Psychological Methods
1996, Vol. 1, No. 2, 154–169

Copyright 1996 by the American Psychological Association, Inc.
1082-989X/96/$3.00

Assignment Methods in Experimentation: When Do Nonrandomized Experiments Approximate Answers From Randomized Experiments?

Donna T. Heinsman and William R. Shadish
University of Memphis

This meta-analysis compares effect size estimates from 51 randomized experiments to those from 47 nonrandomized experiments. These experiments were drawn from published and unpublished studies of Scholastic Aptitude Test coaching, ability grouping of students within classrooms, presurgical education of patients to improve postsurgical outcome, and drug abuse prevention with juveniles. The raw results suggest that the two kinds of experiments yield very different answers. But when studies are equated for crucial features (which is not always possible), nonrandomized experiments can yield a reasonably accurate effect size in comparison with randomized designs. Crucial design features include the activity level of the intervention given the control group, pretest effect size, selection and attrition levels, and the accuracy of the effect-size estimation method. Implications of these results for the conduct of meta-analysis and for the design of good nonrandomized experiments are discussed.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Since Glass (1976) coined the term *meta-analysis,* hundreds of meta-analyses have been done (Cooper & Lemke, 1991; Lipsey & Wilson, 1993). Nearly all have focused on substantive issues such as whether psychotherapy works. But a few meta-analyses have studied methodological issues, such as experimenter expectancy effects (Rosenthal & Rubin, 1978) or publication bias (Simes, 1987). The present study is of this latter type, using meta-analysis to examine the defining feature of the randomized experiment, random assignment to conditions (Fisher, 1925). When certain assumptions are met (e.g., no treatment correlated attrition) and it is properly executed (e.g., assignment is not overridden), random assignment allows unbiased estimates of treatment effects and justifies the theory that leads to tests of significance. We compare this experiment to a closely related quasi-experimental design—the nonequivalent control group design—that is similar to the randomized experiment except that units are not assigned to conditions at random (Cook & Campbell, 1979). Statistical theory is mostly silent about the statistical characteristics (bias, consistency, and efficiency) of this design.

However, meta-analysts have empirically compared the two designs. In meta-analysis, study outcomes are summarized with an effect size statistic (Glass, 1976). In the present case, the standardized mean difference statistic is relevant:

$$d = \frac{M_T - M_C}{SD_p},$$

where M_T is the mean of the experimental group, M_C is the mean of the comparison group, and SD_p is the pooled standard deviation. This statistic

Donna T. Heinsman and William R. Shadish, Center for Applied Psychological Research, Department of Psychology, University of Memphis.

Donna T. Heinsman died January 13, 1996.

This study reports results from the dissertation of Donna T. Heinsman, completed under the supervision of William R. Shadish. We thank Betsy Becker, Beth Devine, Robert Slavin, and Nan Tobler for their generous assistance in helping us obtain copies of studies from their meta-analyses.

Correspondence concerning this article should be addressed to William R. Shadish, Center for Applied Psychological Research, Department of Psychology, University of Memphis, Memphis, Tennessee 38152.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

allows the meta-analyst to combine study outcomes that are in disparate metrics into a single metric for aggregation. Comparisons of effect sizes from randomized and nonrandomized experiments have yielded inconsistent results (e.g., Becker, 1990; Colditz, Miller, & Mosteller, 1988; Hazelrigg, Cooper, & Borduin, 1987; Shapiro & Shapiro, 1983; Smith, Glass & Miller, 1980). A recent summary of such work (Lipsey & Wilson, 1993) aggregated the results of 74 meta-analyses that reported separate standardized mean difference statistics for randomized and nonrandomized studies. Overall, the randomized studies yielded an average standardized mean difference statistic of $d = 0.46$ ($SD = 0.28$), trivially higher than the nonrandomized studies $d = 0.41$ ($SD = 0.36$); that is, the difference was near zero on the average over these 74 meta-analyses. Lipsey and Wilson (1993) concluded that "there is no strong pattern or bias in the direction of the difference made by lower quality methods. In a given treatment area, poor design or low methodological quality may result in a treatment estimate quite discrepant from what a better quality design would yield, but it is almost as likely to be an underestimate as an overestimate" (p. 1193). However, we believe that considerable ambiguity still remains about this methodological issue.

Problems With Past Research

Careful Definition of Key Variables of Interest

Most meta-analyses have examined assignment method differences in studies primarily devoted to substantive questions. As a result, they have often not carefully defined the independent variable (assignment method) and dependent variable (effect size) in the assignment method question. For example, they may report a difference between a set of categories different from randomized and nonrandomized experiments; so Becker (1990) included both nonequivalent control group designs and uncontrolled studies in her nonrandomized group. Other meta-analysts may have coded studies into these categories incorrectly, as did one meta-analyst who confused random sampling with random assignment. Other problems arise with the dependent variable—effect size. For example, when studies report results only as nonsignificant, the meta-analyst may estimate effect size as zero. If such decisions are differentially distributed across random and nonrandomized experiments—as they might be if, for example, randomized experiments are more likely to be published—differences in average effect size might be created or reduced as a result.

What Should Count as a Randomized or Nonrandomized Experiment?

Methods textbooks (e.g., Cook & Campbell, 1979) emphasize that randomized experiments may yield biased estimates if differential attrition occurs; similarly, matching on reliable covariates may aid nonrandomized experiments (Holland, 1986; Rubin, 1974). If so, then comparisons between randomized and nonrandomized experiments should consider both assignment method and these other variables. Hence, while we first examine differences between random and nonrandom assignment, we focus the rest of this article on more practically interpretable comparisons. This reframes the question from "Are there differences between randomized and nonrandomized experiments?" to "Under what conditions do the results of nonrandomized experiments do a better or worse job of approximating the results from randomized experiments?"

Effect Size Variance

Past meta-analyses focused on mean differences between randomized and nonrandomized studies; but Hedges (1983) provided tentative evidence in a small sample that such studies might differ in variability. Hedges (1983) explained that "preexisting differences between groups are not controlled in the quasi experiments. If the studies that did not have random assignment exhibited a distribution of real preexisting differences, then these differences would also be reflected in the distribution of (posttest) effect-size estimates" (p. 393). This issue has not been examined in other meta-analyses, despite its importance for statistical power. We examine it here.

Variables Confounded With Results Over Areas

Even though the average difference between randomized and nonrandomized experiments may be zero (Lipsey & Wilson, 1993), such differences

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

may vary greatly over area, sometimes being positive and sometimes negative. One reason may be that variables such as matching, attrition, and selection may covary with assignment mechanism differently over areas. For instance, in some areas matching may be common in randomized experiments, but it may be more common in nonrandomized experiments in other areas. It is crucial to take such covariates into account before drawing a conclusion about differences between randomized and nonrandomized experiments. We include two kinds of these covariates: (a) those that are intimately tied to experimental design such as differential attrition, matching, the kind of control group used, and the similarity of the comparison group to the treatment group, and (b) other variables that past meta-analyses suggest may strongly influence effect size such as publication status or various measurement characteristics.

