Meta-Analysis of Drug Abuse Prevention Programs

Editor:

William J. Bukoski, Ph.D.

NIDA Research Monograph 170
1997

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
National Institutes of Health

National Institute on Drug Abuse
Division of Epidemiology and Prevention Research
5600 Fishers Lane
Rockville, MD 20857

ACKNOWLEDGMENT

This monograph is based on the papers from a technical review on "Meta-Analysis of Drug Abuse Prevention Programs" held on July 26-27, 1993. The review meeting was sponsored by the National Institute on Drug Abuse.

Opinions expressed in this volume are those of the authors and do not necessarily reflect the opinions or official policy of the National Institute on Drug Abuse or any other part of the U.S. Department of Health and Human Services.

The U.S. Government does not endorse or favor any specific commercial product or company. Trade, proprietary, or company names appearing in this publication are used only because they are considered essential in the context of the studies reported herein.

National Institute on Drug Abuse
NIH Publication No. 97-4147
Printed 1997

NIDA Research Monographs are indexed in the *Index Medicus*. They are selectively included in the coverage of *American Statistics Index, BioSciences Information Service, Chemical Abstracts, Current Contents, Psychological Abstracts,* and *Psychopharmacology Abstracts.*

Click on title or page number to go to page

Table of Contents

Meta-Analysis of Drug Abuse Prevention Research

William J. Bukoski

INTRODUCTION

After nearly 15 years of declining rates in adolescent drug abuse, current epidemiologic research indicates significant increases in the use of a variety of illicit drugs of abuse such as inhalants, marijuana, cocaine, lysergic acid diethylamide (LSD), and phencyclidine (PCP) (Department of Health and Human Services 1994) by children and youth in the 8th, 10th, and 12th grades. Faced with these alarming increases in drug abuse, concerned parents, educators, and community leaders are turning to prevention research to better understand the nature of these recent trends and to guide prevention policy and program development. Critical to effective preventive action at all levels of Government is an assessment of the numerous scientific findings that have been published over the past decade that may indicate which prevention practices are efficacious and which drug abuse prevention strategies need to be considered for implementation in school and community programs in order to bring a halt to increased drug abuse by the Nation's youth.

To assist in this deliberative process, the National Institute on Drug Abuse (NIDA) has consulted with a number of this country's best scientists to analyze prevention research findings from a variety of published studies and to integrate those results into a meaningful and objective meta-analysis in order to identify promising drug abuse prevention strategies and policies. Given the complexities of the published prevention research, it was decided that the meta-analysis of research findings should follow the systematic procedures employed in this methodology and utilize a common standard or metric that would permit the comparison and integration of outcomes across a variety of individual studies (Cook et al. 1992). Central to this process is the calculation of a metric that is called the effect size. The effect size provides, in standard deviation units, an objective and uniform measure of quantitative differences in drug prevention outcomes such as self-reported drug use, knowledge of negative consequences of drug abuse, and antidrug-abuse attitudes that could be attributed to the exposure of the treatment group that had been randomly assigned to an experimental prevention intervention in comparison to a control group that did not receive the program.

To conduct a meta-analysis, researchers identify salient prevention research studies. Using a standardized procedure, they calculate the effect sizes for drug-related outcome measures reported in each study. Given that effect sizes are calculated in units of standard deviation, the measurements are comparable across studies and hence subject to further analysis such as assessing the efficacy of different prevention program strategies. Rather than relying on findings from one study, meta-analysis provides a technically sound method of combining results from a variety of studies in order to identify the extent to which specific types of prevention programs are effective in reducing and preventing adolescent drug abuse.

The technique of meta-analysis provides a systematic and objective assessment of prevention research findings reported by many scientific studies and results in a convergence of higher order information that can only be provided by analysis of an entire body of research findings. Meta-analysis provides a standardized approach to the identification, selection, assessment, and interpretation of the results of a variety of medical, psychiatric, and behavioral research literatures and is particularly valuable in synthesizing research findings from an emerging science, such as drug abuse prevention research.

The practical outcome of NIDA's meta-analysis of prevention research is twofold: programmatic and methodological. Each chapter in this monograph addresses one of these two objectives.

In the first section of the monograph, Tobler presents a meta-analysis of adolescent drug abuse prevention research findings; Schmidt and colleagues provide a meta-analysis of integrity tests for predicting drug and alcohol abuse; and Becker provides an approach for meta-analysis of drug-related risk and protective factors research. In the second section of the monograph, several chapters explore the appropriateness and special methodological considerations that must be addressed when conducting a meta-analysis of the drug abuse prevention research literature. Perry's chapter focuses upon methods to calculate effect sizes; Devine's chapter discusses issues in coding prevention intervention studies; Shadish and Heinsman assess the differences in outcomes produced by experimental versus quasi-experimental studies; Matt explores issues concerning generalized causal inferences related to program effects; Hansen reviews approaches to classifying independent variables and types of correlational relationships between dependent and independent variables; in separate chapters, Lipsey and Hedges discuss potential applications of meta-analysis for policy development; and Bangert-Drowns presents general advantages and potential limitations of conducting and utilizing meta-analysis in drug abuse prevention research.

Collectively these chapters provide a current overview of the efficacy of drug abuse prevention programs and related measurement systems and help define the techniques employed in meta-analysis of drug abuse prevention programs. The monograph provides firsthand guidance in the application of research findings from meta-analysis and appropriate discussion of key technical procedures that should be considered in conducting future meta-analyses of drug abuse prevention research. It also helps to delineate what prevention programs and policies appear to be the most effective in combating drug abuse by adolescents and young adults who may be entering the workplace.

This publication clearly illustrates the value of being able to combine findings from specific high-quality primary research studies into a cohesive summary that better defines what the science of drug abuse prevention offers to guide future program decisionmaking by prevention practitioners. It is expected that future decisions concerning prevention programs and policy at the Federal, State, and community level will be enhanced by practical application of these findings, leading to the implementation of more effective drug abuse prevention strategies at all levels.

REFERENCES

Cook, T.; Cooper, H.; Cordray, D.; Hartmann, H.; Hedges, L.; Light, R.; Louis, T.; and Mosteller, F., eds. *Meta-Analysis for Explanation*. New York: Russell Sage Foundation, 1992.

Department of Health and Human Services. *HHS News* --- *HHS Releases High School Drug Abuse and DAWN Surveys.* Department of Health and Human Services, National Institutes of Health. Press release December 12, 1994.

AUTHOR

William J. Bukoski, Ph.D.
Chief
Prevention Research Branch
Division of Epidemiology and Prevention Research
National Institute on Drug Abuse
Parklawn Building, Room 9A-53
5600 Fishers Lane
Rockville, MD 20857

Meta-Analysis of Adolescent Drug Prevention Programs: Results of the 1993 Meta-Analysis

Nancy S. Tobler

INTRODUCTION

Policy relevant conclusions emerge when meta-analytic techniques are used to achieve consensus out of the inconsistencies found in individual research studies. Extensive search procedures located 120 school-based drug prevention programs that evaluated success on self-reported drug use measures. Hypothesis tests were made of an a priori classification scheme for the type of program. Six types of programs were identified based on content and delivery and were collapsed into noninteractive and interactive programs.

Because programs varied from 20 to 6,000 students, both ordinary least squares regressions (unweighted effect size) and weighted least squares regressions (weighted effect size) were conducted. Six covariates were entered into the regressions: sample size, targeted drug, type of control group, special populations, type of leader, and attrition. The relationship between program content, delivery, and the size of the programs was examined.

Interactive programs were significantly superior to the noninteractive programs in their ability to impact drug use behaviors and were equally successful for tobacco, alcohol, marijuana, and other illicit drugs. The effectiveness of the interactive programs was not only replicated, but increased with a subset of 56 high-quality experimental programs.

A meta-analysis of 143 adolescent drug prevention programs was completed by the author in 1986 and was reported elsewhere (Tobler 1986, 1992*b*. Tobler (1992*b*) is a reanalysis of 91 programs (a subset of the original 143 programs) that measured change solely on drug use outcome measures.[1] This chapter is based on a second data set which includes 120 adolescent drug prevention programs. One publication (Tobler 1993) emphasizes substantive material and gives a thorough description of the types of programs. Two publications report the methodology, inferential statistics, and the major findings (Tobler 1992*a*, 1994). In Tobler (1994) the data were reanalyzed to verify the major findings in Tobler (1992*a*), using a reduced set of relevant variables as covariates so that the number of parameters is more in line with the number of cases.[2] This reanalysis is summarized briefly in this chapter.

