Experiments Versus Quasi-Experiments: Do They Yield the Same Answer?

William R. Shadish and Donna T. Heinsman

Life would be ever so much easier if quasi-experiments yielded just as good causal inferences as randomized experiments. Of course, the term "quasi-experiment" covers a multitude of designs. Here it refers to the workhorse of the quasi-experimental design literature: the nonequivalent control group design that includes a treatment group, a control group not receiving treatment, and a posttest for both, but where the assignment of subjects to conditions is not controlled by the researcher, and is certainly not random.

The latter comparison to randomized experiments is generally of most interest. For the assessment of treatment outcome, randomized experiments are widely acknowledged to have many important advantages. Most salient, statistical theory suggests that randomized experiments yield unbiased estimates of treatment effects. For this reason, randomized experiments are usually viewed as the gold standard against which to compare the results of other methods for assessing treatment outcome. If quasi-experiments did as well as randomized experiments, they could often be substituted for randomized experiments, which in many situations would make the logistics of experimentation considerably easier for both researcher and subject.

Unfortunately, relatively few researchers have tried to compare results from randomized experiments to those from quasi-experiments; those who have explored the issue have found inconsistent results. In the medical and surgical literatures, for example, research suggests that randomized trials of medical innovations yield smaller estimates of the effectiveness of the innovation (Colditz et al. 1988; Gilbert et al. 1978). In psychotherapy research, the findings suggest that random assignment may make little difference to outcome (Smith et al. 1980). Becker's (1990) study of Scholastic Aptitude Test (SAT) coaching found that randomized trials yielded larger effect sizes than quasi-experiments. In reality, of course, each of these studies operationalized the question slightly differently. Some included only sequential assignment of subjects to conditions in the quasi-experimental category, others included uncontrolled studies in that same category, and still others lumped random assignment together

with other factors that may affect internal validity. So it is not clear that these studies all addressed the same question.

Moreover, most of these results come from studies aimed at answering substantive questions such as whether psychotherapy works. The methodological question of whether randomized experiments differ from quasi-experiments has usually been of secondary interest, one of many variables that happened to be coded and reported during exploratory analyses. This leads to two general problems. First, few of these past studies have examined the issue in detail. For example, they generally simply report some categorical test of the difference between randomized and quasi-experiments, rarely exploring variables that might moderate the effects of assignment method, such as whether or not studies were published. Second, these past studies have rarely paid careful attention to defining the independent variable (random versus nonrandom assignment) and dependent variable (effect size) as carefully as might be desired to answer this question. For example, these reviews have often included studies where the assignment process was unclear. While this approach is reasonable to get an estimate of the effect of treatment over all studies, it may cloud a comparison between randomized and quasi-experiments if some studies with ambiguous assignment are included in one of these categories.

Given the importance of the question and the paucity of focused research on the question, therefore, the authors have recently begun using meta-analysis to try to explore this issue further. For example, Heinsman (1993) recently finished a dissertation on this topic using 47 quasi-experiments and 52 randomized experiments from four previous meta-analyses that examined, respectively, the effects of SAT coaching (Becker 1990), ability grouping of children in classrooms (Slavin 1990), presurgical psychoeducational interventions (Devine 1992), and drug use prevention (Tobler 1986). This chapter summarizes Heinsman's (1993) most important results, and then reports the results of some additional analyses of that data.

HEINSMAN'S APPROACH

Methodologically, Heinsman sought to remedy certain problems in past comparisons of randomized to quasi-experiments by ensuring that the independent variable (assignment method) and the dependent variable (effect size) were as clearly described and accurately coded as possible given the constraints of meta-analysis. Regarding the independent variable, random versus nonrandom assignment, Heinsman excluded studies that did not have both a treatment and a

control group, that did not clearly describe the assignment process, or that used haphazard assignment. Regarding the dependent variable, effect size, studies were excluded if at least one accurate effect size could not be computed, if it was not clear which numerical direction on a dependent variable constituted a positive outcome, or if statistics were reported for significant findings but not for nonsignificant findings. Finally, Heinsman only coded variables at posttest rather than followup, and excluded studies that reported data only on dichotomous outcomes. The latter are probably best coded with odds ratios, which are not clearly comparable to standardized mean difference statistics.