Method

Sample

The present study drew from four past meta-analyses that contained both random and nonrandomized experiments on juvenile drug use prevention programs (Tobler, 1986), psychosocial interventions for postsurgery outcomes (Devine, 1992), coaching for Scholastic Aptitude Test performance (Becker, 1990), and ability grouping of pupils in secondary school classes (Slavin, 1990). These four areas were selected deliberately to reflect different kinds of interventions and substantive topics. Although Lipsey and Wilson's (1993) article had not appeared at the time this selection was made, these four meta-analyses reflect the kinds of substantive topics in their larger sample. Specifically, 290 of the 302 meta-analyses in their sample examined health, mental health, or education—exactly the areas examined in the four meta-analyses in the present study (the remaining 12 meta-analyses in their sample examined worksite and organizational interventions). All four meta-analyses also included many unpublished manuscripts, allowing us to examine publication bias effects. In this regard, a practical reason for choosing these four was that previous contacts with three of the four authors of these meta-analyses suggested that they would be willing to provide us with these unpublished documents. These meta-analyses also reported variable results comparing randomized experiments with nonrandomized experiments. Our target was to include about 100 primary studies in the present meta-analysis, more than is included in the vast majority of past meta-analyses (Lipsey & Wilson, 1993) and large enough to allow some modeling of the effects of the variables of interest in this study using regression. From each of these four meta-analyses, we selected randomized and nonrandomized experiments, including dissertations and other unpublished manuscripts, using the following criteria:

1. To maximize effect size, we looked for studies that compared treatments with control conditions rather than with other treatments and that did so at posttest rather than at follow-up. Control conditions included no treatment, wait list, placebo, and treatment as usual. An example of the latter is a study of the effects of presurgical patient education on postsurgical outcome; the comparison condition included everything that is normally done to patients prior to surgery, probably including some education conveyed through the physician or nurse or through the informed consent form.

2. We excluded studies that did not report the statistics required to compute an effect size using a standard formula. We included the several ways of computing a standardized mean difference statistic from mean and standard deviations, from raw data, or from a one-way two-group F test or t test. We sometimes estimated the standardized mean difference statistic from other data such as means and a three-group F test or repeated measures F test, if those formulae seemed likely to approximate effect size well (Smith et al., 1980, Appendix 7). However, we coded these separately for later comparison to the exact methods. We excluded effect sizes reported only as significant or nonsignificant, those codable only from an inexact probability level (e.g., $p < .05$ or $p > .05$), and other estimates that seemed likely to yield quite different answers from the standardized mean difference statistic on posttest means (e.g., estimating effect size from two within-group t tests). We excluded dichotomous outcomes, which should be coded by odds ratios (Shadish & Haddock, 1994), because we could not convert odds ratios to d (Hasselblad & Hedges, 1995, have now published a formula for doing so).

3. We excluded studies in which subject assign-

Table 1
Number of Studies From Four Meta-Analyses

	Random experiments		Nonrandom experiments		
Meta-analysis	Published	Unpublished	Published	Unpublished	Total
Becker (1990): SAT coaching	2	5	3	4	14
Slavin (1990): ability grouping	0	5	4	5	14
Devine (1992): presurgical intervention	18	9	9	5	41
Tobler (1986): drug-use prevention	10	2	13	4	29
Total	30	21	29	18	98

Note. SAT = Scholastic Aptitude Test.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

ment method was not clear or in which the author wrote contradictory things about assignment.

4. We excluded studies using haphazard assignment, such as alternating assignment of subjects to conditions. Such assignment is not formally random but also not obviously biased; and there are too few such studies to examine as a separate category. However, this criterion does not imply that systematic bias will be present in the nonrandomized experiments we did include. It is an empirical question whether other nonrandom selection mechanisms will result in a systematic bias to outcome.

This procedure yielded 98 studies for inclusion, 51 random and 47 nonrandom. These studies allowed computation of 733 effect sizes, which we aggregated to 98 study-level effect sizes. Table 1 describes the number of studies in more detail. Retrieving equal numbers of published and unpublished studies in each cell of Table 1 proved impossible. Selection criteria resulted in elimination of 103 studies, of which 40 did not provide enough statistics to calculate at least one good effect size;[1] 19 reported data only for significant effects but not for nonsignificant ones; 15 did not describe assignment method adequately; 11 reported only dichotomous outcome measures; 9 used haphazard assignment; 5 had no control group; and 4 were eliminated for other reasons (extremely implausible data, no posttest reported, severe unit of analysis problem, or failure to report any empirical results). There is no way to tell how many eliminated studies contributed to random and nonrandom contrasts conducted by past authors. However, given that (a) we eliminated over half the available studies and (b) most meta-analyses report the random and nonrandom contrast for all or nearly all studies in their sample, probably the vast majority of these eliminated studies were used in the past estimates.

Variables Coded

In addition to effect size and substantive area, we coded two sets of variables. The first set (italicized in what follows) are important design variables. First, *assignment method* (random or not) is coded. Second, although the mere presence of a pretest should not affect study outcome, *pretest effect size* may influence study outcome. Third, some studies use *matching* (blocking) or stratifying of subjects at pretest. These terms are not used consistently in the literature (Fleiss, 1986; Keppel, 1991; Kirk, 1982). We used blocking and matching interchangably to refer cases in which the number of units in the block is equal to the number of conditions in the experiment; an example would be rank-ordering subjects on weight at pretest in an experiment with two conditions and then assigning the 2 subjects with the highest weights randomly to condition, and the same with each subsequent pair. With stratifying, the number of units in the strata exceeds the number of conditions, as when subjects are stratified by gender and then randomly assigned to conditions separately from within strata. However, we collapsed these distinctions into one category because of small cell sizes when they were coded separately. Fourth, we coded *total attrition* and *differential attrition* sepa-

[1] One randomized drug-use prevention experiment (Rabinowitz & Zimmerli, 1974) initially included in this meta-analysis was later excluded after it was found to fail this effect-size computation criterion.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

rately for the pair of conditions involved in each effect size. The former is defined as the number dropping out of both conditions divided by the total number assigned to both conditions; the latter computes this separately for each condition, and then differences those rates over conditions. Fifth, we coded *activity in the control group,* with passive controls being no treatment and wait list, and active controls being placebo and treatment as usual; passive controls may yield larger effects. Sixth, we coded whether the *control was internal or external,* the former being drawn from the same population as treatment group subjects (e.g., children from the same school) and the latter from a patently different population (e.g., children in a different city). Randomized controls are always internal, so nonrandomized internal controls might yield results more similar to those from randomized controls. Finally, we coded *self-selection* versus other-selection of subjects to conditions, with randomized experiments always being other selection.

The second category of variables are those that past meta-analyses suggest may predict effect size. First, coding *publication status* allows assessment of publication biases. Second, we coded whether *treatment standardization* occurred; some research indicates standardized treatments may yield larger effects (e.g., Shadish & Sweeney, 1991). Third, we coded *mode of assessment* (self-report or not) for each measure, and its *specificity* (tailored to treatment or not), features that may increase effect size (Smith et al., 1980; Shadish & Sweeney, 1991). Fourth, we coded whether an *exact or inexact effect size* estimate was used. Fifth, we coded *sample size.* Even though we use weighted least squares statistics in which the weight is a function of sample size, sample size itself could still be a significant predictor.