For purposes of brevity, the two different meta-analyses are called 1986 and 1993. A number of differences should be noted. First, in 1986 the type[3] of prevention program was determined on a single dimension: the content or subject matter of the program. In 1993, this was expanded to include two dimensions: the content or subject matter and how the program content is delivered.[4] Second, the 1993 sample of drug prevention programs was limited to school-based prevention programs, whereas the 1986 sample included both school and community-based programs. Third, as adolescent drug use peaked in 1978, the 1993 meta-analysis examined only 1978 to 1990 data (versus 1972 to 1984 in the 1986 meta-analysis). This choice was made to reflect the downward societal trends in drug use (Johnston et al. 1986, 1989). Fourth, the 120 programs in the final 1993 sample all used drug use measures, versus 91 programs in the 1986 meta-analysis. The final set of programs in 1993 included 81 programs identified after the 1986 meta-analysis and 39 from the previous 1986 meta-analysis.

Finally, in 1993 the newest meta-analytical methodology was used to avoid the potential problem of arriving at incorrect conclusions due to inappropriate statistical procedures (Hedges and Olkin 1985; Hunter and Schmidt 1990; Rosenthal 1986). The 1993 results include both the unweighted effect size (UNES) and the weighted effect sizes (WES). Glass and colleagues (1981) defined UNES as the standardized mean difference between the treatment and the control group:

$$ES = (X_e - X_c)/SD_c, (1)$$

where ES = effect size, X_e and X_c are the means for the experimental and control group, respectively, and SD_c is the standard deviation (SD) of the control group. In drug prevention research, parametric statistics are reported[5] which are computed using the pooled SD. To keep effect sizes comparable, it is more appropriate to use statistics which use the pooled SD, such as Cohen's d or its equivalent Hedges' g. The WES is then computed by weighting each effect size by the inverse of the variance, an estimate of the sample size. Hedges' (1986, p. 739) formula for the weighting factor of an individual study is:

$$W_i = [2(n_{ei} + n_{ci})n_{ei}\, n_{ci}] / [2(n_{ei}+n_{ci})^2 + n_{ei}\, n_{ci}\, d_i^2], \quad (2)$$

where W_i = weighting factor of the study, d_i = unweighted effect size, n_{ei} = number in the experimental group, and n_{ci} = number in the control group. Use of WESs is based on the fact that larger samples produce more stable results.

SELECTION CRITERIA

Selection Criteria for 1993

Criteria for inclusion in the 1993 meta-analysis were: (a) school-based drug prevention programs available to all members of the student body (may have included but did not target high-risk youth[6]); (b) reporting of drug use outcome measures; (c) use of a control or comparison group (comparison groups must have both pretest and posttest); (d) grades 6 to 12 (5th grade if incorporated into a middle school and/or longitudinal research was conducted); (e) goals of primary prevention, secondary prevention, and/or early intervention (does not target identified abusive/ compulsive or addicted drug users in treatment[7]); (f) participation of all ethnic groups that comprise the school's population; (g) location in United States and/or Canada; and (h) reported or published after 1977.

Additional Criteria for a Subset of Higher Quality Experimental Studies

The selection of a special subset of programs was made for two reasons. The first reason was to replicate the results with a set of solely experimental studies obtained from the mixed set of experimental and quasi-experimental studies. Many researchers feel that results of programs evaluated with quasi-experimental research designs yield overestimates of program effects; therefore, the analysis of a set of experimental studies will empirically examine this question. Second, as factors other than random or nonrandom assignment can impact evaluation results, a program was chosen that: (a) had a delivery intensity of not less than 4 hours (i.e., 1 week of classes); (b) administered a posttest not less than 3 months after pretest; (c) was not a placebo program even if the placebo program was compared to a control group (i.e., a program with one or more essential components deliberately excluded such as refusal skills); (d) was not compared to another treatment program, (e) had followed individuals in longitudinal research (i.e., no cross-sectional research); and (f) had a measure of control for preexisting differences even if these differences were reported as nonsignificant (i.e., effect sizes could be computed from a change score, covariance adjusted means, or the individual's level of drug use at pretest).

META-ANALYTIC METHODOLOGY

Coding Procedures

A 50-page codebook was compiled that included over 250 variables related to: (a) treatment components (see table 1); (b) participant characteristics (e.g., grade, sex, ethnicity, socioeconomic class); (c) program characteristics (e.g., year, source of publication, goal, targeted drug, funding, location, number involved, number tested, research center); (d) implementation factors (e.g., intensity, duration, boosters, leaders, hours and type of leader training); (e) research methodology (e.g., sampling, assignment, unit of assignment, type of control group, research design, threats to internal validity); (f) test instrumentation (e.g., reliability, test-retest, internal consistency, reactivity of measure); and (g) data analysis (e.g., unit of data analysis, method of effect size calculation). In coding studies, the main focus was on gaining as much information as possible about the programs. If information was missing in the primary report or ambiguities needed clarification, researchers were contacted or additional literature searches were initiated. The principal investigator and two research associates independently coded all the content items. Ambiguous coding interpretations became the topic of discussion in the 2-hour weekly meetings and misinterpretations or errors were corrected.

A second "Manual for Effect Size Calculations" was developed for converting each of the summary statistics encountered (see Tobler 1992*a*, appendix 3). The principal investigator and two doctoral research associates, working independently from those coding content items, conferred about the choice of outcome measures and statistical procedures to use in calculating the effect size. Calculations were aided by a special computer software program (Tobler 1992*a*) and were spotchecked by the principal investigator.

Analysis—A Program

A program is the unit of analysis. In meta-analysis, studies are most often the unit of analysis with one effect size being reported per study (Bangert-Drowns 1986). But in drug prevention program research, some studies (i.e., research projects) compared the efficacy of more than one type of program. As the type of program is the variable of interest, using the study as the unit of analysis would not allow comparisons about the type of program. For example, "a cognitive program, a decisionmaking program and a values-clarification program" were compared in a single experimental study reported by Goodstadt and Sheppard (1983, p. 362). The three different types of alcohol education programs were administered to independent groups of adolescents, thereby contributing three effect sizes, one for each program type.

It was also necessary to insure that only one effect size was contributed to the overall analyses of a single program and a single group of adolescents. Numerous articles or reports were written about a single program. Each of the articles related different information about the same program such as results for different testing periods (i.e., pretest information, immediate posttest, and followups). Often details about the program content, instrumentation, and implementation were included in separate publications. To insure independence of a sample of students, all authors were cross-checked against all other authors in the database to identify duplicate reports on the same group. Sets of articles or reports were then sequenced by pretest, posttest, and followup results and given one program number.

Independence of Outcome Measure

Each outcome measure category estimated the effect of the program based on a different concept. If two or more effect sizes on the same outcome measure were reported for a program, they were averaged and recorded as one effect size. Using this procedure, a student was represented only once in a specific outcome measure category. As results were not averaged across outcome categories, a student could not be represented more than once in the overall analysis for that outcome measure.

Every outcome measure reported at baseline was traced through all testing periods. Frequently, a large number of these measures were not reported in the final results. It was assumed that failure to report on all of the initial measures indicated nonsignificant findings and an effect of zero was assigned, a conservative method.

Independence for Type of Drug

Effect sizes were kept independently for five categories of drugs: cigarettes, alcohol, marijuana, hard drugs (cocaine, heroin, stimulants, inhalants, and tranquilizers), and "all drugs." The "all drugs" category accommodated programs with various combinations of drugs not reported separately. If more than one effect was reported for a category, the mean was reported as a single effect for that category. Each category was kept independently to facilitate later analyses by type of drug. For the main analyses—one effect per program—the results were averaged across types of the drugs. Behavioral intentions were not included as a drug use measure.

Independence for Subpopulations

If results were broken out separately by sex, grade, and/or level of drug use (nonuser, experimental user, user), individual effect sizes were calculated. For example, if three types of outcome measures were reported for boys and girls for three levels of drug use, 18 effects were computed (3 outcomes x 2 sexes x 3 levels). "Because...different students are involved in each of these comparisons, the effect sizes derived from the comparisons are independent" (Giaconia and Hedges 1982, p. 585). To obtain one program effect for the final analysis, the effect size for each subpopulation was averaged. For example, in a program having a positive effect for the boys and a negative effect for girls, the mean effect for the program is zero and does not accurately portray the program's results. Bangert-Drown's (1986) study effect method (one effect per program) does not take into account differential results across subpopulations. Because the WES was used, the weighting factors for the individual subpopulations were also combined into a single weighting factor for the program. But, in this case, the sums of the individual subpopulation weights were computed to be used at the aggregate level (see Tobler 1994).

