizability. In a research domain in which respondent features are related to program effects, the failure to take into account the relevant respondent characteristics in the design would reduce the effect size and statistical power (Lipsey, 1990).

Outcome Constructs and Measurement Features

Outcome Constructs

The importance of the various dependent variable constructs that represent the expected outcomes of an intervention is reflected in the number of meta-analyses that reported breakouts of effect size by outcome construct. Different outcome constructs were associated with roughly 7%, on average, of the variance in effect sizes, almost the same amount as for treatment type (see Table 1). Within an intervention

Table 2
Mean Linear Correlation Between Effect Size and Study Features That Break Out Into Ordered Categories

Study feature	M^a	Range	N^b
Treatment feature			
Intensity or duration	−.02	−.44–.56	81
Respondent features			
Age	−.02	−.65–.35	90
Gender[c]	.00	−.30–.32	34
Measurement feature			
Researcher-developed measure	.10*	−.69–.36	27
Design features			
Comparison group type			
No treatment vs. placebo	.06	−.87–.66	23
No treatment vs. alternative treatment	.18*	−.53–.71	16
Design type			
Random vs. nonrandom	.04*	−.60–.77	76
Comparison vs. pre–post	−.08	−.78–.25	41
Methodological quality	−.06	−.48–.39	65
Sample size	−.08	−.59–.44	69

[a] Bootstrap mean, weighted by the harmonic mean of the number of studies contributing effect sizes to each level of the breakout.
[b] Number of independent meta-analyses contributing to each mean eta-squared.
[c] A positive correlation indicates that larger average effects were observed for males.
* $p < .05$, based on a confidence interval derived from the standard deviation of a bootstrap distribution.

domain, therefore, effects on some of the outcome constructs measured were typically much larger than others. It is not, of course, especially surprising that treatments would have larger effects on some outcome variables than others. However, it is interesting that the amount of differentiation is so great given that researchers presumably measure all these outcomes in expectation of potential effects.

Measurement Operationalization

How an outcome construct is measured may matter as much as what is measured. To examine this possibility, the effect size breakouts for different measurement operationalizations were examined. Unfortunately, only 11 independent meta-analyses reported such breakouts, ranging from narrow to broad differences in how the constructs were operationalized. An example from the narrow end of the continuum is a breakout of the different versions of achievement tests used in studies of programs for teaching biology as inquiry (El-Nemr, 1980). At the broader end of the continuum is a breakout of different indices of recidivism (e.g., official arrest, self-report) in studies of delinquency interventions (Kaufman, 1985).

As shown in Table 1, these breakouts were associated with about the same proportion of variance in outcomes as differences in the constructs measured. The small number of meta-analyses contributing to this analysis and the correspondingly large confidence interval for the bootstrap-weighted mean limit any conclusion that can be drawn regarding this matter. However, this indication that different operationalizations of what is presumed to be the same outcome construct within the same treatment domain can lead to quite different results is disconcerting. This finding may be due in part to differential measurement reliability and validity. Hunter and Schmidt (1990) clearly showed the degradation in effect size attributable to measurement unreliability and invalidity.

Source of Information

Related to how an outcome construct is operationalized is the source of the information for the measure, independent of the construct. Several meta-analyses, mostly in mental health, grouped measures by who provided the information, such as self-report, therapist observation, or physiological measurement. These breakouts did not control for the construct measured and, as such, may be confounded. To the degree present, such a confound would inflate the eta-squared. As shown in Table 1, the source of the information accounts for slightly less variability in

effect size than either the construct or the operationalization of the construct. Given the smaller average magnitude of the eta-squared and its smaller variability relative to that for constructs and operationalizations, it appears that the "who" may be less important than the "what" and "how" of measurement.

Origin of Measure

A final measurement feature examined in many meta-analyses was the origin of the outcome measure, that is, whether it was developed by the researcher or was a preexisting standardized or published instrument. These breakouts were found almost exclusively in meta-analyses of educational interventions and, on average, accounted for slightly less than half as much effect size variance as that associated with different constructs or measurement operationalizations (see Table 1). Because these breakouts involved only two categories (researcher developed vs. standardized or published), it was also possible to examine the direction of the effect (see Table 2). Researcher-developed measures generally yielded higher effect sizes within a given treatment domain than standardized or published measures. The direction of this effect was as anticipated, favoring tests developed specifically for the research study. Such measures may be more likely to tap the relevant aspects of the construct being changed by the intervention than a published measure that is not necessarily well adapted to the circumstances of a particular intervention.

The range of the mean effect size difference between researcher-developed versus standardized or published measures was quite large (Table 3), suggesting that the nature of this relationship is very different in different treatment domains. Closer inspection, however, revealed that only 4 of the 27 mean differences were negative, with one outlier of −1.3. The next largest negative value was much less extreme (−.36). Thus, the balance of evidence suggests that researcher-developed measures yield larger effects.