Reliability

We developed a coding manual for all these variables (available on request) and trained to meet reliability criteria on it. Interrater reliability was assessed by having both authors code all information and one effect size from each of 30 studies. Reliability for continuous variables was assessed by a correlation and for categorical variables by kappa. For continuous variables, reliability was excellent, ranging from .98 to 1.00. On the basis of Fleiss's (1981) criteria, 25 categorical variables had excellent reliability (all greater than .76), and 2 had acceptable reliability (.45 and .63).

Analyses

Analyses followed Hedges and Olkin (1985). All standardized mean difference statistics were corrected for small sample bias. Study level effect sizes were weighted by the inverse of their sampling variability in all analyses. In categorical and regression analyses the appropriate test of significance is a Q statistic distributed as chi-square. Variance component estimates were computed as between-studies variation, a function of total variation minus sampling error.

Results

First, we report several simple comparisons of randomized and nonrandomized experiments. Second, we report a regression analysis that adjusts for variables confounded with assignment method. Finally, we project effect size if both randomized and nonrandomized experiments were equally well designed and implemented.

Simple Categorical Analyses

Table 2 shows that over all 98 studies, experiments in which subjects were randomly assigned to conditions yielded significantly larger effect sizes than did experiments in which random assignment did not take place ($Q = 82.09$, $df = 1$, $p < .0001$). Within area, randomized experiments yielded significantly more positive effect sizes for ability grouping ($Q = 4.76$, $df = 1$, $p = .029$) and for drug-use prevention studies ($Q = 15.67$, $df = 1$, $p = .000075$) but not for SAT coaching ($Q = .02$, $df = 1$, $p = .89$) and presurgical intervention studies ($Q = .17$, $df = 1$, $p = .68$). This yielded a borderline interaction between assignment mechanism and substantive area ($Q = 5.93$, $df = 3$, $p = .12$). We include this interaction in subsequent regressions because power to detect interactions is smaller than power to detect main effects and because such an interaction is conceptually the same as Lipsey and Wilson's (1993) finding that assignment method differences may vary considerably over substantive areas. Finally, as Hedges (1983) predicted, the variance component for nonrandomized experiments was twice as large as the

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 2
Differences Between All Randomized and Nonrandomized Experiments

	Randomized experiments			Nonrandomized experiments		
Sample	No. of studies	Average effect size	Variance component	No. of studies	Average effect size	Variance component
Over all studies	51	0.28*	0.06*	47	0.03	0.12*
Area 1: SAT coaching	7	0.36*	0.13*	7	0.37*	0.14*
Area 2: ability grouping	5	−0.02	0.02*	9	−0.23*	0.05*
Area 3: presurgical intervention	27	0.27*	0.05*	14	0.24*	0.04*
Area 4: drug-use prevention	12	0.30*	0.05*	17	0.15*	0.10*

Note. SAT = Scholastic Aptitude Test.
* $p < .05$.

variance component for randomized experiments in the overall sample. Within areas, variance components were equal in two areas but larger for nonrandomized experiments in two others. Hence nonrandom assignment may result in unusually disparate effect size estimates, creating different means and variances.

To judge from these results, one might conclude that nonrandomized experiments yield poor estimates of treatment effectiveness in some areas. But the data in Table 2 are not the best test of this hypothesis. In fact, they may not even be a good test, despite the fact that they are the test most often used in past meta-analyses. The data in Table 2 are confounded with many other variables that influence effect size. For example, many methods textbooks might suggest that a more appropriate comparison is between randomized experiments that had no differential attrition from conditions and nonrandomized experiments that used matching of groups and had no differential attrition. Table 3 reports such a comparison, which did not change results substantially. The problem with Table 3 is that the practice of experimentation typically involves many design features that are often used differently in different substantive areas or are confounded differentially with the use of random assignment across and within areas. Table 4 gives one good example, in which effect sizes for randomized and nonrandomized experiments are broken down within area according to whether an active control (placebo or treatment as usual) or passive control (no treatment or wait list) was used. Note that passive controls yielded larger effects than did active controls, not surprising because giving control participants no active intervention should result in a treatment–control discrepancy larger than that when controls received some active intervention. More important, the use of these controls is differentially confounded with area. For instance, the SAT coaching studies all used passive controls, but the ability grouping studies all used active controls. Even within areas these controls were used differentially. Active controls were used more often in randomized experiments for presurgical interventions but were used more often in nonrandomized experiments for drug-use prevention studies. With all this confounding, it is no wonder Tables 2 and 3 did not yield very clear answers. More variables must be cross-tabulated against assignment method and area to explore such confounds. Problematically, there is a limit to the number of variables that can be cross-tabulated, for the size of the table proliferates beyond practicality and contains empty cells. A more efficient method is regression.

Regression Models

A baseline is the initial regression predicting effect size from assignment method, area, and their interaction. Two of the three effects were significant along with the overall equation itself (R = .66, Q = 322.44, df = 7, p < .00001). The area effect significance test was Q = 159.54 (df = 3, p < .00000), and for assignment was Q = 21.92 (df = 1, p < .00000). The raw regression weight for assignment method was b = .33 (SE = .037). The interaction term—a

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 3
Differences Between Practically Interpretable Randomized and Nonrandomized Experiments

	Randomized experiments with no differential attrition			Nonrandomized experiments that used matching with no differential attrition		
Sample	No. of studies	Average effect size	Variance component	No. of studies	Average effect size	Variance component
Over all studies	25	0.29*	0.09*	7	−0.16	0.52*
Area 1: SAT coaching	1	0.54*	—	3	0.25	0.32*
Area 2: ability grouping	2	−0.23	0.03	3	−0.49*	0.18
Area 3: presurgical intervention	17	0.30*	0.09*	1	1.22	—
Area 4: drug-use prevention	5	0.53*	0.04*	0	—	—

Note. SAT = Scholastic Aptitude Test.
* $p < .05$.

measure of the extent to which the difference between randomized and nonrandomized experiments varies over substantive area—contributed $Q = 5.93$ points to the overall regression chi-square, nonsignificant ($p = .12$) at 3 degrees of freedom. The model specification test was significant ($Q = 421.06$, $df = 90$, $p < .0001$), with a Birge ratio of $R_B = 4.68$, so that these effect sizes have over four times more between-studies variation than might be expected given the within-study sampling error (Hedges, 1994).