Pooling Effect Sizes Over Test Intervals for a Single Program

Effect sizes were computed for each subpopulation for all testing periods reported. The exact number of months from pretest to posttest and/or followup was coded. A categorical variable was created: (a) 1 to 12 months, (b) 13 to 24 months, (c) 25 to 36 months, and (d) greater than 37 months. If more than one test was given in an interval, the average was reported. This occurred frequently in the first time interval as many programs gave a posttest and followup test within 12 months. None of the time intervals included all of the programs, so it was necessary to consider pooling effects across test intervals. However, analyses were first conducted to determine if effects decreased or increased with time. Three statistical procedures were used. First, a repeated-measures multivariate analysis of variance (MANOVA) was found to be nonsignificant for programs (N = 4) with results in all four time periods. A second repeated-measures MANOVA for programs (N = 12) in the first and the fourth time intervals was also found nonsignificant. Further inspection showed that equal numbers of programs reported increases in effect size over time as those reporting decreases in effect size over time. Third, scatterplots of 118 programs[8] compared each time period with each other. The scatterplots also supported the pooling of effects sizes (for greater detail, see Tobler 1992*a*).

A second aggregation produced a final single effect for a program by averaging the effects for the time intervals reported (Tobler 1994). This method maintains the statistical independence for each program.

Choice of Covariate Adjusted Means

Effect sizes are usually computed on the final unadjusted posttest results (Glass et al. 1981; Smith et al. 1980). Unadjusted means can only be used when random assignment resulted in truly equivalent treatment and control groups. Undoing the covariate adjusted scores to obtain the unadjusted means, as proposed by Smith and colleagues (1980) and Glass and colleagues (1981), would remove all the control built into the data analysis to correct for the problem of preexisting differences. In fact, the best-designed programs that initially blocked on preexisting drug use would be penalized the most. As the purpose of meta-analysis is to show program effects, not preexisting differences, the program effect sizes were computed from the covariate adjusted means reported by the researcher. Also, including quasi-experimental (nonrandom assignment) studies necessitates working with change scores; an assumption of no preexisting differences between groups at pretest cannot be made. Additionally, the unit of random assignment for experimental programs was intact social units, either classrooms (27 percent) or schools (53 percent), rather than individuals (27 percent). Only 43 percent of those studies randomly assigning intact units had more than six experimental and six control units, which leaves preexisting differences a major problem. As a final consideration, test-retest reliabilities are needed to compute unadjusted posttest scores whether analysis of covariance summary statistics or pretest/posttest means and SDs are available for effect size computations. Test-retest values were not reported in 81 percent of the studies in Tobler (1992*a*). Convention rules for estimating test-retest reliabilities were developed by Smith and colleagues (1980); however, these are gross estimates, either underestimating or overestimating the actual effect size.

Windsorizing

Based on a precedent set by Lipsey (1992) in a meta-analysis of juvenile delinquency treatment, a decision was made to windsorize the weighting factor. This was accomplished by limiting the weighting factor for the larger programs to a maximum and increasing the weighting factor of the smaller programs. This decision was necessary as the sample of students in a program varied from 20 to about 6,000. The weighting factor is the inverse of the variance, which is approximately four times smaller than the number of participants in the program. Twenty-one programs had weighting factors under 25 (100 tested students or less), while six programs had weighting factors near or above 1,000 (i.e., 4,000 tested students). Without windsorizing, the largest programs would be given 40 times the weight of the smaller programs, allowing one large study to completely overshadow the results of the smaller programs. To reduce the 40:1 ratio to a more reasonable 8:1 ratio, the weighting factors under 30 were windsorized up to 30 and the larger programs over 250 were limited to 250. The number present at each test was used to determine the weighting factor.

Other Decisions

When frequencies, proportions, or percentages were the only data reported, probit transformations (Cohen and Cohen 1983) were used to compute the effect size. The use of probit transformations with change scores is discussed elsewhere (Tobler 1985). Where parametric statistics were reported, the effect sizes were calculated using documented formulas (Tobler 1992*a*, appendix 3). When reports stated that results were significant, a 0.05 level of significance was assumed and the corresponding t levels computed. If only a statement of nonsignificance was reported, a p value of 0.50 was assigned (i.e., an effect size of zero). This is a conservative method for estimation of effect sizes. Had researchers given the actual p value, even though not significant, it would lead to an effect size greater than zero.

INDEPENDENT VARIABLE—TYPE OF PROGRAM

Program Content

Program content was coded for 30 items and collapsed into 7 major domains: knowledge, affective, drug refusal skills, generic skills, safety skills, extracurricular activities, and others (see table 1). The content items were coded either yes or no; therefore, the coding scheme did not reflect the relative time spent on a particular content area. This necessitated a subjective decision about the amount of emphasis placed on a particular content item before categorization.

Group Process

The methods and techniques used to deliver the program content have been given little emphasis in the review literature. The terms "interactive" and "noninteractive" were chosen to emphasize what actually happened in the classroom. The type of group process or delivery method was incorporated into the definition of the type of program in the 1993 scheme. "When we observe how a group is handling its communication, i.e., who talks how much or who talks to whom, we are focusing on group process" (Edwards 1972, pp. 182-183). As drug prevention programs are carried out with either the whole class or in smaller groups, a group classification system based upon Toseland and Rivas's topology (1984, pp. 20-22) was specifically revised to describe the classroom processes operating in school-based drug prevention programs (table 2). Four types of groups (A to D) were identified and were ranked on a continuum, beginning with group A (table 2, left column) which had little or no adolescent interaction (i.e., didactic presentations). Each group included progressively greater degrees of interaction among the group members, with group D being the most interactive.

Determination of the Type of Program

Once the decisions about the content and type of group were made, the two dimensions were combined to determine the type of program. Using items listed in table 3, a decision was made about which combination of the five major content domains (knowledge, affective, refusal skills, generic skills, and safety skills[9]) best portrayed the program content. The second choice was the type of group (right column, table 3). The overall context of the entire program was taken into consideration before making a final determination about the type of program. For example, for Drug

TABLE 1. *Major content in adolescent drug prevention programs.*

KNOWLEDGE

Knowledge of drug effects
Knowledge of media and social influences
Knowledge of actual drug use by peers (normative education)

AFFECTIVE

Self-esteem and feelings
Personal insight and self-awareness
Attitudes, beliefs, and values

REFUSAL SKILLS

Drug-related refusal skills
Public commitment activities
Cognitive behavioral skills
Support systems/networking with nondrug-using adolescents

GENERIC SKILLS

Communication skills
Assertiveness skills
Decisions/problemsolving skills
Coping skills
Social/dating skills
Goal-setting
Identifying alternatives

SAFETY SKILLS

Skills to protect self in a drug-related situation
Skills to protect other peers in a drug-related situation
Drinking/driving safety

EXTRACURRICULAR ACTIVITIES

Paid job activities or training
Organized sports
Organized cultural activities
Nondrug leisure time activities
Volunteer work in the community

OTHER

Peer counseling/facilitating/helping
Homework exercises
Rewards, token economy, and reinforcement
Parent involvement
Communitywide coordination and involvement

SOURCE: Reprinted with permission: Tobler, N. Updated meta-analysis of adolescent drug prevention programs. In: Montoya, C.; Ringwalt, C.; Ryan, B; and Zimmerman, eds. *Evaluating School-Linked Prevention Strategies: Alcohol, Tobacco and Other Drugs*. San Diego: USCD Extension, University of California, 1993. pp. 71-86.