Design Features

Effect size breakouts related to study design most often described one of four different study features: type of comparison group, design type, sample size, and methodological quality. Each of these accounted for an average of roughly 2% to 5% of the effect size variance (see Table 1).

Type of Comparison Group

One set of breakouts contrasted the mean effect size for studies comparing treatment versus a no-treatment control group with that for studies comparing treatment versus a placebo control group. Another set contrasted the mean effect size for studies comparing treatment versus an alternative treatment. The eta-squared value (see Table 1) showed that, overall, the proportion of effect size variance associated with type of comparison group averaged about 5%; much more variance was associated with treatment-alternative treatment comparisons in the cases in which this

Table 3
Difference Between Mean Effect Sizes for Study Features That Break Out Into Two Categories

Study feature	M[a]	Range	N[b]
Measurement feature			
Researcher-developed measure	.13*	−1.3–0.8	27
Design features			
Comparison group type			
No treatment vs. placebo	.13*	−1.0–1.6	23
No treatment vs. alternative treatment	.26*	−1.0–1.6	18
Design type			
Random vs. nonrandom	.03	−1.1–0.8	80
Comparison vs. pre–post	−.13*	−1.6–0.5	47
Methodological quality	−.06	−.70–.64	41
Sample size	−.18*	−1.0–0.7	65

[a] Bootstrap mean, weighted by the harmonic mean of the number of studies contributing effect sizes to each level of the breakout.
[b] Number of independent meta-analyses contributing an eta-squared.
* $p < .05$, based on a confidence interval derived from the standard deviation of a bootstrap distribution.

breakdown was reported. The direction and magnitude of the relationships, as indexed by the average correlations and mean effect size differences for these breakouts (Tables 2 and 3), indicate, as expected, that studies contrasting treatment with no treatment yielded higher effect sizes than those that used either a placebo or alternative treatment as the control condition.

Design Type

Breakouts were examined for randomized versus nonrandomized assignment to experimental groups and for comparison group versus one-group pre–post designs. Comparison group designs were either randomized or nonrandomized and are distinguished from one-group pre–post designs in that the latter do not have a control condition. The mean eta-squared for comparison group designs (randomized and nonrandomized) versus the one-group pre–post design was three times that for the randomized versus nonrandomized designs (.06 vs. .02; see Table 1). Although the confidence intervals for these two estimates overlapped slightly, the finding suggests that the effect size estimates produced by randomized and nonrandomized comparison group designs are more similar to each other within an intervention area than estimates from either compared with those of one-group pre–post designs.

The mean correlation between randomized (coded 1) versus nonrandomized (coded 0) design type and effect size was .04 (Table 2), showing that randomized designs tended to yield slightly higher effect sizes. The magnitude of this effect can be seen in the overall mean effect size difference of .03 (Table 3), which was not significantly different from zero. It does not appear, therefore, that nonrandom comparison group type designs are greatly biased on average relative to randomized designs. However, it is important to recognize that the contrast here is between randomized and nonrandomized designs as they occur in typical intervention research. In practice, randomized designs often fall short of the ideal because of differential attrition, contamination of the control group, and other validity threats that degrade the initial randomization. The contrast between studies that were initially randomized versus studies that were initially nonrandomized, therefore, may not represent a large difference in the internal validity actually obtained at the conclusion of the studies. In addition, examination of the ranges for the correlations and the mean effect size differences (see Tables 2 and 3) shows that there was often substantial bias associated with nonrandomized designs within specific treatment domains. Thus, nonrandomized designs may yield quite different observed effects relative to randomized designs, but the difference is almost as likely to represent an upward as a downward bias.

The bias of one-group pre–post designs relative to comparison group designs was examined by combining the effect sizes from the randomized and nonrandomized designs, when reported separately, into a single category and contrasting it with the mean effect size for the one-group pre–post designs. The correlation between these two categories and effect size was −.08 (see Table 2), and the mean magnitude of the effect size difference was about −.13 (see Table 3). The one-group pre–post design, therefore, generally overestimates treatment effects relative to comparison designs, and, in some treatment domains, the bias is quite large.

Methodological Quality

Breakouts on the quality of study methods as rated by the meta-analysts in their coding were also examined, but these focused heavily on internal validity and, hence, overlapped the issue of type of design. Features that were coded in those ratings included type of assignment (e.g., P. A. Cohen, 1980), degree of differential attrition (e.g., Samson, Borger, Weinstein, & Walberg, 1984), and equivalence of groups at pretest (Sweitzer & Anderson, 1983). The effect size breakouts by method quality accounted for roughly 2% of the variability in study outcome (see Table 1). The direction and magnitude of bias associated with poorer quality studies represented in the correlation between method quality ratings and effect size (see Table 2) was slightly negative but nonsignificant; higher quality studies tended to have smaller effect sizes. The mean effect size difference between the high- and low-quality categories (see Table 3) showed the same pattern, also nonsignificant. This null finding is counter to the general belief that low method quality leads to biased results. It appears that low method quality functions more as error than as bias, reducing the confidence that can be placed in the findings but neither consistently over- nor underestimating program effects.