We then added the first and second sets of predictor variables, significantly improving model fit to $R = .80$ ($Q = 474.60$, $df = 20$, $p < .00001$); regression coefficients are reported in Table 5. Effect size was higher with low differential and total attrition, with passive controls, with higher pretest effect sizes, when the selection mechanism did not involve self-selection of subjects into treatment, and with exact effect size computation measures. In Table 5, the interaction term and the main effect for assignment were nonsignifi-

Table 4
Differences Between Randomized and Nonrandomized Experiments Taking Activity of Control Group Into Account

	Random experiments			Nonrandom experiments		
Sample	No. of studies	Average effect size	Variance component	No. of studies	Average effect size	Variance component
Area 1: SAT coaching						
Passive controls	7	0.36*	0.13*	7	0.37*	0.14*
Active controls	0			0		
Area 2: ability grouping						
Passive controls	0			0		
Active controls	5	−0.02	0.02*	9	−0.23*	0.05*
Area 3: presurgical intervention						
Passive controls	13	0.29*	0.03*	9	0.50*	0.02*
Active controls	15	0.25*	0.07*	5	0.10	0.00
Area 4: drug-use prevention						
Passive controls	11	0.30*	0.04*	12	0.22*	0.05*
Active controls	1	0.00		5	−0.14*	0.20*

Note. SAT = Scholastic Aptitude Test.
* $p < .05$.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 5
Regression Statistics for Final Equation

Variable	Q statistic[a]	Raw regression coefficient	Standard error
Self- vs. other-report	2.91	0.093	.054
Percentage differential attrition	11.59	−1.056*	.310
Specific vs. general measure	0.80	0.061	.069
Published vs. unpublished	1.68	−0.062	.048
Passive vs. active control group	45.33	0.332*	.049
Exact vs. approximate effect size	13.33	0.235*	.064
Sample size	0.07	0.000[b]	.000[b]
Pretest effect size	46.72	1.140*	.167
Random vs. nonrandom assignment	2.80	0.082	.049
Standardized treatment vs. not	1.92	0.054	.039
Self- vs. other-selection into conditions	6.53	−0.126*	.049
Use of matching-stratifying or not	1.24	0.059	.053
Percentage total attrition	4.02	−0.282*	.141
Internal vs. external control group	0.00	−0.001	.043
Main effect for area	24.60	[c]	[c]
Area × Assignment interaction	3.84	[c]	[c]

[a] Tested against chi-square at 1 degree of freedom except for the main effect for area and for the Area × Assignment interaction, which are tested at 3 degrees of freedom. [b] Regression coefficient was a nonsignificant 2.38×10^{-6}; standard error was 1.68×10^{-5}. [c] No single regression coefficient is available because these two effects are each tested by entering a set of three dummy variables.
* $p < .05$.

cant; the main effect for area remained significant. The model specification test was still rejected (Q = 268.90, df = 77, p < .00001), with a Birge ratio of R_B = 3.49.

Various explorations of outliers, interactions, and effect size transformation yielded four robust regression findings. First, the interaction between area and assignment mechanism was never significant. Second, the main effect for assignment method hovered around the .05 significance level, with an unstandardized regression weight that suggested that random assignment adds between 0.05 and 0.10 to the standardized mean difference statistics that would occur in a nonrandomized experiment. When we take confidence intervals into account, the true effect in Table 5 may range from a high of 0.176 to a low of −0.016. Third, the area main effect was usually significant. Fourth, of the remaining predictors, the ones that were always significant were higher pretest effect size and the use of passive control groups, both of which increased effect size.

Projecting the Results of an Ideal Comparison

Given these findings, one might ask what an ideal comparison between randomized and nonrandomized experiments would yield. We simulate such a comparison in Table 6 using the results in Table 5, projecting effect sizes using predictor

Table 6
Projection of Difference Between Hypothetically Equated Randomized and Nonrandomized Experiments

Sample	Randomized experiment	Nonrandomized experiment
SAT coaching	0.90	1.01
Ability grouping	0.98	0.97
Presurgical education	1.20	1.15
Drug use	0.95	0.86

Note. SAT = Scholastic Aptitude Test.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

values that equate studies at an ideal or a reasonable level. The projections in Table 6 assume that both randomized and nonrandomized experiments used passive control groups, internal control groups, and matching; allowed exact computation of *d*; had no attrition; standardized treatments; were published; had pretest effect sizes of zero; used $N = 1{,}000$ subjects per study; did not allow self-selection of subjects into conditions; and used outcomes based on self-reports and specifically tailored to treatment. Area effects and interaction effects between area and assignment were included in the projection. Note that the overall difference among the eight cell means has diminished dramatically in comparison with Table 2. In Table 2, the lowest cell mean was −0.23 and the highest was 0.37, for a range of 0.60. The range in Table 6 is only half as large (0.34). The same conclusion is true for the range within each area. In Table 2 that range was 0.01 for the smallest difference between randomized and nonrandomized experiments (SAT coaching) to 0.21 for the largest difference (drug-use prevention). In Table 6, the range was 0.11 (SAT coaching), 0.01 (ability grouping), 0.05 (presurgical interventions), and 0.09 (drug-use prevention). Put a bit more simply, nonrandomized experiments are more like randomized experiments if one takes confounds into account.

Discussion

Table 6 suggests that if randomized and nonrandomized experiments were equally well designed and executed, they would yield roughly the same effect size. These results are supportive of the theory of quasi-experimentation (Cook & Campbell, 1979), which until now has emphasized the conceptual logic of quasi-experimental design more than empirical demonstrations that the designs give trustworthy answers. However, we have good reason to think that, in practice, randomized and nonrandomized designs are not equally well designed and executed. Indeed, in many cases it will be very difficult to do so. If so, do the present results have any implications for practice? We think so.

Practice of Field Research

Four of the six significant predictors in Table 5 (pretest effect size levels, total and differential attrition, and self- vs. other-selection) suggest two ways to improve the design and conduct of some quasi-experiments. First, when possible, do not let participants in quasi-experiments select into or out of conditions. The more they self-select, the more biased will be the results. At the start of quasi-experiments, avoid as much self-selection of participants into conditions as possible. When this is not feasible, the focus should be on minimizing self-selection differences between conditions, for example, by selecting a control group from those who applied for treatment too late to be included. Minimizing self-selection will be feasible in other cases when participants do not self-select into treatment (e.g., students rarely self-select into classrooms) or the researcher partly controls final selection into conditions (e.g., selecting matched controls from a presumably similar population). Such designs are preferred. Researchers should also follow standard advice for minimizing attrition (e.g., Capaldi & Patterson, 1987; Cohen et al., 1993). When none of this is feasible, and self-selection and attrition are prominent, we should not expect to obtain a good estimate of effect.

Second, big differences between groups on the outcome variable at pretest will lead to big differences on that variable at posttest. So researchers should minimize pretest differences when possible using techniques such as matching on reliable covariates (Cook & Campbell, 1979) or on propensity scores (Rosenbaum & Rubin, 1985). Researchers should also consider adjusting posttest scores for pretest differences during analysis. Of course, to do these things one must have a pretest on the outcome variable or on some proxy variable correlated with it. So include such pretests when possible.

A fifth significant predictor in Table 5, use of active rather passive control group, does not suggest normative advice. In our data, quasi-experiments with passive controls more closely resembled randomized experiments. But that is just a description that does not lead to a prescription. Unlike attrition or pretest differences, which are strongly suspected to lead to bias on theoretical grounds, there is nothing inherently biased or unbiased about the activity level of a control group. Use of either control is appropriate for both randomized and nonrandomized experiments. The choice should reflect the substantive question of interest.

Practice of Meta-Analysis

This study also has two implications for the practice of meta-analysis. First, ignoring the difference

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

between assignment methods is a bad idea. To judge from Lipsey and Wilson (1993), most meta-analyses ignore this difference. Table 2 shows the potentially harmful effects of doing so; the unadjusted mean effect size estimates can differ over assignment method, as can the variances, so the results may be biased and less powerful. Meta-analysts should compare results from the two designs and not combine effect sizes over those two designs if results are significantly different.