TABLE 2. *Four group types.*

	Group A	Group B	Group C	Group D
Aim	To educate: knowledge gain	To educate: intrapersonal competence; self-awareness, self-esteem building, feelings, values, "affective education"	To develop: interpersonal skills; relationships with others; increase feeling of acceptance through positive peer interactions	To develop: intrapersonal and/or interpersonal growth.
Purpose	Learning through didactic presentation	Some didactic presentations; group discussions; individually oriented experiential activities	To increase communications and social skills; improve interpersonal relationships through structured exercises, role plays; and interpersonal experiential activities	To identify member's potentials; self-awareness, insight, and interpersonal development through discussion and dynamic group process
Leadership	Leader as teacher	Leader as teacher; provider of structure for group	Leader as facilitator of the group’s activities; provide structure	Leader as facilitator and role model; group members take responsibility for group's direction
Focus	Individual knowledge	Individual growth	Focus on group as a medium for interaction; involvement of all individuals	Either member or group focus; individual growth through the group experience
Structure	Highly structured; passive participation	Structured; passive and some active participation	Structured; active participation	Limited structure; open ended, active participation

TABLE 2. *Four group types (continued).*

	Group A	Group B	Group C	Group D
Bond	None	Common interest in learning; skills development; bond limited	A common activity, enterprise or situation; bond between members	Common goals among members; contract to use the group to grow; bond between members
Composition	Typical school class	Similarity of educational or skill level	Can be diverse or homogeneous	Can be quite diverse. Based on members' ability to work towards growth and development
Communication patterns	Leader to student	Leader to member, didactic; sometimes member to member during discussions; self-disclosure low	Member to member; often represented in activity and nonverbal behavior; self-disclosure low to moderate	Highly interactive; members often take responsibility for communication in the groups; self-disclosure moderate to high

SOURCE:Reprinted with permission: Tobler, N. Updated meta-analysis of adolescent drug prevention programs. In: Montoya, C.; Ringwalt, C.; Ryan, B.; and Zimmerman, R., eds. *Evaluating School-Linked Prevention Strategies: Alcohol, Tobacco and Other Drugs.* San Diego: USCD Extension, University of California, 1993. pp. 71-86.

Awareness Resistance Education programs (Project DARE), a determination was made that relatively more emphasis was placed on intrapersonal content in the affective and generic domains than on the interpersonal skills such as drug refusal (Ringwalt et al. 1990). Project DARE's content would be coded knowledge, affective, some refusal skills, and some generic skills, and was classified as knowledge-plus-affective because the emphasis on refusal and interpersonal skills was limited. The second choice, the group process, also placed Project DARE under knowledge-plus-affective as it was most typically delivered in a noninteractive group B setting. Twenty-six distinct types of programs were identified, consolidated into the six major

subcategories, and further collapsed into two overarching categories: noninteractive programs and
interactive programs.

DEPENDENT VARIABLE

The drug use outcome measures were paper-and-pencil self-reports given confidentially in a classroom setting and were often accompanied by physical tests (i.e., saliva). The reliability of confidential self-reports of cigarette use has been documented (Murray et al. 1987; O'Malley et al. 1983; Pechacek et al. 1984) as have measures of other illicit and licit drug use (Oetting and Beauvias 1990; Single et al. 1975).

DATA ANALYSIS

Ordinary and Weighted Least Squares Regression

Ordinary least squares (OLS) regression analyses were used for the unweighted effect size. For the weighted effect size, weighted least squares (WLS) regression analyses were conducted as detailed in Hedges and Olkin (1985). This procedure weights each program effect size by the sample size of that program. The significance testing is conducted at the program level (SPSS 1990).

To account for the differences in the effectiveness of a type of program, other variables related to program success must be considered. For example, recent smoking programs have been highly successful and the possibility exists that their success is the result of targeting cigarettes and not the type of program used. Multiple regression procedures make available methods for computing the unconfounded effect for the type of program by partialing out the effect of all the covariates (i.e., holding constant the effect of the covariates). A discussion of each covariate is included in the following sections.

Dummy Coding for Categorical Variables

In the present analyses, the dependent variable (effect size) and one covariate (sample size) are continuous variables. The remaining six predictor variables are categorical. The independent variable, type of program, is categorical, as are the five covariates: type of control group, experimental design, special populations, targeted drug(s), and

TABLE 3. *Type of program by content and process.*

Content	Process
NONINTERACTIVE: KNOWLEDGE ONLY	
Knowledge	Group A
Knowledge	Film/theater
Knowledge + Attitudes	Group A
Drinking + Driving	Group A
Drinking + Driving	Scare tactics
AFFECTIVE ONLY	
Affective	Group B *ECM
Affective	Group B
KNOWLEDGE PLUS AFFECTIVE	
Knowledge + Affective	Group B
Knowledge + Affective + Attitudes + Values	Group B
Knowledge + Affective + Decisions	Group B
Knowledge + Affective + Generic	Group B
Knowledge + Affective + Refusal + Generic	Group B
Knowledge + Affective + Generic + Community	Group B
Drinking + Driving	Group B
INTERACTIVE: SOCIAL INFLUENCES	
Knowledge + Refusal	Group C
Knowledge + Refusal + Community**	Group C
Drinking + Driving	Group C
COMPREHENSIVE LIFE SKILLS	
Knowledge + Refusal + Generic	Group C
Knowledge + Refusal + Generic + Community**	Group C
Drinking + Driving	Group C
OTHERS	
Knowledge + Norm-changing	Group C
Knowledge + Affective	Group C
Knowledge + Affective	Group D
Knowledge + Affective + Generic	Group C
Knowledge + Affective + Refusal + Generic	Group D

KEY: * = Effective classroom management for teachers; ** = Total community effort supporting the school-based program.

SOURCE:Reprinted with permission: Tobler, N. Updated meta-analysis of adolescent drug prevention programs. In: Montoya, C.; Ringwalt, C.; Ryan, B.; and Zimmerman, R., eds. *Evaluating School-Linked Prevention Strategies: Alcohol, Tobacco and Other Drugs.* San Diego: UCSD Extension, University of California, 1993. pp. 71-86.

leaders. The type of program or independent variable is comprised of two clusters of programs: noninteractive and interactive. Therefore, it was dummy coded, 1 or 0, to identify group membership. Three other covariates were comprised of binary clusters: type of control group, experimental design, and special populations. Two covariate variables were comprised of a cluster of more than two dummy variables. For example, leaders consisted of a cluster of four different types of leaders: teachers, same age or older age peer leaders, mental health professionals, and all others. Teachers were designated as the reference group and were coded 0, 0, 0. The peer leaders were coded 1, 0, 0; mental health professionals were coded 0, 1, 0; and all others were coded 0, 0, 1. In dummy coding, the degrees of freedom for a variable are K-1; therefore, a binary variable uses one degree of freedom. Three degrees of freedom are used for the leaders variable, which is composed of a cluster of four dummy variables.

Regression Equation

To examine the effects due to the primary independent variable (type of program) without the confounding effects of the covariates, it was necessary to remove the proportion of variance attributed by each covariate. Each of the covariate clusters was entered into the regression equation before the primary independent variable. The sequence of entry for the covariates was arbitrary as no order was hypothesized. The effects of the six confounding covariates were removed before computing the covariate adjusted means for two types of programs.

To keep the number of parameters in line with the number of cases, interactions were not included. Partial confirmation for this is given by the fact that the two-way analyses of variance (ANOVAs) for each covariate with the primary independent variable had no significant second-order interaction effects. Finally, the OLS residuals were examined for outliers. Six outliers were identified and removed, leaving a sample of 114 programs.

Of interest is the extent that a covariate accounts for program success. It is important to answer questions such as, "Which is more highly associated with program success, the type of program or the drug targeted by the program?" The increment to R^2, which is the proportion of variance accounted for by a covariate, can be used to determine the relative impor-tance of a variable for predicting program efficacy. No attempt was made to independently analyze any of the levels within the categorical covariates. For the primary independent variable, the magnitude of the change in R^2 can be determined when this variable is entered into an equation that already contains the covariates (i.e., partialing out the effect of all the covariates).

Hypothesized Covariates Omitted

The variables identified as potent predictors of program success were chosen based on previous research (Tobler 1986) and a review of the literature. Sex, initial level of drug use, booster sessions, implementation factors, and the research center were all eliminated as covariates because only a limited number of programs reported results broken out for this information (frequencies are reported in Tobler 1992*a*).

Two additional hypothesized variables, grade and program intensity, were eliminated based on the analyses reported in Tobler (1992*a*). Each variable was nonsignificant in all 16 regression analyses and contributed R^2 increments of less than 2 percent.

Six Covariates Included

Sample Size. The effect sizes for the programs with large sample sizes were found to be smaller (Tobler 1992*a*); therefore, the weighting factor, which is an approximate estimate of the sample size, was entered as a continuous variable.

Type of Control Group. Treatments compared to a no-treatment control group were found to have higher effect sizes than those compared to a standard health curriculum/another treatment (Tobler 1986, 1992*a*). The reference category was treatments compared to a health class control.