It is interesting to note that these exclusion criteria eliminated a large number of studies (perhaps as many as half) that were included by the authors of the four meta-analyses used as a database in Heinsman's study (Becker 1990; Devine 1992; Slavin 1990; Tobler 1986). This is not, of course, to criticize those authors for their inclusion criteria; their purposes—to review substantive questions—could be answered adequately with the inclusion criteria they used. Heinsman's exclusion of studies using haphazard assignment is probably irrelevant to their purposes; such studies may not be easily classified as random or quasi-experiments, but they are certainly controlled outcome studies that address the substantive question. On the other hand, Heinsman's need to exclude this many studies does suggest that the estimates of differences between random and quasi-experiments those four authors provided may not be as accurate as Heinsman's, whose exclusion criteria were explicitly designed to provide the best answer to a limited methodological question. More generally, the same conclusion would probably hold for nearly any other study that reports differences between random and quasi-experiments (e.g., Smith et al. 1980). To the extent those studies reported such differences as secondary, exploratory analyses, their estimates are probably modestly suspect as well.

Overall Results

Overall, Heinsman (1993) found that the weighted average effect size of randomized experiments ($d+ = 0.42^*$) was significantly higher than the effect size for quasi-experiments ($d+ = 0.03$). (In this chapter, an effect size or a variance component that is significantly different from zero is marked with an asterisk). This finding was replicated in two of the four substantive areas (drug use prevention and ability grouping), with the other two areas yielding no difference between the two assignment methods. In a series of exploratory regression analyses, Heinsman tried to eliminate the assignment effect by including predictor variables, including second- and third-order interaction terms, that might account for the

variance in effect sizes. The effect was greatly reduced but could not be eliminated, even though 84 percent of the variance in effect sizes was explained with 37 predictors in the largest regression equation.

These results seem to suggest strongly that—on the average—randomized experiments may yield slightly larger effect sizes than quasi-experiments. Of course, this is an average main effect conclusion, whereas the presence of significant interaction terms in Heinsman's regression analysis raises the classic problem of whether it is still permissible to interpret the main effect. The authors think it is worth noting the main effect while cautioning future meta-analysts that it may be an unwise practice to assume that one can lump results from random and quasi-experiments together into a single substantive analysis. The test for differences between random and quasi-experiments should always be made first in the meta-analysis, and subsequent analyses should take the distinction into account if a significant difference is found.

Following up on a hypothesis suggested by Hedges (1983), Heinsman also examined variance component differences between randomized and quasi-experiments. Specifically, in a sample of 12 random and 12 quasi-experiments concerning the effects of open education, Hedges found that quasi-experiments yield larger variance components than randomized experiments. Hedges hypothesized that this might be due to a failure of quasi-experiments to equate groups at pretest. The hypothesis certainly seems plausible, but Heinsman was unable to replicate this effect using the 46 sample studies with pretest information. The variance component for quasi-experiments ($\&^2(\quad) = 0.12^*$) was significantly larger than zero, but not much larger than the variance component for randomized experiments ($\&^2(\ ) = 0.09^*$), which was also significantly larger than zero. In the four subareas, all the variance components were again significantly different from zero, with those for randomized experiments being quite similar in magnitude to those from quasi-experiments.

Despite this failure to replicate Hedges's (1983) finding, this hypothesis needs to be tested in future research. After all, the size of the variance component may reflect the effects of fixed-effects covariates, and a fairer test would partial those effects out before computing the final variance component figures. This could probably be done by predicting residual effect sizes after removing the effects of covariance in a regression equation, and then recomputing the variance components.

Heinsman also examined pretest effect sizes and the relationship between pretest and posttest effect sizes in randomized versus quasi-

experiments. The aim was to see whether differences between randomized and quasi-experiments at posttest might be accounted for by corresponding differences at pretest. Unfortunately, the findings were rather complex: Average pretest effect sizes were not significantly different in comparing 21 randomized ($d+ = 0.08^*$; $\&^2(\ ) = 0.00$) versus 25 quasi-experiments ($d+ = 0.04$; $\&^2(\ ) = 0.00$), at least within the sample of 46 studies that had pretest data, and the variance components were zero at pretest, as would be expected. Further, pretest effect sizes correlated positively and significantly ($r = 0.68^*$) with posttest effect size. Unfortunately, the sample of 46 studies with a pretest also showed no difference at posttest between randomized ($d+ = 0.28^*$; $\&^2(\ ) = 0.02^*$) and quasi-experiments ($d+ = 0.26^*$; $\&^2(\ ) = .06^*$), taking away the very effect the authors wanted to explain. By contrast, the sample of 66 studies without pretests showed a large difference between randomized ($d+ = 0.50^*$; $\&^2(\ ) = 0.11^*$) and quasi-experiments ($d+ = -0.09^*$; $\&^2(\ ) = 0.20^*$), but pretests were not available to test the authors' hypothesis. Especially given Heinsman's finding of significant covariation between pretest and posttest effect sizes, however, this hypothesis clearly needs further study.