Second, when differences between randomized and nonrandomized experiments occur, the meta-analyst must decide how to proceed. One option is to omit the nonrandomized experiments, especially if few of them are available. However, we prefer that meta-analysts explore the sources of assignment method differences and adjustments for the bias. For example, most meta-analysts do not code pretest effect size estimates. But the unstandardized regression weight for pretest effect size was around 1.0 in the present study, so that whatever differences one has at pretest will be reflected at posttest, plus treatment effects. One might explore several adjustments. The simplest is to subtract pretest effect size from posttest effect size (Wortman, 1992). Or one could include pretest effect size as a covariate in regression. Problematically, many studies do not report pretest data. With the subtraction method, this implies that some estimates will be adjusted for pretest data and some will not. With regression, imputation of missing pretest data might be necessary (Rubin, 1987) to ensure that the correlation matrix is positive definite. Similar explorations could be done for other significant predictors in Table 5 as well. Of course, given the preliminary nature of the present study, we canot be sure which predictors should be included in such adjustments. We can be sure only that an adjustment will often be warranted.

Table 5 lists another significant predictor, that effect sizes were higher when the exact standardized mean difference statistic could be computed than when some approximation was used (see also Ray, 1994). Indeed, this predictor is noteworthy because a criterion for selection of studies into the present research concerned effect-size computation method. The range of methods we used was narrow, which should have reduced the likelihood of finding a computation method effect. Of course, the ideal solution would be for authors of primary studies to report enough statistics to allow computation of *d*. But even if reporting is improved, some cases will have to rely on approximations, if for no other reason than to deal with the backlog of existing reports. Two prominent myths in meta-analysis appear to be that the bulk of approximate methods are captured by Smith et al.'s (1980) appendix of methods and that the accuracy of approximate methods is not much cause for concern. To the contrary, as meta-analysis has proliferated, our experience is that seemingly minor adaptations of standard methods can yield very different results, that many more methods are invented by hard-pressed meta-analytic practitioners than are subject to critical peer scrutiny, and that the statistical properties of these methods are largely unknown. The gap between meta-analytic practitioners and statisticians on this problem is suprising, with the former frequently searching for good information about accurate effect-size estimation methods and with the latter viewing the matter as mundane. The gap needs to be closed.

Caveats to Present Results

Generalization Question. This meta-analysis used the *d* statistic in its analysis, but primary experiments often use more complex analyses that take pretest scores into account, such as covariance or gain score analyses. So we do not know whether our results on *d* statistics, which are generally unadjusted for pretest scores, would generalize to results from these more complex analyses. For example, differences between randomized and nonrandomized experiments found in Table 2 might disappear if one used covariance-adjusted statistics instead of *d*. However, the results in Table 6 indirectly suggest that statistics that take pretests into account should help nonrandomized studies come closer to the answer provided by randomized studies; that is, our results should generalize. But this suggestion is only indirect because Table 6 is based on between-studies adjustments for pretest differences, whereas primary researchers use within-study adjustments for pretest differences across subjects. The distinction is important because work in multilevel models (of which meta-analysis is a special case) suggests that the strength and occasionally direction of a regression coefficient can change when going from the within-study level to the between-studies level (Raudenbush & Willms, 1991).

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Model Specification Tests. The regression reported in Table 5 did not account for all the predictable variance in effect size estimates, so the resulting regression coefficients may not be accurate. Optimistically, numerous explorations of these data did not change the interpretation of key coefficients. Ultimately, the stability of these coefficients with additional predictors is an empirical question. The addition of substantive variables (e.g., type of treatment and subject characteristics) could help account for variance. After all, we usually hope that most variance is due to such substantive variables and that variance due to methodological sources is small. Hence it is sobering that a model consisting entirely of method variables would account for so much variance in the present study, not that it would account for so little. Substantive experts could also help with area-specific methodological codes, such as codes about prototypical selection biases in each area. An example is a recent meta-analysis of neonatal survival among at-risk infants born in or out of a neonatal care unit (Ozminkowski, Wortman, & Roloff, 1988). These studies are all nonrandomized for ethical reasons. The substantive experts knew that birthweight was an important selection variable, and the methodologist knew how selection biases might be manifested differently over studies. Together, they created a selection bias factor for adjusting the results to better approximate what might happen in randomized experiments.

Conclusion

Methodologists often write as if their topic is only theoretical, not subject to empirical study. Statisticians sometimes study methods using Monte Carlo techniques, but they have less direct relationship to actual research practice. More general empirical inquiries are needed. We hope this article helps illustrate how important and productive it can be to use meta-analysis to do so, but we can also use surveys and interviews (e.g., Dennis, 1988), experiments (e.g., Braverman, 1988), secondary analyses of existing data (e.g., Bowering, 1984), and case studies (e.g., Cook & Walberg, 1985). Elsewhere, we labeled such work *the empirical program of methodology* (Shadish & Heinsman, in press). We hope such work can breathe new life into methodology as a specialty distinct from statistics.

References

References marked with an asterisk indicate studies included in the meta-analysis.

*Albert, W., & Simpson, R. (1985). Evaluating an educational program for the prevention of impaired driving among grade 11 students. *Journal of Drug Education, 15,* 57–71.

*Alderman, D. L., & Powers, D. E. (1980). The effects of special preparation on SAT-verbal scores. *American Educational Research Journal, 17* 239–253.

*Archuletta, V., Plummer, O. B., & Hopkins, K. D. (1977). *A demonstration model for patient education: A model for the project "Training Nurses to Improve Patient Education."* Boulder, CO: Western State Commission for Higher Education.

Becker, B. J. (1990). Coaching for the Scholastic Aptitude Test: Further synthesis and appraisal. *Review of Educational Research, 60,* 373–417.

*Best, J. K. (1981). Reducing length of hospital stay and facilitating the recovery process of orthopedic surgical patients through crisis intervention and pastoral care. (Doctoral dissertation, Northwestern University, 1981). *Dissertation Abstracts International, 42,* 3631B.

*Bicak, L. J. (1962). *Achievement in eighth grade science by heterogeneous and homogeneous classes.* Unpublished doctoral dissertation, University of Minnesota.

*Botvin, G., & Eng, A. (1982). The efficacy of a multicomponent approach to the prevention of cigarette smoking. *Preventative Medicine, 11,* 199–211.

Bowering, D. J. (Ed.). (1984). *Secondary analysis of available data bases.* San Francisco: Jossey-Bass.

Braverman, M. T. (1988). Respondent cooperation in telephone surveys: The effects of using volunteer interviews. *Evaluation and Program Planning, 11,* 135–140.

*Breidenstine, A. G. (1936). The education achievement of pupils in differentiated and undifferentiated groups. *Journal of Experimental Education, 5,* 91–135.

*Budd, S., & Brown, W. (1974). Effect of a reorientation technique on postcardiotomy delirium. *Nursing Research, 23,* 341–348.

*Burke, K. B. (1986). *A model reading course and its effects on the verbal scores of eleventh and twelfth grade students on the Nelson Denny Test, the Preliminary Scholastic Aptitude Test, and the Scholastic Aptitude Test.* Unpublished doctoral dissertation, Georgia State University. (University Microfilms No. 86-26152)

Capaldi, D., & Patterson, G. R. (1987). An approach to the problem of recruitment and retention rates for

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

longitudinal research. *Behavioral Assessment, 9,* 169–177.