Experimental Design. A categorical variable was made for studies that had acceptable attrition (with or without differential dropout) and unacceptable attrition (with or without differential dropout). The reference category was acceptable attrition. This binary variable was derived from the empirical findings reported (Tobler 1992*a*). A decision tree was used which involved three choices: assignment, attrition, and differential dropout.

The results showed that no differences in effect sizes were observed for random (0.17) versus nonrandom (0.16) assignment. Whether differential dropout occurred from treatment or control was missing in 61.7 percent of the reports (Tobler 1992*a*); these studies were grouped with those reporting differential dropout (a conservative method). As a result of the complex empirical results for experimental design, it was decided that the experimental design was best represented by the binary variable of acceptable and unacceptable attrition. The retention rates for school-based drug prevention studies were compiled as part of a meta-analysis of 85 longitudinally followed cohorts (Hansen et al. 1990). This data provided normative attrition rates for drug prevention research. Attrition was coded as acceptable if it was on the mean or above (12 months from pretest) and unacceptable if below the mean.

Special Populations. The literature reports that most research has been conducted primarily in schools with > 50 percent white populations. In Tobler (1992*a*), schools with > 50 percent minority or problem students were found significantly more successful than those with > 50 percent white populations in a number of regressions for the 114 programs. The reference category was schools with > 50 percent white populations.

Targeted Drug. Three categories existed for this dummy variable: smoking programs, alcohol programs, and substance abuse and/or generic drug prevention programs. The generic drug programs have outcome measures for cigarettes, alcohol, marijuana, and all other drugs. Therefore, the effect size must be seen as an average of the results for all drugs tested, whereas smoking and alcohol programs tested a single drug. It was not possible to examine the results for a single drug in the generic programs and still use the study effect method (one effect per program). The reference group was smoking programs.

Leaders. Four categories of leaders were entered for this block: teachers, peer leaders, mental health specialists,[10] and others (e.g., research staff, health educators, and various outside professionals). The reference group was teachers.

RESULTS

As the aim of this chapter is to connect the descriptive statistics and the qualitative information to the inferential analyses, this section briefly discusses the inferential statistics.[11] The mean effect sizes are presented for only the unadjusted means as the focus is descriptive. It should be noted that the difference between the unadjusted mean effect sizes and the covariate adjusted mean effect sizes was very small (approximately 0.01). The results are reported for both UNES and WES. The WES is meta-analytically sound as the effect size has been weighted for the sample size. The UNES is reported even though it is considered meta-analytically unsound, because it provides a way to examine the results for the smaller programs without their being overpowered by the larger programs. Two problems specific to drug prevention program research makes this necessary: the wide range in sample sizes (20 to 4,000 tested youth), and the limited number of evaluated programs, which precludes separate meta-analyses for the smaller programs and larger programs.

Nature of Programs Located

The strength of a meta-analysis depends on the comprehensiveness of the sample of programs. Extensive search procedures located 595 studies of adolescent prevention programs. The 120 programs that passed the selection criteria came from 90 research studies: single programs were reported in 69 studies; more than one type of program was reported in 21 studies.[12]

Five hundred and five studies (84.9 percent) did not pass the selection criteria. Eighty-three studies (14.1 percent) that were evaluated lacked a control or comparison group; 113 studies (19.2 percent) were descriptive in nature with subjective conclusions. Another group of 30 (5.1 percent) studies was excluded after passing the original selection criteria. These 30 studies initially appeared to have an experimental or quasi-experimental research design, but on further reading it was found that they were not implemented according to the original plan and/or they lacked the necessary statistical data to calculate an effect size.

Other studies were excluded as follows: 132 were not school based and/or did not target high-risk youth[13]; 45 evaluated grades under fifth or college students; 30 did not have drug use outcome measures; 21 were published earlier than 1978; 13 were published before 1978 and also did not have drug use measures; 17 were not implemented in the United States or Canada; 2 evaluated only the teachers; and finally, time did not allow the inclusion of 19 eligible studies.

The 143 drug prevention programs in the 1986 meta-analysis were included in the tally above. Only 39 programs passed the new selection criteria. Because comparisons will be made between the 1986 and 1993 meta-analyses, the reasons for excluding 104 of the 143 programs are also listed: 13 did not meet the more stringent research design qualifications,
13 lacked drug use measures, 29 were published before 1978, 24 were published before 1978 and also lacked drug use measures,
22 targeted high-risk youth and/or were community based, 2 were eliminated due to insufficient time, and 1 report was a duplicate of the same group (i.e., second posttest).

Effect Size by Type of Program

The 70 interactive school-based adolescent drug prevention programs were effective in changing adolescent drug use behaviors (UNES = 0.247, 0.95 confidence interval (CI) = 0.18 to

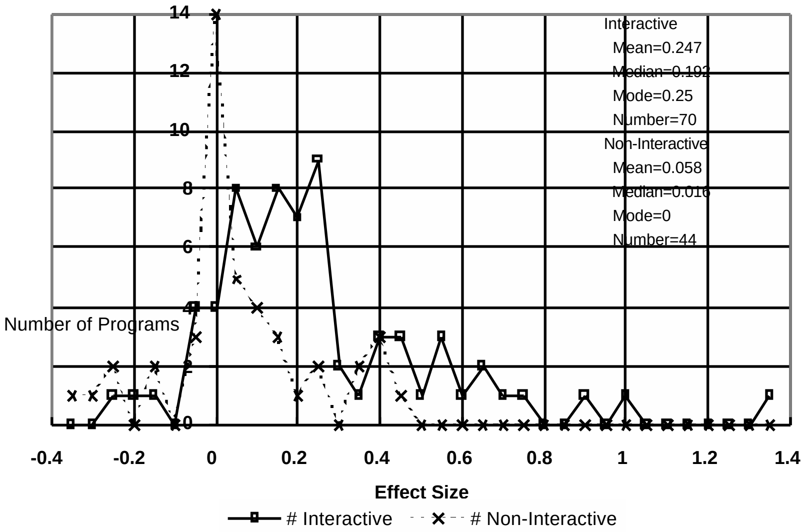
31; WES = 0.164, 0.95 CI = 0.14 to 18) while the 44 noninteractive programs were essentially ineffective (UNES = 0.058, 0.95 CI = 0.00 to 0.11, WES = 0.075, 0.95 CI = 0.05 to 0.10). A further analysis of the subset of 56 high-quality experimental programs included in the larger set of 114 programs[14] showed even higher effect sizes for the interactive programs (UNES = 0.317, 0.95 CI = 0.22 to 0.41; WES = 0.214, 0.95 CI = 0.19 to 0.24; N = 38) and still lower effect sizes for the noninteractive programs (UNES = 0.017, 0.95 CI = -0.07 to 0.11; WES = 0.043, 0.95 CI = 0.00 to 0.09; N = 18). In all four regressions, the interactive programs were significantly better than the noninteractive programs (P = 0.002 for the unweighted OLS for 114 programs; P = 0.001 for the unweighted OLS for 56 programs; P = 0.009 for the weighted WLS for 114 programs; and P = 0.015 for the weighted WLS for 56 programs).

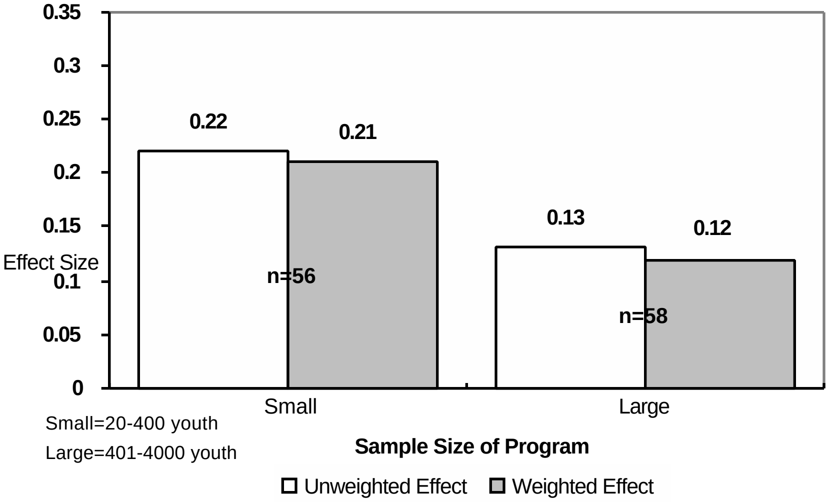
Effect Size Distribution by Noninteractive and Interactive Programs

Figure 1 gives the frequency distribution of the UNES for the two types of programs. The distributions for the interactive and noninteractive programs are shown separately and are strikingly different. The 44 noninteractive programs have a mean of 0.058, a median of 0.016, a mode of zero, a range of -0.35 to 0.45, and 0.95 CIs of 0.00 to 0.11. This stands in contrast to the 70 interactive programs which have a mean of 0.247, a median of 0.192, a mode of 0.25, a range of -0.24 to 1.34 and 0.95 CIs of 0.18 to 0.31. When the two separate distributions are compared to the combined distribution for all 114 programs, it can be seen that the noninteractive programs were responsible for the mode, while the interactive programs contribute more to the positive skew.