Tangentially, it is worth commenting on the pretest effect size data itself. First, consider the randomized experiments. In theory, the mean effect size and variance components at pretest should be zero in randomized experiments. But the mean effect size, although small, is significantly larger than zero. Possible explanations include sampling error; attrition, that is, reporting of pretest data only on subjects who completed the experiment; investigators' decision to rerandomize if initial randomization favors control subjects, or not rerandomize if initial differences favor treatment subjects; or indicating random assignment that actually was not done. However, none of these points can easily be addressed using meta-analytic methodology. Second, consider the quasi-experiments. Their average effect size and variance components are both reported as zero. For whatever reason, these investigators seem to equate groups at pretest as do the randomized experiments, at least on observed measures (not necessarily expectations). Further research is currently underway to see if this might partly be due to the use of matching. If so, quasi-experiments that matched ought to have zero effect size at pretest, while those that did not would exceed zero. It will then be interesting to explore posttest scores by the same breakdown to see if there is any evidence of the regression to the mean that methodologists claim might occur in quasi-experiments as a result of matching on pretest scores.

Consequently, Heinsman (1993) concluded that random assignment tends to increase the size of standardized mean difference statistics

relative to nonrandom assignment, and that this effect could not be eliminated (although it could be made much smaller) even by trying to capitalize on chance as much as possible in the selection of covariates in a regression equation. Note that this result was also found by Becker (1990) using a similar methodology. A more extensive report of this work can be found in Heinsman and Shadish (in press). (Incidentally, another University of Memphis student (Ragsdale) is doing essentially the same study for a master's thesis with the entire sample of 100 studies from the marital and family psychotherapy research literature. This study will replicate Heinsman's findings on a different literature, one that has traditionally shown no difference between randomized versus quasi-experiments.)

Analyses on Heinsman's Drug Use Prevention Sample

One of the four areas in the Heinsman (1993) study was drug use prevention, using 30 studies from Tobler's (1986) meta-analysis on that topic. Heinsman found that the results from the overall analysis replicated consistently in this subsample. For this area, the overall weighted least squares (WLS) average effect size for 13 randomized studies was d+ = 0.51*, compared to d+ = 0.15* for 17 quasi-experiments, the difference being highly significant. The variance component for the randomized experiments was $\&^2(\ ) = 0.13$*, compared to $\&^2(\ ) = 0.10$* for quasi-experiments—both significantly different from zero but not substantially different from each other. The only other finding from Heinsman's analysis that is worth mentioning is that differences between randomized and quasi-experiments appeared only on measures of knowledge, attitude, and the like; measures of behavior showed no difference between randomized and quasi-experiments, no doubt at least partly because both effect sizes were zero—that is, the interventions did not seem to affect actual behavior.

Heinsman (1993) did not apply the kind of regression analyses used with the overall sample to the drug use prevention subsample. Hence the data were reanalyzed with the same purpose as before—to see if the effect favoring randomized experiments could be made to disappear. Again, the authors could not make it disappear. Potential covariates were selected by conducting 15 individual regressions; in each regression the effect size was predicted from assignment, from the covariate, and from the interaction of assignment with the covariate. This yielded 17 possible predictors that were entered into a WLS regression predicting effect size. As expected given the small sample size relative to the number of predictors, the multiple correlation was quite high at R = 0.96 and was highly significant ($Q_r = 657.43$, df = 17, $p < .001$; $Q_e = 54.12$, df = 12, $p < 0.001$). As in the overall analysis, the predictor for

random assignment was still positive (= 0.55*) and significant using the difference between Qr with and without assignment as a predictor (Q_{diff} = 85.23, df = 1, p < 0.001). Of course, the small number of studies involved in this analysis necessitates caution in interpreting the effects. But it is worth noting that the conclusion is the same as that found in the overall analysis: Random assignment increases the size of the effect, and this effect cannot be eliminated even when trying to capitalize on chance to do so. Of course, the same caveat mentioned earlier applies here; one must interpret the main effect for randomized experiments cautiously in the presence of significant interactions in the regression equation.