*Chiotti, J. F. (1961). *A progress comparison of ninth grade students in mathematics from three school districts in the state of Washington with varied methods of grouping.* Unpublished doctoral dissertation, University of Northern Colorado.

Cohen, E. H., Mowbray, C. T., Bybee, D., Yeich, S., Ribisl, K., & Freddolino, P. P. (1993). Tracking and follow-up methods for research on homelessness. *Evaluation Review, 17,* 331–352.

Colditz, G. A., Miller, J. N., & Mosteller, F. (1988). The effect of study design on gain in evaluation of new treatments in medicine and surgery. *Drug Information Journal, 22,* 343–352.

*Collins, N. W., & Moore, R. C. (1970). The effect of a preanesthetic interview on the operative use of thiopental sodium. *Anesthesia and Analgesia, 49,* 872–876.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings.* Chicago: Rand-McNally.

Cook, T. D., & Walberg, H. J. (1985). Methodological and substantive significance. *Journal of School Health, 55,* 340–342.

Cooper, H. M., & Lemke, K. M. (1991). On the role of meta-analysis in personality and social psychology. *Personality and Social Psychology Bulletin, 17,* 245–251.

*Curran, R. G. (1988). *The effectiveness of computerized coaching for the Preliminary Scholastic Aptitude Test (PSAT/NMSQT) and the Scholastic Aptitude Test (SAT).* Unpublished doctoral dissertation, Boston University. (University Microfilms No. 88-14377)

*Davis, H. S. (1973). The role of crisis intervention in the patient's recovery from elective surgery (Doctoral dissertation, Northwestern University, 1973). *Dissertation Abstracts International, 36,* 3490B.

*Delong, R. D. (1971). Individual differences in patterns of anxiety arousal, stress-relevant information and recovery from surgery. (Doctoral dissertation, University of California, Los Angeles, 1970). *Dissertation Abstracts International, 32,* 554B.

Dennis, M. L. (1988). *Implementing random field experiments: An analysis of criminal and civil justice research.* Unpublished doctoral dissertation, Northwestern University, Evanston, IL.

Devine, E. C. (1992). Effects of psychoeducational care with adult surgical patients: A theory probing meta-analysis of intervention studies. In T. D. Cook, H. M. 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. 35–82). New York: Russell Sage Foundation.

*Devine, E. C., O'Conner, F. W., Cook, T. D., Wenk, V. A., & Curtin, T. R. (1988). Clinical and financial effects of psychoeducational care provided by staff nurses to adult patients in the post-DRG environment. *American Journal of Public Health, 78,* 1293–1297.

*Drews, E. M. (1963). *Student abilities, grouping patterns and classroom interaction* (Cooperative Research Project No. 608). Washington, DC: U.S. Department of Health, Education, and Welfare.

*DuPont, P., & Jason, L. (1984). Assertiveness training in a preventive drug education program. *Journal of Drug Education, 14,* 369–378.

*Duryea, E. (1983). Utilizing tenets of innoculation theory to develop and evaluate a preventive alcohol education intervention. *Journal of School Health, 53,* 250–257.

*Dusewicz, R., & Martin, M. (1981). *Impacts of a Georgia drug abuse prevention program: Final evaluation report.* Philadelphia, PA: Research for Better School, Inc. (ERIC Document Reproduction Service No. ED 10569)

*Duthler, T. B. (1979). An investigation of the reduction of psychological stress in patients facing surgical removal of tumors. (Doctoral dissertation, University of Missouri, 1979). *Dissertation Abstracts International, 40,* 4477B.

*Dziurbejko, M. M., & Larkin, J. C. (1978). Including the family in preoperative teaching. *American Journal of Nursing, 79,* 1892–1894.

*Ebel, H., Katz, D., & Rosen, A. (1975). Effect of a marijuana drug-education program: Comparison of faculty-elicited and student-elicited data. *Journal of Drug Education, 5,* 77–85.

*Egbert, L. D., Battit, G. E., Welch, C. E., & Bartlett, M. K. (1964). Reduction of postoperative pain by encouragement and instruction of patients. *New England Journal of Medicine, 270,* 825–827.

*Evans, F. R., & Pike, L. W. (1973). The effects of instruction for three mathematics item formats. *Journal of Educational Measurement, 10,* 257–272.

*Felton, G., Huss, K., Payne, E. A., & Srsic, K. (1976). Preoperative nursing intervention with the patient for surgery: Outcomes of three alternative approaches. *International Journal of Nursing Studies, 13,* 83–96.

*Fick, W. W. (1963). The effectiveness of ability grouping in seventh grade core classes. *Dissertation Abstracts International, 23,* 2753.

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

*Field, P. B. (1974). Effects of tape-recorded hypnotic preparation for surgery. *International Journal of Clinical and Experimental Hypnosis, 22,* 54–61.

Fisher, R. A. (1925). *Statistical methods for research workers.* Edinburgh: Oliver and Boyd.

*Fitzgerald, O. (1978). Crisis intervention with open heart surgery patients. (Doctoral dissertation, American University, 1978). *Dissertation Abstracts International, 39,* 127A.

*Flaherty, G. G., & Fitzpatrick, J. J. (1978). Relaxation technique to increase comfort level of postoperative patients: A preliminary study. *Nursing Research, 27,* 352–355.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions.* New York: Wiley.

Fleiss, J. L. (1986). *The design and analysis of clinical experiments.* New York: Wiley.

*Florell, J. L. (1971). Crisis intervention in orthopedic surgery. (Doctoral dissertation, Northwestern University, 1971). *Dissertation Abstracts International, 32,* 3633B.

*Ford, S. (1974). *Grouping in mathematics: Effects on achievement and learning environment.* Unpublished doctoral dissertation, Yeshiva University.

*Frankel, E. (1960). Effects of growth, practice and coaching on Scholastic Aptitude Test scores. *Personnel and Guidance Journal, 38,* 713–719.

*Gersick, K., Grady, K., & Snow, D. (1988). Social-cognitive skill development with sixth graders and its initial impact on substance use. *Journal of Drug Education, 18,* 55–70.

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

*Hart, R. R. (1980). The influence of a taped hypnotic induction treatment procedure on the recovery of surgical patients. *International Journal of Clininal and Experimental Hypnosis, 28,* 324–332.

Hasselblad, V., & Hedges, L. V. (1995). Meta-analysis of screening and diagnostic tests. *Psychological Bulletin, 117,* 167–178.

*Hayward, J. (1975). *Information—A prescription against pain.* London: Whitefriars Press.

Hazelrigg, M. D., Cooper, H. M., & Borduin, C. M. (1987). Evaluating the effectiveness of family therapies: An integrative review and analysis. *Psychological Bulletin, 101,* 428–442.

Hedges, L. V. (1983). A random effects model for effect sizes. *Psychological Bulletin, 93,* 388–395.

Hedges, L. V. (1994). Fixed effects models. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 285–299). New York: Russell Sage Foundation.