Sample Size

Beyond the small sample bias for effect size statistics demonstrated by Hedges (1981), we would not expect studies that varied in sample size, other things

equal, to produce different effect sizes, only a difference in the precision with which those effect sizes were estimated. However, Table 1 shows that sample size was associated with about 4% of the effect size variability across studies within a treatment domain. Furthermore, the correlation and mean effect size difference (Tables 2 and 3) indicated that larger samples tended to yield smaller effects. The small sample bias demonstrated by Hedges cannot account for this difference; it is negligible in sample sizes above 20, and the great majority of studies in the sample size breakouts used larger samples than that.

A plausible explanation is that smaller studies may represent more tightly controlled implementations and evaluations of interventions and thus are more homogeneous with regard to both respondent populations and treatment delivery (Yeaton & Sechrest, 1981). This homogeneity would mean less variability on the dependent variable within a study and possibly stronger effects, with a corresponding increase in observed effect sizes. It is also possible that the differences associated with sample size reflect publication bias. Larger studies have greater power to detect small effects, and statistically significant findings may have greater likelihood of being submitted and accepted for publication. Published studies, in turn, are easier to locate and thus likely to be overrepresented in meta-analyses (Begg, 1994; Kraemer, Gardner, Brooks, & Yesavage, 1998; Lipsey & Wilson, 1993).

A Composite Model of Treatment Effect Estimates

Early in this article, a simple model was proposed in which observed intervention effects were viewed as a function of (a) substantive features of the intervention under study (e.g., treatment type, respondent characteristics), (b) features of the study methods (e.g., research design), and (c) stochastic error, particularly sampling error. The analyses reported here provide rough estimates of the proportions of observed effect size variance contributed by various specific substantive and methodological study features. We turn now to the task of combining that information to generate an order-of-magnitude estimate of the overall proportion of effect size variance associated with the interventions being studied relative to that stemming from other sources.

To accomplish this, we first identified those variance sources from the prior analysis that involved substantial conceptual overlap and selected the one with the broader scope. For instance, under measurement features (see Table 1), "source of information" and "researcher-developed measure" are not likely to be orthogonal study features, and both, in turn, are likely to be related to "operationalization." In this case, we judged measurement operationalization to cover the broadest range of measurement variations and dropped the other two categories from consideration. Similarly, under treatment features, we dropped the "treatment component" category in favor of the more global representation in "treatment type." For respondent features, no single category is more encompassing than the others, so, for this case, we simply averaged the eta-squared values across all of them.

For summary purposes, we assume, rather generously, that there is little covariation among the eta-squared indices for the conceptually distinct categories or, at least, that any covariation is not highly disproportionate across pairs of categories. On that basis, we can construct a rough estimate of the relative proportions of variance associated with substantive and methodological study features by adding together, within these respective groupings, the mean proportion of variance associated with each of the selected study features (Figure 1).

An overall estimate of the proportion of effect size variance attributable to subject-level sampling error was derived from 117 meta-analyses that reported sample size information for the studies included in the analysis. Across these meta-analyses, sampling error accounted for as little as 1% of the variance and as much as 100%. The mean was 26%, with the 25th and 75th percentiles at 7% and 39%, respectively. Thus, within the typical treatment domain represented in this sample, about 25% of the observed variability in effect sizes can be attributed to sampling error associated with the study-level subject samples.

With the different average proportion of variance estimates grouped according to whether they are related to substantive features of the intervention, methods, or sampling error, more than 70% of the effect size variance was represented in the composite model. The remainder constituted a residual category and was included as such for completeness. The resulting composite model is presented in Figure 1. As an initial estimate of the relative influence of various study features on observed outcomes, we believe this model has heuristic value, but, of course, it is necessarily approximate. The inclusion of additional study features and better estimation of the statistically indepen-

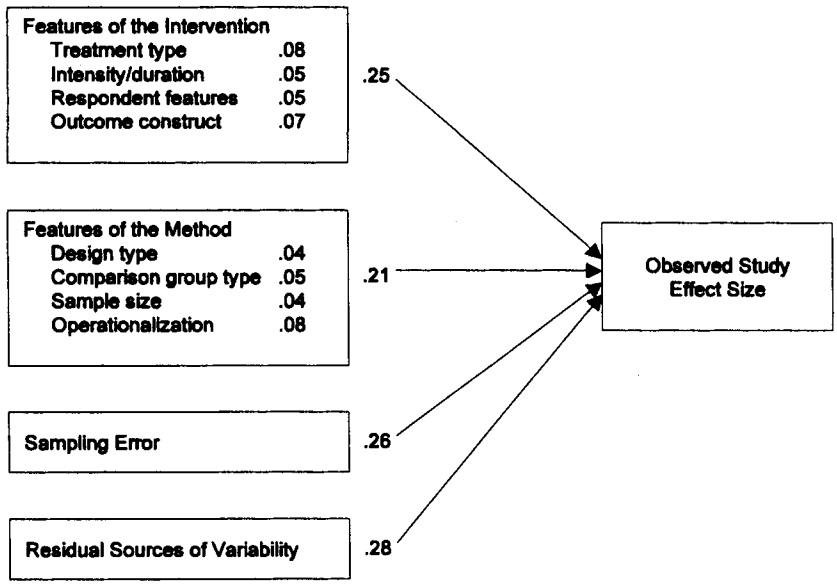

Figure 1. Proportion of variance in observed study effect size explained by selected study features.

dent contribution of each feature to the variance in observed effects would almost certainly change the model to some unknown extent.