SEPARATING THE EFFECT

If there is a main effect, however, a logical next step might be to try to explain the effect. One way to do this is to try to separate the effect into different parts, each part being routed through a different mediator variable. The method used for this analysis has been presented several times in recent years (Shadish 1992; Shadish and Sweeney 1991), and subjects meta-analytic data to linear structural modeling techniques. At the outset, of course, it must be acknowledged that this analysis should be viewed as highly exploratory and tentative for many reasons. Some of those reasons have to do with the ambiguities associated with mediational models in correlational data, and others have to do with whether the particular statistical approach taken is the most appropriate for modeling meta-analytic data. These objections have real merit, even though the present chapter may not be the place to discuss them in detail (but see Becker and Schram 1993; Shadish 1992, 1996; Shadish and Sweeney 1991). For present purposes, the analysis has two objectives: to shed some light on possible explanations for any effect that random assignment may have on study outcomes, and to stimulate more thinking in meta-analysis about how such mediational models might best be pursued.

The initial model is presented in figure 1; this model was fit in a structural equations modeling program (Bentler 1992), using as input a WLS covariance matrix that was generated from a computerized regression program (SPSS 1990). This model approached but did not reach an acceptable overall fit (0^2 = 28.65, df = 6, p < 0.001; comparative fit index (CFI) = 0.844; consistent version of Akaike's information criterion (CAIC) = -4.92). This model consisted of four mediational paths; for ease of interpretation, only the significant paths are included in figure 1, along with the standardized path coefficient for that path. In the first path, the mediator was whether or not the control group in the treatment-control comparison was active or passive. Passive controls included

no-treatment and wait-list control groups with the subjects receiving little or no intervention, and active control groups were placebo and treatment-as-usual controls with subjects receiving an intervention of some sort. Results suggested that randomized experiments used passive controls more often than did quasi-experiments, and the use of such controls increased overall effect size. The net effect is that randomized experiments yield larger effect sizes.

The second path used internal versus external control as a mediator. An internal control is one selected from the same pool of subjects, such as students from the same grade levels in the same schools; an external control is selected from a pool of subjects that is patently different from those in the treatment group, such as students in another city. Results suggested that randomized experiments were much more likely to use internal controls—indeed, they use them definitionally—whereas quasi-experiments used external controls as well. It was hypothesized that results from studies with external controls might be less likely to resemble randomized trials, but the use of such controls was unrelated to effect size in the present model.

The third path included self-selection versus other-selection into treatment as the mediator. Results suggested subjects do not self-select into treatment in randomized trials—again, this is definitional—while they do sometimes self-select into treatment in a quasi-experiment. Self-selection, in turn, seems to decrease study effect size. Hence quasi-experiments end up producing lower effect sizes as a result.

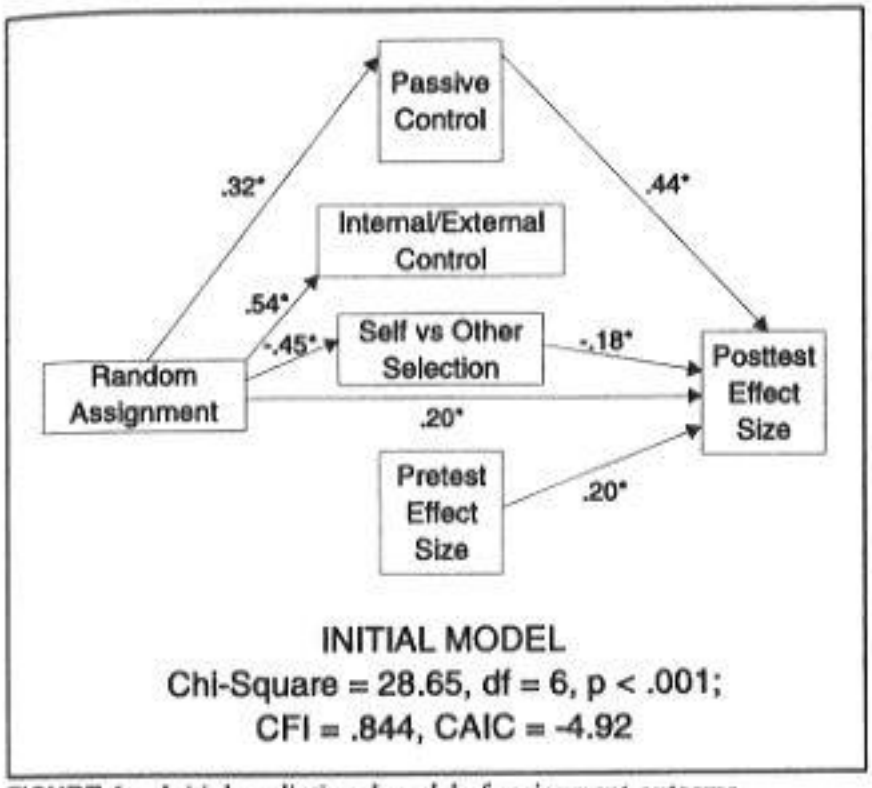

FIGURE 1. *Initial mediational model of assignment-outcome relationship.*

Presumably this mediator needs some explanation as well, and the explanation will presumably hinge on the nature of the selection processes in each area; this is a matter that future researchers can follow up.