Effect Size by Sample Size

There is a limitation, however. All the large scale implementations (i.e., 400 to 4,000 tested youth), whether interactive or noninteractive programs, exhibited a leveling of effectiveness (see figure 2). The smaller programs (i.e., 20 to 400 tested youth) had an UNES of 0.22 (0.95 CI = 0.14 to 0.31) and a WES of 0.21 (0.95 CI = 0.18 to 0.25). The large scale programs had an UNES of 0.13 (0.95 CI = 0.08 to 0.17) and a WES of 0.12 (0.95 CI = 0.10 to 0.14). This is a ds of 0.09.


Interactive
Mean=0.247
Median=0.192
Mode=0.25
Number=70
Non-Interactive
Mean=0.058
Median=0.016
Mode=0
Number=44
Number of Programs
14
12
10
8
6
4
2
0
-0.4
-0.2
0
0.2
0.4
0.6
0.8
1
1.2
1.4
Effect Size
# Interactive
# Non-Interactive


0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Effect Size
0.22
0.21
0.13
0.12
n=56
n=58
Small
Large
Small=20-400 youth
Large=401-4000 youth
Sample Size of Program
Unweighted Effect
Weighted Effect

Effect Size by Posttest Time Interval

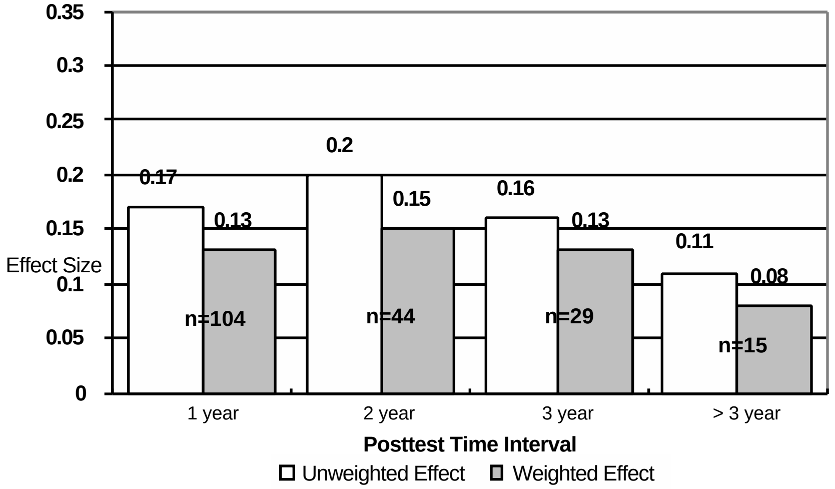
The results at each of the four test intervals (1 year, 2 years, 3 years, and > 3 years) are given in figure 3. The magnitude of the effect sizes were maintained over the first 3 years and showed a slight decrease for the fourth interval (greater than 3 years). Most probably, the minor variations across time were due to the different sets of programs included in each interval. No single time interval included all the programs. Ninety percent of the programs reported test results within the first year; a sharp drop was observed for the second year, with only 34 percent of programs represented; 25 percent reported third year results; and only 15 percent took posttests at a period greater than 3 years. The total number does not equal 114 because many programs administered multiple posttests.

To alleviate concerns about decay of program effectiveness over time, the four OLS and WLS regressions were rerun using the first posttest results as the dependent measure regardless of the length of time from pretest. To control for effectiveness decay over time, the length of time from pretest to the first posttest was entered as an additional continuous covariate along with the original six covariates.

The results of OLS and WLS regression based on the first posttest were almost identical to those based on the average effect size across time intervals. Using the first posttest results, the interactive programs were significantly better than the noninteractive programs: P = 0.003 (first posttest) versus P = 0.002 (across time) for the unweighted OLS for 114 programs; P = 0.005 (first posttest) versus 0.009 (across time) for the weighted WLS for 114 programs; P = 0.002 (first posttest) versus P = 0.001 (across time) for the unweighted OLS for 56 programs; and P = 0.015 (first posttest) versus 0.015 (across time) for the weighted WLS for 56 programs. The effect size across time and the effect size for the first posttest showed identical patterns for the interactive and noninteractive programs. The only noteworthy observation was that the noninteractive programs were underrepresented for the second year, third year, and the fourth interval (greater than 3 years).

Effectiveness of the Six Major Subgroups

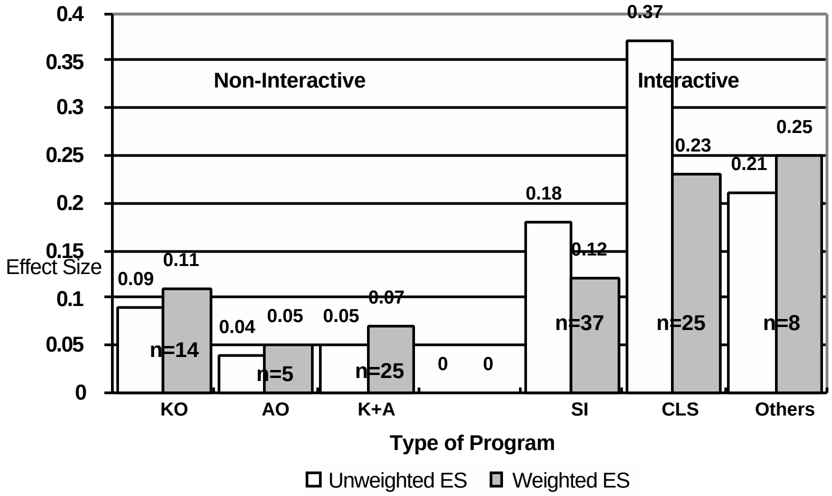
Set of 114 Programs. The noninteractive programs included three subgroups: knowledge only (KO), affective only (AO), and knowledge-plus-affective (K+A) programs (see figure 4). The three interactive programs subgroups were social influences (SI), comprehensive life skills



(CLS), and others (see right side of figure 4). In the set of 114 programs, the highest effect sizes were obtained by the CLS programs. The others programs had the second highest effect sizes, but also the largest confidence interval. The third most effective subgroup was the SI programs, which also had the tightest confidence intervals.

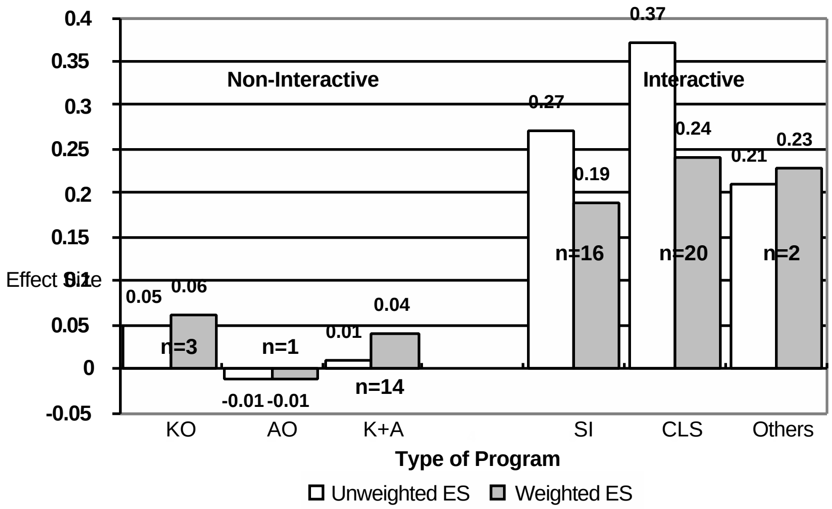
All three subgroups of interactive programs had higher effect sizes than the three subgroups of noninteractive programs.

Subset of 56 Programs. Differences were observed when comparing the 114 programs (figure 4) with the subset of 56 programs (figure 5). For the interactive programs, the effect sizes for the SI programs were substantially higher in the subset of 56 programs than in the 114 programs, while the CLS and others programs remained about the same. The three noninteractive programs showed lower effect sizes for the well-controlled programs (subset of 56).