Some credibility is lent to the estimates in Figure 1 by virtue of their surprising similarity to the results from a very thorough partitioning of the effect size variance within a single meta-analysis of more than 400 studies of intervention with juvenile delinquents (Lipsey, 1992, 1997). Using multiple regression analysis to estimate the independent contribution of different groups of study features, Lipsey found that characteristics of the treatments, such as treatment type, dosage, and client type, accounted for only 22% of effect size variability. About 25% of the variance was associated with method features (e.g., group equivalence, control type, attrition, and measurement characteristics), and sampling error accounted for 27%. These results have less generality than those shown in Figure 1, because they are based on 1 meta-analysis rather than 319. On the other hand, they provide a more complete accounting of study features and better estimates of their independent contributions to effect size variance. The relatively close agreement between these two attempts to partition across-study effect size variance provides some assurance that the model presented in Figure 1 provides a reasonable, albeit approximate, summary of the relative influence of different groups of study features on the observed outcomes.

Perhaps the most striking aspect of the summary constructed in Figure 1 is the rough parity between the proportion of effect size variance associated with substantive features of the intervention and that associated with features of the study methods. This suggests that methodological choices made by the researcher have nearly as much influence on observed effect sizes as the features of the intervention phenomena under study. Moreover, although the residual category is unlikely to entirely represent variance associated with undocumented substantive study features, even if a large portion of it did, the total would still be surprisingly modest; only about half the variation in observed outcome would then be associated with differences in substantive intervention characteristics.

The rather large role of sampling error in observed study effects is also notable. It accounts for roughly 25% of the total variability, which is quite large relative to the portion associated with substantive features of the intervention. This indicates that, for any substantive study feature to influence an observed effect, it must compete not only with distortions associated with study method but with considerable noise from sampling error.

Discussion

This study started with the question, What is the role of method in treatment effect estimates? More specifically, what is the influence of method features

relative to substantive features on observed study outcomes? Ideally, we would have found that the typical range of methods used by researchers was associated with little variation in observed effect sizes relative to that associated with substantive features of the intervention. Although the effect size variability attributable to specific methodological and measurement features was small to modest on average (e.g., eta-squares ranging from .02 to .10), that attributable to substantive features was of similar magnitude. Thus, the effect sizes observed in a typical treatment effectiveness study are in large part a function of method and sampling error. Viewed as a signal-to-noise ratio, the signal is relatively small and the random and nonrandom noise is relatively large.

Implications for Treatment Effectiveness Research

The indications of substantial instability in observed treatment effects found in the analyses presented here have particular importance for the interpretation of the findings from a single study. If the magnitude of the effect observed in the study is mainly a reflection of specific design and measurement choices, plus a substantial random component, different findings between studies are as likely to be the result of method differences or random variation as the result of substantive differences. Thus, a single study will not typically provide a trustworthy indication of the effectiveness of a particular treatment (Schmidt 1992, 1996). Until the stability and generalizability of an effect across acceptable study methods and samples are established, evidence about a treatment effect is weak.

Within the inherent limits of a single study, one important design element is the procedure for assigning subjects to experimental conditions. The benefits of random assignment for ensuring internal validity are well known, and the randomized clinical trial is generally viewed as the ideal design for assessing the effectiveness of an intervention (e.g., Boruch, 1997; Cook & Campbell, 1979). It was, therefore, surprising to find virtually no difference, on average, between the results from nonrandomized comparison group designs and those from randomized designs. Nonetheless, this finding cannot be interpreted as evidence for the equivalence of randomized and nonrandomized designs for providing estimates of treatment effects. A more likely explanation is that within some treatment domains the selection bias in one nonrandomized comparison is offset by an opposite bias in another such comparison. Cook and Leviton (1980) argued that it is plausible that selection bias acts as error rather than bias in many treatment domains, neither consistently over- nor underestimating effect sizes. A balanced distribution of selection biases across studies is by no means assured in any treatment domain, however, and large differences between random and nonrandom comparison group designs were found in some of the meta-analyses examined here. Similar differences were found by Heinsman and Shadish (1996) in four selected treatment domains they analyzed very closely. They found that randomized designs produced larger average effects than nonrandomized designs, although this difference was substantially reduced when they controlled for other study differences.