The fourth path used pretest effect sizes as a mediator. As figure 1 clearly shows, pretest effects sizes are modestly but significantly and positively related to posttest effect size. But no significant relationship existed between assignment method and pretest effect size. This lack of relationship is the same result found in Heinsman's (1993) related analysis to the same effect. A cautionary reminder about this variable as implemented in figure 1, however, is that one must recall that pretest effect sizes were not present for about half the studies. The authors used mean substitution to estimate the missing pretests for this model, and have some reason to think such missing data estimates are not very good. Hence this path should be regarded with caution.

A final point about figure 1 is that it contains a significantly positive direct path between random assignment and posttest effect size. The

addition of this path significantly improves the fit of the model. Substantively, this means that the four mediator variables in figure 1 are not themselves capable of fully explaining the effect of random assignment on effect size. This replicates the conceptual conclusions from Heinsman's (1993) regression analysis, but with a different analytic strategy.

However, since the initial model did not fit acceptably well, it was modified slightly in a series of specification searches to try to obtain a better fit. The resulting model is presented in figure 2, and it fit acceptably well ($0^2 = 13.21$, df = 7, $p < 0.067$; CFI = 0.945; CAIC = -25.96). Of course, one should be doubly cautious about interpreting this subsequent model because it suffers from all the flaws of the first model plus those associated with capitalization on chance in the specification search. Nonetheless, it is also worth noting that this final model closely resembles the initial model, and that the paths common to both models have largely the same coefficient values. This suggests that one can interpret the model with at least a modicum of confidence that it is not entirely due to chance.

Four findings from this final model are worth noting. First, three paths from the initial model remained the same in the final model: randomized experiments more frequently used passive controls, which increased effect size; they also used other-selection into treatment more often, which also increased effect size; and pretest effect size was unrelated to assignment method, but was positively related to posttest effect size. Second, the path involving use of internal versus external controls was dropped; parts of this path were not significant in the initial model, so this does not depart much from the initial model. Third, a new path was added through the use of exact effect size methods as a mediator. An exact effect size method is one that yields Cohen's d; inexact methods try to approximate Cohen's d, for example by using information from a three-group F-ratio to estimate the pooled standard deviation of Cohen's d when means and sample sizes are available but standard deviations are not. Results suggest that exact methods yield smaller effect sizes, and that randomized experiments are less likely to allow use of exact methods, so that the overall effect is to increase effect sizes from randomized experiments. Fourth, note that the direct effect of random assignment on posttest effect size is still positive and significant.