One Type of Program Eliminated: Drinking/Driving

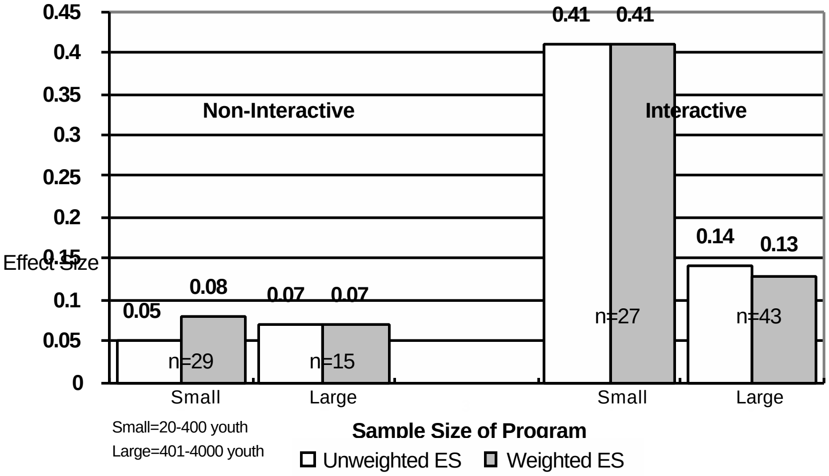
It was hypothesized that drinking/driving programs should constitute a major type of program based solely on the program content (Goodstadt, personal communication, November 10, 1989). These programs emphasize an individual's responsibility in alcohol-related accidents or deaths, personal safety relative to driving with someone who has been drinking, and responsibility for providing safe transportation for friends who have been drinking (e.g., designated drivers). Only 10 drinking/driving programs were located. Within this subgroup, there were five different types of programs: KO, KO with fear tactics, K+A, SI , and CLS programs. Surprisingly, the effect sizes ranged from -0.18 to 0.30, showing extreme heterogeneity; this indicates that other factors were operating besides program content. When reclassified by the program delivery method (group A to D), the effect sizes matched those of programs having similar delivery methods (i.e., same types of groups). The limited number of drinking/driving programs precludes any definitive analyses, but the fact that the empirical results show that the type of group is more important than the content in determining the type of program validates the inclusion of the type of group process in the classification scheme.



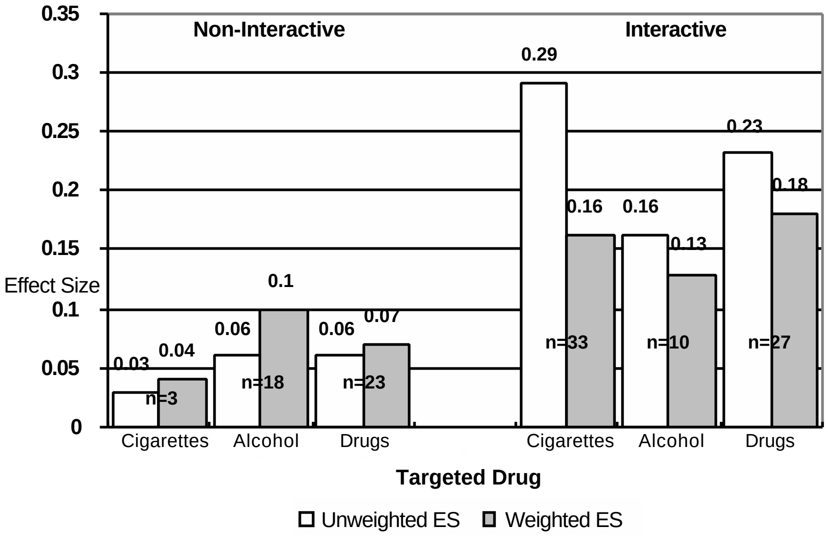
Six Covariates by Type of Program

The UNES and WES were presented for only the larger set of 114 programs in order to maintain enough programs in a category. Without knowing the distribution for each of the categories, the interpretation of a category mean with a limited number of programs becomes very tenuous. For example, if 9 programs show small but consistent positive effects while the 10th program is highly negative, the resulting mean may be zero or even slightly negative. Therefore, the results for the set of 56 programs are described only when they differ from the 114 programs.

Sample Size. The sample size, a continuous covariate, was found significant in all four regressions reported in Tobler (1994), as was the independent variable, type of program. However, the proportion of variance accounted for by sample size was lower than the independent variable, but much higher than any of the other five covariates. When broken into groups of small or large size programs within the noninteractive and interactive categories, the small interactive programs were extremely successful (figure 6). Both their UNES and WES were 0.41, whereas the small noninteractive programs achieved only an UNES of 0.05 and WES of 0.08. A large drop in effectiveness was observed when the small, highly successful interactive programs were implemented on a larger scale. Still, the large interactive programs were twice as successful as large noninteractive programs. (Notice that the UNES and WES were nearly equivalent in each of the four groups.) For the remaining covariates, the WES will be much lower, reflecting the loss of effectiveness when implemented on a large scale.

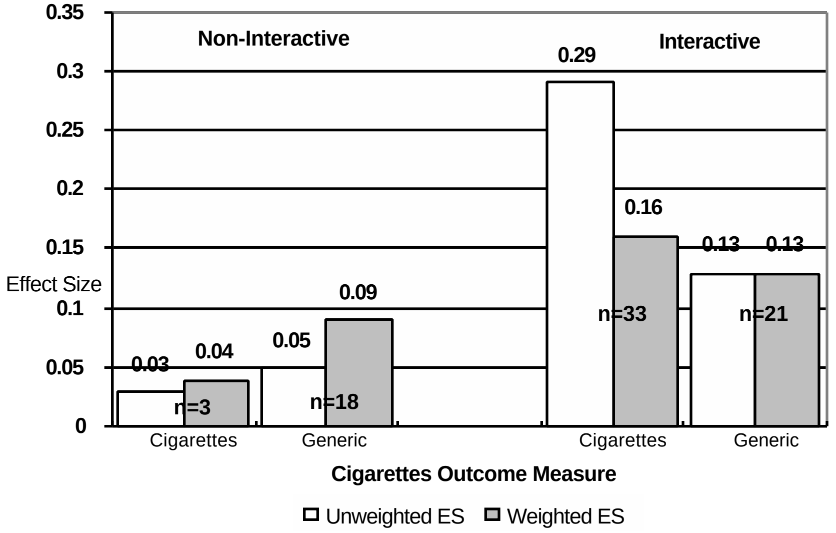


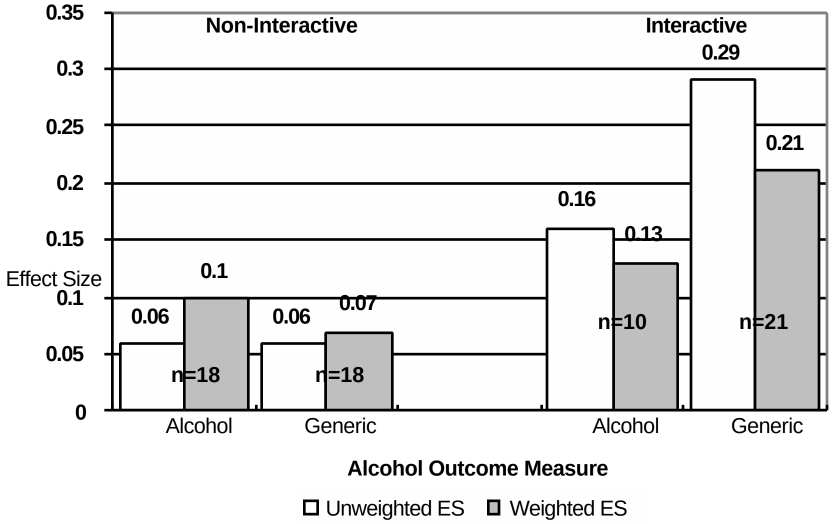
Targeted Drug. Figure 7 shows the results of smoking, alcohol, and generic drug prevention programs by type of program. This covariate was significant in both of the unweighted OLS regressions (56 and 114 programs), which indicates that the smaller smoking programs were more effective than the larger generic drug prevention programs. Relative to the size of the programs, the interactive smoking programs had an UNES almost twice as high as its WES. Relative to targeting a drug, the interactive programs had an UNES for the smoking programs that was larger than either the alcohol or the generic drug programs; all the WESs were about the same. In other words, the smaller programs were more effective in all cases, although the differentials between the UNES and WES were not as large for the alcohol programs or the generic drug programs. For the noninteractive programs by program size, a smaller differential was observed between the UNES and WES and it was in the opposite direction. The differences between the smoking, alcohol, and generic programs were minimal.