In a similar spirit, LaLonde and Maynard (1987) and Fraker and Maynard (1987) compared effect estimates from an experimental study of an employment training program with that from a quasi-experimental study. They found that not only did the quasi-experimental study produce different results, but the results varied with the statistical model used. Thus, within the employment training domain, a quasi-experimental design produced results inconsistent with the findings from a randomized study even when using sophisticated statistical methods. On the other hand, in the classic case of the field trials for the Salk polio vaccine, a randomized clinical trial and a quasi-experimental design component found similar positive effects (Francis et al., 1955; Meldrum, 1998). A nonrandomized design will be vulnerable to selection bias, but whether significant bias typically occurs is an empirical question. The meta-analytic evidence currently available on this point is far from conclusive but does suggest that selection bias need not be large relative to the many other influences on the magnitude of a treatment effect estimate.

In any event, it is worth noting that, although the proportion of variance associated with design type is smaller than that associated with most of the substantive features of the intervention (Table 1 and Figure 1), it is not a great deal smaller. Design type, therefore, generally contributes a significant amount of "noise" relative to the "signals" the researcher is attempting to detect. What is perhaps more interesting is that the way outcome measures are operationalized appears to be associated with at least as much variation in observed effects as type of design. Indeed, the estimates in Table 1 and Figure 1 show that the operationalization has a larger relationship with effect

size than either of the two design features examined (design type and comparison group type). The number of meta-analyses reporting effect size breakouts according to how the outcome measure was operationalized, however, was much smaller than the number reporting about design type, so the resulting mean eta-squares estimate has a more limited empirical basis.

Nonetheless, the suggestion that the operationalization of the outcome variable may have as much influence on the study findings as the method of assignment to conditions raises important questions that warrant further investigation. Issues related to the quality and appropriateness of outcome measurement are not extensively discussed in the literature on experimental methods for studying treatment effectiveness. Correspondingly, the selection of the operationalization for the dependent variable is generally not discussed or explained in any depth in reports of treatment research. These practices are consistent with the assumption that this matter is not especially problematic and, hence, need not receive great attention in the design of the research. The findings presented here from those meta-analyses that break out effect sizes for different operationalizations of an outcome variable give a contrary indication; this matter may be quite problematic and could well deserve considerably more attention from researchers and methodologists.

Implications for Meta-Analysis

The findings presented here suggest that outcomes observed in a treatment study are, to a considerable extent, a function of specific features of the study methods and often of rather specific features of the intervention itself. Under these circumstances, meta-analysis is not only a relatively precise and effective way to summarize the findings of a body of treatment research and, in the process, gain the statistical power advantages of the combined sample size of the constituent studies, but it is also the means by which the generalizability and stability of those findings are investigated and the respective influence of method, substance, and stochastic error is disentangled in the assessment of treatment effects (Cook, 1993).

One straightforward implication of this situation is the importance of meta-analysts attending to between-study differences and using appropriate analytic frameworks to assess them (e.g., Hedges & Olkin, 1985; Hunter & Schmidt, 1990). Mean effect size values, or breakouts only for major treatment types and outcome constructs, without examination of the amount and sources of variation in the effect sizes contributing to those means, may be very misleading if interpreted as treatment effects. Moreover, this task must be approached in a sophisticated way using multivariate analysis to help disentangle the relationship of different study features, especially method features, with effect sizes. This is critical given the correlational nature of meta-analytic data. Design features and outcome operationalizations are often related to treatment type, duration, respondent characteristics, and other such substantive features of the intervention. Differences in mean effect size that are observed between, for example, different treatment types may actually result from differences in method that are confounded with those treatment types.

Meta-analysis allows for statistical techniques to be used to control for the influence of study method features so that less confounded estimates can be derived for treatment effects. Such controls do not eliminate the possibility that observed differences are a function of unmeasured nuisance variables, but they do reduce the plausibility that the observed effects are the result of confounds with readily identifiable features of study method. Two examples of meta-analyses that applied such techniques to modeled or adjusted-for-method effects before interpreting substantive differences reinforce the conclusion just presented. Lipsey and Wilson (1998) and Shadish (1992) used different approaches to controlling for method features before interpreting substantive differences between mean effect sizes, and both found substantial method effects.

Limitations of This Study

The principal limitation of this study is that, of necessity, the study features of interest could be represented only in broad categories. For example, the distinction between randomized and nonrandomized comparison group designs includes a broad range of design types (e.g., randomization with matching, nonrandomization with post hoc matching), and any of these may have greater or lesser attrition subsequent to the assignment. A more differentiated coding and reporting of study features by the meta-analysts whose results were examined here would have permitted a fuller and more detailed accounting of variance sources. This, in turn, might have increased the proportion of variance found to be associated with study features and reduced the unexplained residual variance.