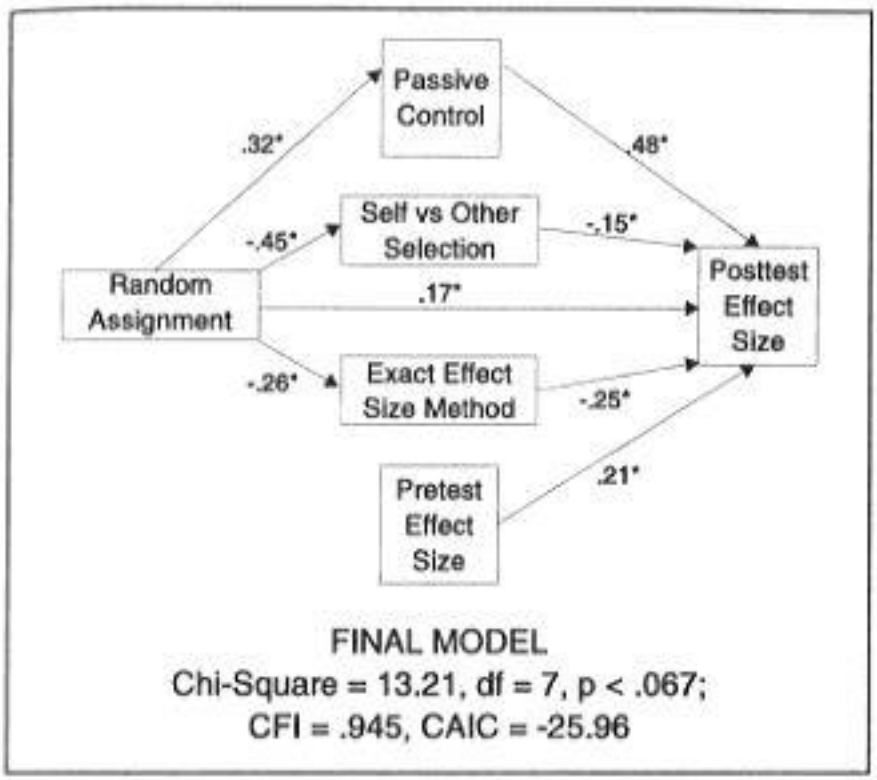

FIGURE 2. *Final mediational model of assignment-outcome relationship.*

Discussion of Figure 2

What is particularly gratifying about this final model is that it makes good conceptual sense and at the same time points to areas for potential future research on this topic. Perhaps the most intuitively sensible path is the one involving use of passive controls. Those who write about methodology have speculated for years that passive controls should yield larger effects than active controls (e.g., Cook and Shadish 1986), so it is gratifying to see the results support the hypothesis. Indeed, this is one of those conclusions that in retrospect seems so obvious that readers of this chapter might rightly respond, "I could have told you that."

The self-versus-other selection path points to the theoretically obvious role that selection bias almost certainly plays in the outcomes of quasi-experiments. The challenge here is mostly one for future research: other than knowing selection bias must somehow be involved, this particular coding reveals relatively little about the mechanisms underlying the bias. Researchers need to develop better ways to measure these mechanisms, ideally methods

that are codable in meta-analysis to the extent that authors provide sufficient information in their reports. The path that was dropped from the initial to the final model (involving the use of internal versus external controls) was such a code, and showed some promise even though it did not survive in the final model. However, selection bias is also quite likely to involve mechanisms that vary from substantive area to substantive area, so that area-specific codes would also be worth developing, especially in meta-analyses sampling larger numbers of studies from one area than in the present analysis.

The path involving method of effect size calculation involves a variable that the authors have wondered about for years. Almost everyone who actually conducts a meta-analysis complains about poor reporting in primary studies. Nowhere is this more crucial than in poor reporting of the statistics to compute effect sizes, for without effect sizes there is no dependent variable at all. As a consequence, meta-analysts have developed a set of techniques to allow computing effect sizes under adverse circumstances; these techniques range from those that are well thought out and statistically justified to those that are best described as ad hoc. It is not surprising that different estimates may result from such approximations. This, combined with the fact that such approximations are widely used in meta-analytic practice, suggests that statisticians would do a great service to the field by investigating this matter further. But the matter can also be investigated empirically; a student at the University of Memphis wrote a dissertation on the topic. That student selected about 150 studies from the authors' database allowing computation of exact effect sizes, and then computed all possible approximate effect sizes on the same data in order to compare exact versus approximate bias, both in mean and variance components (Ray 1995). Ray's (1995) results confirm that these inexact methods can yield quite different answers.

Elsewhere Shadish (1992) has noted that the fit statistics yielded by common structural equation modeling programs are somewhat different in interpretation from those yielded by the meta-analytic statistics proposed by Hedges and Olkin (1985). In essence, the difference is that the statistics do not take into account possible random effects in the population effect size(s), whereas the Hedges-Olkin statistics do take them into account. Thus, even though the authors' fit statistics suggest the model might be compatible with the data, random effects cannot be tested using this method. It would be possible to approach this matter by testing models like those in figures 1 and 2 using ordinary regression analyses, modified as Hedges and Olkin suggest, to obtain fit statistics that take random effects into account. The procedure would be the same as the regression analyses Heinsman (1993) conducted, reported earlier in

this chapter. However, mediational models such as those in figures 1 and 2 cannot be represented with just one regression equation. In the case of figure 2, for example, four regression equations would be needed to represent the significant paths. While Hedges-Olkin fit indices could be computed separately for each of those four equations, there is as yet no way to cumulate those fit statistics to provide a test of the overall fit of the model. This problem needs further attention by statisticians.

Methodologically, the procedures used here differ from those reported by Becker (this volume) in ways worth noting. Becker cumulates covariance estimates from individual studies that provide such estimates. Instead, this procedure used raters to generate data about each study, and then directly computed covariances among relevant variables in the model. Shadish (1992) has referred to this as a difference between "study-generated" and "rater-generated" data, and has discussed the two methods in more detail elsewhere (Shadish 1992). As described, Becker's (this volume) approach has significant advantages when it is possible; however, it is not always possible. Relatively few studies report the covariances of interest, whereas raters can usually generate codes for at least some of those variables. Further, the kinds of variables the authors examined (e.g., kind of assignment or the type of control group used) do not lend themselves to within-study covariances because they frequently do not vary within a study. The current approach offers significant advantages over Becker's in these situations, and so is especially appropriate when the model involves study-level variables such as those examined in the present study. Shadish (1996) elaborates these matters.