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

*Hegyvary, S. T., & Chamings, P. A. (1975). The hospital setting and patients care outcomes. *Journal of Nursing Administration, 5,* 36–42.

*Hill, B. J. (1982). Sensory information, behavioral instructions and coping with sensory alternation surgery. *Nursing Research, 31,* 17–21.

*Holden-Lund, C. (1988). Effects of relaxation with guided imagery on surgical stress and wound healing. *Research in Nursing and Health, 11,* 235–244.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81,* 945–960.

*Horan, J., & Williams, J. (1982). Longitudinal study of assertion training as a drug abuse prevention strategy. *American Educational Research Journal, 19,* 341–351.

*Hunt, L. (1978). *The effects of preoperative counseling on postoperative recovery—Open heart surgery patients.* Unpublished doctoral dissertation, University of Toronto, Toronto, Ontario, Canada.

*Hurd, P., Johnson, C., Pechacek, T., Bast, P., Jacobs, D., & Luepker, R. (1980). Prevention of cigarette smoking in seventh grade students. *Journal of Behavioral Medicine, 3,* 15–28.

*Jackson, J., & Calsyn, R. (1977). Evaluation of a self development approach to drug education: Some mixed results. *Journal of Drug Education, 7,* 15–28.

*Johnson, C., Graham, J., Hansen, W., Flay, B., McGuigan, K., & Gee, M. (1987). *Project Smart after three years: An assessment of sixth-grade and multiple-grade implementations.* Unpublished manuscript. (Available from C. Anderson Johnson, University of Southern California, Institute for Prevention Research, 35 North Lake Avenue, Suite 200, Pasadena, CA 91101.)

*Johnson, J. E., Juller, S. S., Endress, M. P., & Rice, V. H. (1978). Sensory information, instruction in a coping strategy, and recovery from surgery. *Research in Nursing and Health, 1,* 4–17.

*Kearney, A., & Hines, M. (1980). Evaluation of the effectiveness of a drug prevention education program. *Journal of Drug Education, 10,* 127–134.

*Keefauver, L. W. (1976). *The effects of a program of coaching on Scholastic Aptitude Test scores of high school seniors pretested as juniors.* Unpublished doctoral dissertation, University of Tennessee at Knoxville. (University Microfilms No. 77-3651)

Keppel, G. (1991). *Design and analysis: A researcher's*

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

handbook (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

*Kerckhoff, A. C. (1986). Effects of ability grouping in British secondary schools. *American Sociological Review, 51,* 842–858.

*King, I., & Tarsitano, B. (1982). The effect of structured and unstructured preoperative teaching: A replication. *Nursing Research, 31,* 324–329.

Kirk, R. E. (1982). *Experimental design: Procedures for the behavioral sciences* (2nd ed.). Belmont, CA: Brooks/Cole.

*Kline, R. E. (1964). A longitudinal study of the effectiveness of the track plan in the secondary schools of a metropolitan community (Doctoral dissertation, St. Louis University, 1963). *Dissertation Abstracts International, 25,* 324. (University Microfilms No. 64-4257)

*Laschewer, A. D. (1986). *The effect of computer assisted instruction as a coaching technique for the Scholastic Aptitude Test preparation of high school juniors.* Unpublished doctoral dissertation, Hofstra University.

*Leigh, J. M., Walker, J., & Janaganathan, P. (1977). Effect of preoperative anaesthetic visit on anxiety. *British Journal of Medicine, 2,* 987–989.

*Levesque, L., Grenier, R., Kerouac, S., & Reidy, M. (1984). Evaluation of a presurgical group program given at two different times. *Research in Nursing and Health, 7,* 227–236.

*Lieberman, M., & DeVos, E. (1982). *Adventure-based counseling: Final evaluation report.* Unpublished manuscript. (Available from Marcus Lieberman and Edward DeVos, Harvard University, Graduate School of Education, Cambridge, MA 02138.)

*Lindeman, C. A., & Stetzer, S. L. (1973). Effect of preoperative visits by operating room nurses. *Nursing Research, 22,* 4–16.

*Lindeman, C. A., & Van Aernam, B. (1971). Nursing intervention with the presurgical patient—The effect of structured and unstructured preoperative teaching. *Nursing Research, 20,* 319–332.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48,* 1181–1209.

*Luepker, R., Johnson, C., Murray, D., & Pechacek, T. (1983). Prevention of cigarette smoking: Three-year follow-up of an education program for youth. *Journal of Behavioral Medicine, 6,* 53–62.

*Malfetti, J., Simon, K., & Homer, M. (1977). *Development of a junior high school module in alcohol education and traffic safety.* New York: Safety Research and Education Project, Teachers College, Columbia University.

*Martin, W. B. (1959). Effects of ability grouping on junior high school achievement (Doctoral dissertation, Arizona State College, 1958). *Dissertation Abstracts International, 20,* 2810. (University Microfilms No. 59-1108).

*Martin, W. H. (1927). *The results of homogeneous grouping in the junior high school.* Unpublished doctoral dissertation, Yale University.

*McAlister, A. (1983, May). *Approaches to primary prevention.* Paper presented at National Academy of Sciences National Research Council Conference on Alcohol and Public Policy, Washington, DC.

*Mogan, J., Wells, N., & Robertson, E. (1985). Effects of preoperative teaching on postoperative pain: A replication and expansion. *International Journal of Nursing Studies, 22,* 267–280.

*Myers, E. (1974). The effects of a drug education curriculum based on a casual approach to human behavior. *Journal of Drug Education, 4,* 309–316.

*O'Rourke, T. W. (1973). Assessment of the effectiveness of the New York State Drug Curriculum Guide with respect to drug knowledge. *Journal of Drug Education, 3,* 57–66.

*O'Rourke, T., & Barr, S. (1974). Assessment of the effectiveness of the New York State Drug Curriculum Guide with respect to drug attitudes. *Journal of Drug Education, 4,* 347–356.

*Ortmeyer, J. A. (1978). *Anxiety and repression coping styles, and treatment approaches in the interaction of elective orthopedic surgical stress.* (Doctoral dissertation, Northwestern University, 1977). *Dissertation Abstracts International, 38,* 5536A.

Ozminkowski, R. J., Wortman, P. M., & Roloff, D. W. (1988). Inborn/outborn status and neonatal survival: A meta-analysis of non-randomized studies. *Statistics in Medicine, 7,* 1207–1221.

*Perry, C., Killen, J., Telch, M., Slinkard, L., & Danaher, B. (1980). Modifiying smoking behavior of teenagers: A school based intervention. *American Journal of Public Health, 70,* 722–725.

*Peterson, R. L. (1966). *An experimental study of the effects of ability grouping in grades 7 and 8.* Unpublished doctoral dissertation, University of Minnesota.

Rabinowitz, H., & Zimmerli, W. (1974). Effects of a health education program on junior high students' knowledge, attitudes, and behavior concerning tobacco use. *Journal of School Health, 44,* 324–330.

Raudenbush, S. W., & Willms, J. D. (Eds.). (1991).

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Schools, classrooms, and pupils: International studies of schooling from a multilevel perspective. Orlando, FL: Academic Press.

Ray, J. W. (1994). *An evaluation of the agreement between exact and inexact effect sizes in meta-analysis.* Unpublished doctoral dissertation, University of Memphis.