Increased detail and precision could have been attained, of course, if we had coded the contributing studies de novo rather than relying on what was coded and reported in meta-analyses of those studies. As a practical matter, such an effort would have narrower coverage than the 16,525 studies included in the 319 meta-analyses we examined. The broad scope of our approach has the advantages of efficiency and generality but at the cost of less detail.

A second limitation of this study, implicit in the nature of the research process, is that it is correlational and, therefore, unable to test directly the causal influence of study features on observed outcomes. The experiment on experiments described early in this article would be required to support such causal inferences. Using the natural variation of study features within a treatment domain as an analog to the experiment on experiments is the only practical approach to the research question we have attempted to address, but it has inherent problems. For instance, features of study methods within a treatment domain may well be confounded with substantive differences between studies and may either inflate or deflate the observed relationship between study features and effect size. Because of this possible confounding, a clean partitioning of the effect size variance was not attainable, even for the broad categories of study features we examined. What we have been able to present, therefore, represents a first approximation to the partitioning of effect size variance. More refined estimates will be possible when the practice of meta-analysis gives greater attention to coding and reporting study features and uses more sophisticated techniques for estimating their independent contributions to effect size.

These limitations must also be addressed through better reporting of research methods at the primary study level. Too often the descriptions of methods reported in treatment studies are vague and thus do not allow for careful description and differentiation in meta-analysis coding. Better reporting would enable meta-analysts to code studies into more tightly defined design categories, providing more useful information on the potential biases of specific design choices within that treatment domain. We believe these efforts are justified by the findings presented here. Within the limitations of the current state of study reporting and meta-analysis coding and analysis, these findings give empirical support to the view that the particulars of study method can have as large an influence on the findings as the particulars of the intervention under study and, moreover, that the latter may have far less influence than generally assumed. Better understanding of the nature and magnitude of these influences is essential to improving our methods for studying the effects of psychological, educational, and behavioral interventions.

Reference

References marked with an asterisk indicate studies included in the meta-analysis that were not listed in the Lipsey and Wilson (1993) study.

*Angert, J. F., & Clark, F. E. (1982, May). *Finding the rose among the thorns: Some thoughts on integrating media research.* Paper presented at the meeting of the Association for Educational Communications and Technology, Dallas, TX. (ERIC Document Reproduction Service No. 223 192)

Begg, C. B. (1994). Publication bias. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 399–409). New York: Russell Sage Foundation.

Boruch, R. F. (1997). *Randomized experiments for planning and evaluation: A practical guide.* Thousand Oaks, CA: Sage.

*Chidester, T. R., & Grigsby, W. C. (1984). A meta-analysis of the goal setting-performance literature. In *Academy of Management Proceedings* (pp. 202–206). Briarcliff Manor, NY: Academy of Management.

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

Cohen, P. A. (1980). Effectiveness of student-rating feedback for improving college instruction: A meta-analysis of findings. *Reseach in Higher Education, 13,* 321–341.

Conrad, K. J. (Ed.). (1994). Critically evaluating the role of experiments. *New Directions for Program Evaluation, 63.*

Cook, T. D. (1993). A theory of the generalization of causal relationships. *New Directions for Program Evaluation, 57,* 39–82.

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

Cook, T. D., Cooper, H., Cordray, D. S., Hartmann, H., Hedges, L. V., Light, R. J., Louis, T. A., & Mosteller, F. (Eds.). (1992). *Meta-analysis for explanation: A case book.* New York: Russell Sage Foundation.

Cook, T. D., & Leviton, L. C. (1980). Reviewing the litera-

ture: A comparison of traditional methods with meta-analysis. *Journal of Personality, 48,* 449–472.

Cook, T. D., & Shadish, W. R. (1994). Social experiments: Some developments over the past fifteen years. *Annual Review of Psychology, 45,* 545–580.

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

Dennis, M. L. (1990). Assessing the validity of randomized field experiments: An example from drug abuse treatment research. *Evaluation Review, 14,* 347–373.

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans.* Philadelphia, PA: Society for Industrial and Applied Mathematics.

El-Nemr, M. A. (1980). A meta-analysis of the outcomes of teaching biology as inquiry (Doctoral dissertation, University of Colorado, 1979). *Dissertation Abstracts International, 40,* 5813A. (University Microfilms No. 80-11274)

Ferguson, G. A. (1966). *Statistical analysis in psychology and education* (2nd ed.). New York: McGraw-Hill.

Fraker, T., & Maynard, R. (1987). The adequacy of comparison group designs for evaluations of employment-related programs. *Journal of Human Resources, 22,* 194–227.

Francis, T., Jr., Korns, R., Voight, R., Boisen, M., Memphill, F., Napier, J., & Tolchinski, A. (1955). An evaluation of the 1954 poliomyclitis vaccine trials: Summary report. *American Journal of Public Health, 45* (Suppl.), 1–50.

Glass, G. V. (1976). Primary, secondary and meta-analysis of research. *Educational Researcher, 5,* 3–8.