DISCUSSION

Overall, these results seem to suggest that the answers provided by randomized experiments may be at least modestly different from those provided by quasi-experiments. The size of the difference was substantial in the largest cases, especially in the drug use prevention studies where the effect size was over three times larger for randomized compared to quasi-experiments. But because the analyses indicate that at least some of this difference may be an artifact of covariates, a more conservative estimate is warranted. Extrapolating from figures 1 and 2, which seem to yield the most conservative estimate of the impact of randomization, the unstandardized version of the path coefficients in those figures suggests that randomization might increase effect size by about 0.15 units of d. Even this small value is nontrivial—especially when one is dealing with findings that may be as close to zero as yielded by the quasi-experiments in this study; an increment of even that

modest magnitude might well mean the difference between detecting versus not detecting an effect.

This overall result has at least two kinds of implications; one is methodological. Given the role of selection bias in quasi-experiments, more investigation is needed on the nature of such biases so that researchers can explore the circumstances under which quasi-experimental controls might well approximate randomized controls. The other implication is for meta-analysts. Given the authors' findings, the common practice in most literatures of combining randomized and quasi-experiments is questionable at best. This is a situation in which theory suggests that one of the two methods—the randomized experiment—is likely to yield a better answer than the other. If the two differ, then lumping them together produces a more biased estimate than keeping them separate. While one does not wish to discourage meta-analysts from exploring results yielded by quasi-experiments, it is important that they exercise caution in doing so. When differences between the two methods are found, they ought to provide separate estimates of treatment effectiveness for each of the two methods in order to avoid biased estimates.

But these results are clearly far too preliminary to place great faith in at this point. Further research may, for example, show that the finding favoring randomized experiments may prove to be artifactual, a result of covariates not included in the present study. More seriously, there are good reasons to think that there may be some variation in the finding over substantive areas. In the authors' data, two of the four areas showed no significant differences between randomized and quasi-experiments in simple univariate tests. Although overall regression analysis purported to take this into account through inclusion of various interaction terms involving the substantive area, one cannot be confident of the results. In fact, a preliminary regression analysis on the subset of 41 studies from Devine's (1992) patient education data still suggested no significant effect for random assignment to conditions. Furthermore, it must also be recalled that when this question has been investigated with medical and surgical interventions, results suggest that quasi-experiments yield larger effect sizes than randomized experiments, or just the opposite of the present findings. More generally, it would seem that any effect size differences that might emerge between randomized versus quasi-experiments would have to be due primarily to selection bias. Selection bias, in turn, seems almost certain to involve significant area-to-area variation. So, despite findings reported in this chapter, it is quite likely that the answer to the basic question will vary from subject area to subject area.

The problems faced by meta-analysts are legion, mostly having to do with the many potential confounds of the randomized versus quasi-experiment distinction (especially those that themselves may interact with substantive area) and the many variables that can hardly be coded. For example, 85 percent of the randomized studies made no mention of what random number generator was used, and 77 percent did not say who did the random assignment. In fact, meta-analysis has obvious limitations of this sort that have no easy remedy. Meta-analytic investigations of this question need to be complemented by studies that examine these variables more directly, such as Dennis' (1988) dissertation. Problematically, of course, these methodological studies—meta-analytic or direct—cost money to do, but are rarely fundable in their own right.

Finally, it is important to return to a point alluded to earlier in this study when trying to explain the significant positive effect size at pretest in random experiments. It was said that perhaps the experiments weren't really random. In point of fact, it is very difficult to know whether the authors of the research used random assignment to conditions. One problem is that randomization may be something researchers say to get published or funded, knowing full well that the actual procedure was not or will not be truly random. Another explanation is faulty implementation of random assignment. To judge from research (Dennis 1988), implementation problems are frequent, but rarely mentioned in published form. In fact, Dennis' research suggests that the authors of publications are often not even aware of these implementation problems because, for example, random assignment may have been conducted by a secretary who was not in frequent contact with the author. Another explanation appears to be that some researchers may not understand what random assignment means and how it should be done. The author of one study considered for inclusion in this study, for example, said subjects were randomly assigned to conditions, but later also said that subjects chose which group to enter based on which group fit their schedule. Other authors have said that they randomly assigned, but also that after random assignment they moved subjects from one cell to the other in order to balance some important characteristic such as gender or age. One wonders how often these things occur without being mentioned in published form!