*Reading, A. E. (1982). The effects of psychological preparation on pain and recovery after minor gynecological surgery: A preliminary report. *Journal of Clinical Psychology, 38,* 504–512.

*Reichbaum, L. S. (1984). *The use of relaxation training to reduce stress and facilitate recovery in open-heart surgery patients.* Unpublished doctoral dissertation, University of Pittsburgh.

*Reynolds, A. J., & Oberman, G. O. (1987, April). *An analysis of a PSAT preparation program for urban gifted students.* Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C.

*Roberts, S. O., & Oppenheim, D. B. (1966). *The effects of special instruction upon test performance of high school students in Tennessee* (CB RDR 66-7, No. 1, and ETS RB 66-36). Princeton, NJ: Educational Testing Service.

*Rock, D. A. (1980). Disentangling coaching effects and differential growth in the FTC coaching study. In S. Messick (Ed.), *The effectiveness of coaching for the SAT: Review and analysis of research from the fifties to the FTC* (Research Report No. 80-8. pp. 123–135). Princeton, NJ: Educational Testing Service.

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *American Statistician, 39,* 33–38.

Rosenthal, R., & Rubin, D. B. (1978). Interpersonal expectancy effects: The first 345 studies. *The Behavioral and Brain Sciences, 3,* 377–386.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology, 66,* 688–701.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: Wiley.

*Sadler, O., & Dillard, N. (1978). A description and evaluation of TRENDS: A substance abuse education program for sixth graders. *Journal of Educational Research, 71,* 171–175.

*Sarvela, P., & McClendon, E. (1987). An impact evaluation of a rural youth drug education program. *Journal of Drug Education, 17,* 213–231.

*Schimitt, F. E., & Woolridge, P. J. (1973). Psychological preparation of surgical patients. *Nursing Research, 22,* 108–116.

*Schinke, S., & Gilchrist, L. (1983). Primary prevention of tobacco smoking. *Journal of School Health, 53,* 416–419.

*Schlegel, R. (1977–1978). The role of persuasive communication in drug dissuasion. *Journal of Drug Education, 7,* 279–290.

*Scovill, N., & Brummer, J. (1988). *Preadmission preparation for total hip replacement patients.* Unpublished manuscript.

Shadish, W. R., & Haddock, C. K. (1994). Combining estimates of effect size. In H. M. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 261–281). New York: Russell Sage Foundation.

Shadish, W. R., & Heinsman, D. T. (in press). Experiments versus quasi-experiments: Do you get the same answer? In W. J. Bukoski (Ed.), *Meta-analysis of drug abuse prevention programs* (NIDA Research Monograph) Washington DC: Superintendent of Documents.

Shadish, W. R., & Sweeney, R. (1991). Mediators and moderators in meta-analysis: There's a reason we don't let dodo birds tell us which psychotherapies should have prizes. *Journal of Consulting and Clinical Psychology, 59,* 883–893.

Shapiro, D. A., & Shapiro, D. (1983). Comparative therapy outcome research: Methodological implications of meta-analysis. *Journal of Consulting and Clinical Psychology, 51,* 42–53.

*Shekleton, M. (1983). *The effect of preoperative instruction in coughing and deep breathing exercises on postoperative ventilatory function.* (Doctoral dissertation, Rush University, 1983). *Dissertation Abstracts International, 45,* 04B.

Simes, R. J. (1987). Confronting publication bias: A cohort design for meta-analysis. *Statistics in Medicine, 6,* 11–29.

Slavin, R. (1990). Achievement effects of ability grouping in secondary schools: A best-evidence synthesis. *Review of Educational Research, 60,* 471–499.

Smith, M. L., Glass, G. V., & Miller, T. I. (1980). *The benefits of psychotherapy.* Baltimore: Johns Hopkins University Press.

*Solomon, A. J. (1973). The effect of psychotherapeutic interview on the physical results of thoracic surgery. (Doctoral dissertation, California School of Professional Psychology, 1973). *Dissertation Abstracts International, 34,* 2319B.

*Stoakes, D. W. (1964). *An educational experiment with*

This document is copyrighted by the American Psychological Association or one of its allied publishers.
This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

the homogeneous grouping mentally advanced and slow learning students in the junior high school, Unpublished doctoral dissertation, University of Colorado.

*Stuart, R. (1974). Teaching facts about drugs: Pushing or preventing. *Journal of Educational Psycology, 66,* 189–201.

*Swisher, J., Warner, R., & Herr, E. (1972). Experimental comparison of four approaches to drug abuse prevention among ninth and eleventh graders. *Journal of Counseling Psychology, 19,* 328–332.

*Thompson, G. W. (1974). The effects of ability grouping upon achievement in eleventh grade American history. *Journal of Experimental Education, 42,* 76–79.

Tobler, N. S. (1986). Meta-analysis of 143 adolescent drug prevention programs: Quantitative outcome results of program participants compared to a control or comparison group. *Journal of Drug Issues, 16,* 537–567.

*Van Steenhouse, A. L. (1978). *A comparison of three types of presurgical psychological intervention with male open heart surgery patients.* (Doctoral dissertation, Michigan State University, 1978). *Dissertation Abstracts International, 39,* 1449A.

*Wallace, L. M. (1984). Psychological preparation as a method of reducing the stress of surgery. *Journal of Human Stress, 10,* 62–76.

*Weisheit, R., Hopkins, R., Kearney, K., & Mauss, A. (1982). Substance abuse, nonconformity, and the inability to assign problem responsibility. *Journal of Drug Issues, 12,* 199–209.

*Weiss, O. F., Weintraub, M., Sriwatanakul, K., & Lasagna, L. (1983). Reduction anxiety and postoperative analgesic requirements by audiovisual instruction. *Lancet, 1,* 43–44.

*Wells, J. K., Howard, G. S., Nowlin, W. F., & Vargas, M. J. (1986). Presurgical anxiety and postsurgical pain and adjustment: Effects of a stress innoculation procedure. *Journal of Consulting and Clinical Psychology, 54,* 831–835.

*Whitla, D. K. (1962). Effect of tutoring on Scholastic Aptitude Test scores. *Personnel and Guidance Journal, 41,* 32–37.

*Wilcutt, R. E. (1969). Ability grouping by content topics in junior high school mathematics. *Journal of Experimental Education, 34,* 20–32.

*Wilson, J. F. (1981). Behavioral preparation for surgery: Benefit or harm? *Journal of Behavioral Medicine, 1,* 79–102.

Wortman, P. M. (1992). Lessons from the meta-analysis of quasi-experiments. In F. B. Bryant, J. Edwards, R. S. Tindale, E. J. Posavac, L. Heath, E. Henderson, & Y. Suarez-Balcazar (Eds.), *Methodological issues in applied social psychology* (pp. 65–81). New York: Plenum.

*Zuman, J. P. (1988, April). *The effectiveness of special preparation for the SAT: An evaluation of a commercial coaching school.* Paper presented at the annual meeting of the American Education Research Association, New Orleans, LA. (ERIC Document Reproduction Service No. ED 294 900)

Received October 30, 1995
Revision received November 3, 1995
Accepted November 28, 1995 ■