Hays, W. L. (1988). *Statistics* (4th ed.). Fort Worth, TX: Holt, Rinehart & Winston.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6,* 107–128.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* Orlando, FL: Academic Press.

Heinsman, D. T., & Shadish, W. R. (1996). Assignment methods in experimentation: When do nonrandomized experiments approximate the answers from randomized experiments? *Psychological Methods, 1,* 154–169.

*Hembree, R. (1985). Model for meta-analysis of research in education with a demonstration in mathematics education. Effects of hand held calculators (Doctoral dissertation, University of Tennessee, 1984). *Dissertation Abstracts International, 45*(10), 3087A. (University Microfilms No. 84-29597)

*Henk, W. A., & Stahl, N. A. (1984, November). *A meta-analysis of the effect of notetaking on learning from lecture.* Paper presented at the meeting of the National Reading Conference, St. Petersburg Beach, FL. (ERIC Document Reproduction Service No. 258 533)

*Horak, W. J. (1985, April). *A meta-analysis of learning science concepts from textual materials.* Paper presented at the meeting of the National Association for Research in Science Teaching, French Lick Springs, IN. (ERIC Document Reproduction Service No. 256 629)

*Horon, P. F., & Lynn, D. D. (1980). Learning hierarchies research. *Evaluation in Education, 4,* 82–83.

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

Kaufman, P. (1985). *Meta-analysis of juvenile delinquency prevention programs.* Unpublished master's thesis, Claremont Graduate School, Claremont, CA.

Kazdin, A. E. (Ed.). (1992). *Methodological issues and strategies in clinical research.* Washington, DC: American Psychological Association.

*Klauer, K. J. (1984). Intentional and incidental learning with instructional texts: A meta-analysis for 1970–1980. *American Educational Research Journal, 21,* 323–339.

Kraemer, H. C., Gardner, C., Brooks, J. O., III., & Yesavage, J. A. (1998). Advantages of excluding underpowered studies in meta-analysis: Inclusionist versus exclusionist viewpoints. *Psychological Methods, 3,* 23–31.

LaLonde, R., & Maynard, R. (1987). How precise are evaluations of employment and training programs: Evidence from a field experiment. *Evaluation Review, 11,* 428–451.

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

Lipsey, M. W. (1992). Juvenile delinquency treatment: A meta-analytic inquiry into the variability of effects. In T. D. Cook, H. Cooper, D. S. Cordray, H. Hartmann, L. V. Hedges, R. J. Light, T. A. Louis, & F. Mosteller (Eds.), *Meta-analysis for explanation: A casebook* (pp. 83–127). New York: Russell Sage Foundation.

Lipsey, M. W. (1997). What can you build with thousands of bricks? Musings on the cumulation of knowledge in program evaluation. *New Directions for Evaluation, 76,* 7–24.

Lipsey, M. W., & Cordray, D. S. (2000). Evaluation methods for social intervention. *Annual Review of Psychology, 51,* 345–375.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of

psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48,* 1181–1209.

Lipsey, M. W., & Wilson, D. B. (1998). Effective intervention for serious juvenile offenders: A synthesis of research. In R. Loeber & D. Farrington (Eds.), *Serious and violent juvenile offenders: Risk factors and successful interventions* (pp. 313–395). Thousand Oaks, CA: Sage.

*Luiten, J., Ames, W., & Ackerson, G. (1980). A meta-analysis of the effects of advance organizers on learning and retention. *American Educational Research Journal, 4,* 211–218.

*Luiten, J. W. (1980). Advance organizers in learning. *Evaluation in Education, 4,* 49–50.

Lunneborg, C. E. (1985). Estimating the correlation coefficient: The bootstrap approach. *Psychological Bulletin, 98,* 209–215.

*Lyday, N. L. (1984). A meta-analysis of the adjunct question literature (Doctoral dissertation, Pennsylvania State University, 1983). *Dissertaton Abstracts International, 45*(1), 129A. (University Microfilms No. 84-09065)

McNemar, Q. (1966). *Psychological statistics* (3rd ed.). New York: Wiley.

Meldrum, M. (1998). "A calculated risk": The Salk polio vaccine fields trials of 1954. *British Medical Journal, 317,* 1233–1236.

*Mento, A. J., Steel, R. P., & Karren, R. J. (1987). A meta-analytic study of the effects of goal setting on task performance: 1966–1984. *Organizational Behavior and Human Decision Processes, 39,* 52–83.

Mooney, C. Z., & Duval, R. D. (1993). *Bootstrapping: A nonparametric approach to statistical inference.* Newbury Park, CA: Sage.

*Moore, D. W., & Readence, J. H. (1984). A quantitative and qualitative review of graphic organizer research. *Journal of Educational Research, 78,* 11–17.

*Ogles, B. M., Lambert, M. J., Weight, D. G., & Payne, I. R. (1990). Agoraphobia outcome measurement: A review and meta-analysis. *Psychological Assessment, 2,* 317–325.