The good news in all this, of course, is that such questions are grist for the mill to be ground out in future research. Perhaps such questions, illustrated by the present research and studies like it, are the beginnings of a latent research area that one might call the empirical program of methodology. After all, most methodologists have tended to write about their topic as if it were entirely a theoretical matter, not subject to empirical investigation. What empirical research exists has tended to be done mostly by

statisticians, most often using Monte Carlo techniques that are informative but may have less direct relationship to research done in actual practice. Meta-analytic inquiries such as the present one, as well as the more direct empirical studies that examine methodological practices as they occur when research is implemented, are badly needed to complete the understanding of effective research techniques.

REFERENCES

Becker, B.J. Coaching for the Scholastic Aptitude Test: Further synthesis and appraisal. *Rev Educ Res* 60:373-417, 1990.

Becker, B.J., and Schram, C.M. Models in research synthesis. In: Cooper, H.M., and Hedges, L.V., eds. *The Handbook of Research Synthesis*. New York: Russell Sage Foundation, 1993.

Bentler, P.M. *EQS: Structural Equations Program Manual*. Los Angeles: BMDP Statistical Software, Inc., 1992.

Colditz, G.A.; Miller, J.N.; and Mosteller, F. The effect of study design on gain in evaluation of new treatments in medicine and surgery. *Drug Inform J* 22:343-352, 1988.

Cook, T.D., and Shadish, W.R. Program evaluation: The worldly science. *Ann Rev Psychol* 37:193-232, 1986.

Dennis, D.L. "Implementing Random Field Experiments: An Analysis of Criminal and Civil Justice Research." Ph.D. dissertation, Northwestern University, Evanston, IL, 1988.

Devine, E.C. Effects of psychoeducational care with adult surgical patients: A theory probing meta-analysis of intervention studies. In: Cook, T.D.; Cooper, H.M.; Cordray, D.S.; Hartmann, H.; Hedges, L.V.; Light, R.J.; Louis, T.A.; and Mosteller, F., eds. *Meta-Analysis for Explanation: A Casebook*. New York: Russell Sage Foundation, 1992. pp. 35-82.

Gilbert, J.P.; McPeek, B.; and Mosteller, F. Statistics and ethics in surgery and anesthesia. *Science* 78:684-689, 1978.

Hedges, L.V. A random effects model for meta-analysis. *Psychol Bull* 93:388-395, 1983.

Hedges, L.V., and Olkin, I. *Statistical Methods for Meta-Analysis*. Orlando, FL: Academic Press, 1985.

Heinsman, D.T. "Effect Sizes in Meta-Analysis: Does Random Assignment Make a Difference?" Ph.D. dissertation, University of Memphis, Memphis, TN, 1993.

Heinsman, D.T., and Shadish, W.R. Assignment methods in experimentation: When do nonrandomized experiments approximate the results from randomized experiments? *Psychol Methods*, in press.

Ray, J.W. "An Evaluation of the Agreement Between Exact and Inexact Effects Sizes in Meta-Analysis." Ph.D. dissertation, University of Memphis, Memphis, TN, 1995.

Shadish, W.R. Do family and marital psychotherapies change what people do? A meta-analysis of behavioral outcomes. In: Cook, T.D.; Cooper, H.M.; Cordray, D.S.; Hartmann, H.; Hedges, L.V.; Light, R.J.; Louis, T.A.; and Mosteller, F., eds. *Meta-Analysis for Explanation: A Casebook*. New York: Russell Sage Foundation, 1992. pp. 129-208.

Shadish, W.R. Meta-analysis and the exploration of causal mediating processes: A primer of examples, methods, and issues. *Psychol Methods* 1:1-19, 1996.

Shadish, W.R., and Sweeney, R. Mediators and moderators in meta-analysis: There's a reason we don't let dodo birds tell us which psychotherapies should have prizes. *J Consult Clin Psychol* 59:883-893, 1991.

Slavin, R. Achievement effects of ability grouping in secondary schools: A best-evidence synthesis. *Rev Educ Res* 60:471-499, 1990.

Smith, M.L.; Glass, G.V.; and Miller, T.I. *The Benefits of Psychotherapy*. Baltimore: Johns Hopkins University Press, 1980.

SPSS. *SPSS Reference Guide*. 1990. Available from author, 444 N. Michigan Avenue, Chicago, IL 60611.

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

ACKNOWLEDGMENT

The author gratefully acknowledges the contributions of the late Donna T. Heinsman, Ph.D., to this research and to the field.

AUTHORS

William R. Shadish, Ph.D.
Professor
Department of Psychology
Memphis State University
Memphis, TN 38152

Donna T. Heinsman, Ph.D.
Deceased

Click here to go to page 165