*Parham, J. L. (1983). A meta-anlaysis of the use of manipulative materials and student achievement in elementary school mathematics (Doctoral dissertation, Auburn University, 1983). *Dissertation Abstracts International, 44*(1), 96A. (University Microfilms No. 83-12477)

*Powell, G. (1980, December). *A meta-analysis of the effects of "imposed" and "induced" imagery upon word recall.* Paper presented at the meeting of the National Reading Conference, San Diego, CA. (ERIC Document Reproduction Service No. 199 644)

*Readence, J., & Moore, D. W. (1981). A meta-analytic review of the effect of adjunct pictures on reading comprehension. *Psychology in the Schools, 18,* 218–224.

*Redfield, D. L., & Rousseau, E. W. (1981). A meta-analysis of experimental research on teacher questioning behavior. *Review of Educational Research, 51,* 237–245.

*Roberts, A. R., & Camasso, M. J. (1991). The effect of juvenile offender treatment programs on recidivism: A meta-analysis. *Notre Dame Journal of Law, Ethics & Public Policy, 5,* 421–441.

Rosenthal, R. (1991). *Meta-analytic procedures for social research* (2nd ed.). Newbury Park, CA: Sage.

Samson, G. E., Borger, J. B., Weinstein, T., & Walberg, H. J. (1984). Pre-teaching experiences and attitudes: A quantitative synthesis. *Journal of Research and Development in Education, 17,* 52–56.

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 Method, 1,* 115–129.

Shadish, W. R., Jr. (1992). Do family and martial psychotherapies change what people do? A meta-analysis of behavioral outcomes. In T. D. Cook, H. Cooper, D. S. Cordray, H. Hartmann, L. V. Hedges, R. J. Light, T. A. Louis, & F. Mosteller (Eds.), *Meta-analysis for explanation: A casebook* (pp. 129–208). New York: Russell Sage Foundation.

Shadish, W. R., & Ragsdale, K. (1996). Random versus nonrandom assignment in psychotherapy experiments: Do you get the same answer? *Journal of Consulting and Clinical Psychology, 64,* 1290–1305.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 32,* 752–760.

Stine, R. (1990). An introduction to bootstrap methods: Examples and ideas. In J. Fox & J. S. Long (Eds.), *Modern methods of data analysis* (pp. 325–373). Newbury Park, CA: Sage.

*Stone, C. L. (1983). A meta-analysis of advance organizer studies. *Journal of Experimental Education, 51,* 194–199.

Sweitzer, G. L., & Anderson, R. D. (1983). A meta-analysis of research on science teacher education practices associated with inquiry strategy. *Journal of Research in Science Teaching, 20,* 453–466.

*Tubbs, M. E. (1986). Goal setting: A meta-analytic exami-

nation of the empirical evidence. *Journal of Applied Psychology, 71,* 474–483.

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design* (3rd ed.). New York: McGraw-Hill.

*Wood, R. E., Mento, A. J., & Locke, E. A. (1987). Task complexity as a moderator of goal effects: A meta-analysis. *Journal of Applied Psychology, 72,* 416–425.

Yeaton, W. H., & Sechrest, L. (1981). Critical dimensions in the choice and maintenance of successful treatments: Strength, integrity, and effectiveness. *Journal of Consulting and Clinical Psychology, 49,* 156–167.

Received August 18, 1998
Revision received June 11, 2001
Accepted July 25, 2001 ■

Call for Nominations

The Publications and Communications Board has opened nominations for the editorships of ***Journal of Experimental Psychology: Animal Behavior Processes, Journal of Personality and Social Psychology: Personality Processes and Individual Differences, Journal of Family Psychology, Psychological Assessment,*** and ***Psychology and Aging*** for the years 2004–2009. Mark E. Bouton, PhD, Ed Diener, PhD, Ross D. Parke, PhD, Stephen N. Haynes, PhD, and Leah L. Light, PhD, respectively, are the incumbent editors.

Candidates should be members of APA and should be available to start receiving manuscripts in early 2003 to prepare for issues published in 2004. Please note that the P&C Board encourages participation by members of underrepresented groups in the publication process and would particularly welcome such nominees. Self-nominations are also encouraged.

Search chairs have been appointed as follows:

- Lucia A. Gilbert, PhD, and Linda P. Spear, PhD, for ***JEP: Animal***
- Sara Kiesler, PhD, for ***JPSP: PPID***
- Susan H. McDaniel, PhD, and Mark I. Appelbaum, PhD, for the ***Journal of Family Psychology***
- Lenore W. Harmon, PhD, for ***Psychological Assessment***
- Randi C. Martin, PhD, and Joseph J. Campos, PhD, for ***Psychology and Aging***

To nominate candidates, prepare a statement of one page or less in support of each candidate. Address all nominations to the appropriate search committee at the following address:

Karen Sellman, P&C Board Search Liaison
Room 2004
American Psychological Association
750 First Street, NE
Washington, DC 20002-4242

The first review of nominations will begin December 14, 2001.