Generic Drug Programs. The interactive programs were nearly four times as effective as the noninteractive programs for UNES size and about three times as effective for WES. Note that generic drug programs tested for cigarettes, alcohol, marijuana, and all other illicit drugs; a composite score was necessary to maintain one effect size per program. Furthermore, no independent analysis was made of the levels of dummy coded variables within each categorical covariate (i.e., between smoking, alcohol, and generic programs). Therefore, further analyses were necessary to determine the effectiveness of the generic drug programs with cigarettes and with alcohol. To accomplish this, the cigarette score was extracted from the composite score (i.e., mean for cigarettes, alcohol, marijuana, and other drugs) reported by the generic programs and then compared to the results for the smoking programs (figure 8). A similar procedure was used for alcohol (figure 9).

Smoking Programs. Whether a smoking program or a drug prevention program, the interactive programs were significantly superior to the non-interactive programs (Q_B = 7.95, Q_{cr} = 3.841) in reducing cigarette use. Therefore, any further comparisons must be made within the categories of interactive or noninteractive programs. Within the 54 interactive programs, the 33 smoking programs were not significantly superior to the 21 generic drug prevention programs (Q_B = 1.62, Q_{cr} = 3.841). The small interactive smoking programs were extremely beneficial, as seen in the UNES magnitude (figure 15). However, the WES for the interactive smoking programs was not much better than the interactive generic programs, although both were higher than the noninteractive programs. The reverse was true for the generic noninteractive programs, which had a higher WES than UNES, but only three noninteractive programs targeted cigarettes. However, these measurements were performed on the set of 114 programs with all their inherent problems; therefore, to validate these results a further analysis was made using the 56 high-quality experimental programs.[15]





Unfortunately, only 38 programs tested cigarettes in the subset of 56 programs. The interactive programs had a sufficient number of programs (i.e., 14 smoking and 15 generic) for comparison; the noninteractive programs had only 1 smoking and 8 generic programs. The interactive smoking programs were highly successful and significantly superior to the interactive generic programs (Q_B = 33.38, Q_{cr} = 3.841). The effect sizes for the smoking programs (UNES =

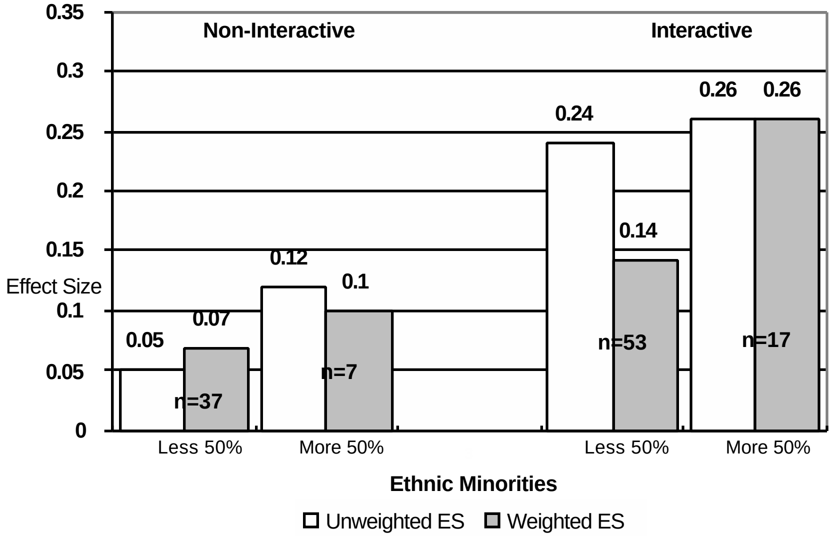
0.48, WES = 0.32) were much higher than those obtained for cigarettes in the generic programs (UNES = 0.11, WES = 0.12). The higher UNES versus WES for the smoking programs indicates that the smaller smoking programs were more successful than the larger smoking programs. Apparently the size of the program was not a factor for the generic drug programs, as the UNES and WES were almost equivalent.

Alcohol Programs. The interactive programs were significantly superior to the noninteractive programs (Q_B = 23.42, Q_{cr} = 3.841) for alcohol use based on the alcohol outcome measure for the set of 114 programs (figure 9). For alcohol programs, it should be noted that only 32 percent of the interactive programs targeted alcohol versus 50 percent of the noninteractive programs. Within the noninteractive programs there were no significant differences between programs that targeted alcohol and the generic approaches (Q_B = 0.98, Q_{cr} = 3.841); the interactive programs showed that generic programs were slightly superior to alcohol programs (Q_B = 4.670, Q_{cr} = 3.841).

Again, because of the potential sources of bias in the set of 114 programs, the subset of 56 high-quality experimental programs was examined. The interactive programs were superior to the noninteractive programs (Q_B = 20.6, Q_{cr} = 3.841). Only one of the four subcategories, interactive generic programs, was large enough for reliable conclusions; the noninteractive alcohol, the noninteractive generic, and the interactive alcohol programs had less than 10 cases, which makes these categories vulnerable to spurious findings. Surprisingly, the generic interactive programs achieved an effect size for alcohol (UNES = 0.29, WES = 0.21, N = 15) that was approximately twice as high as the results for cigarettes using generic interactive programs.

Special Populations. The F change for schools having greater than 50 percent ethnic minorities was statistically significant in only one regression: WLS regression for 114 programs (P = 0.009). The UNES and WES were identical (0.26) for the interactive programs in schools having greater than 50 percent minorities (figure 10); this similarity indicates that the larger programs produced results equivalent to the smaller programs. Within the interactive programs, the opposite was true for schools with less than 50 percent minorities. A much lower WES (difference of 0.10) than UNES indicates that when the interactive programs were implemented with a white student population on a large scale, they did not do as well. The noninteractive programs were also slightly more successful in schools with greater than 50 percent ethnic minorities.

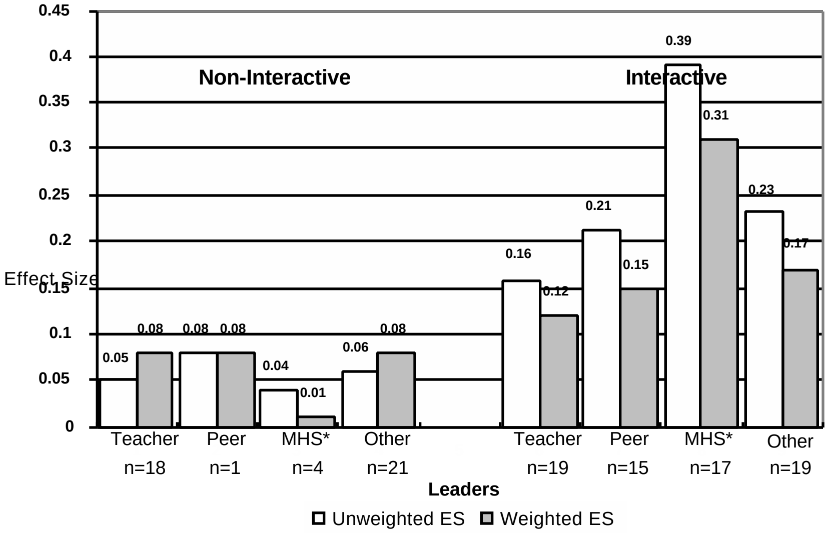
Leaders. No statistically significant findings were obtained for the F change for leaders in any of the four regressions, and the increment to R^2 was below 2 percent. For the interactive programs, mental health specialists were the most effective and teachers were the least effective (figure 11). Same age/older age peer leaders and other professionals were slightly more successful than the teachers within the interactive programs. The pattern was the same within the noninteractive programs with the exception of mental health specialists, who were the least successful leaders (N = 4). A degree of confidence can be placed in these findings because the four types of leaders were almost equally represented within the interactive programs, providing an excellent opportunity for comparison.



Experimental Design. The F change for experimental design was not significant in any of the four regressions. Unacceptable attrition was only slightly related to smaller effect sizes within the interactive programs, while it accounted for very little difference in the noninteractive programs (figure 12).

Type of Control/Comparison Group

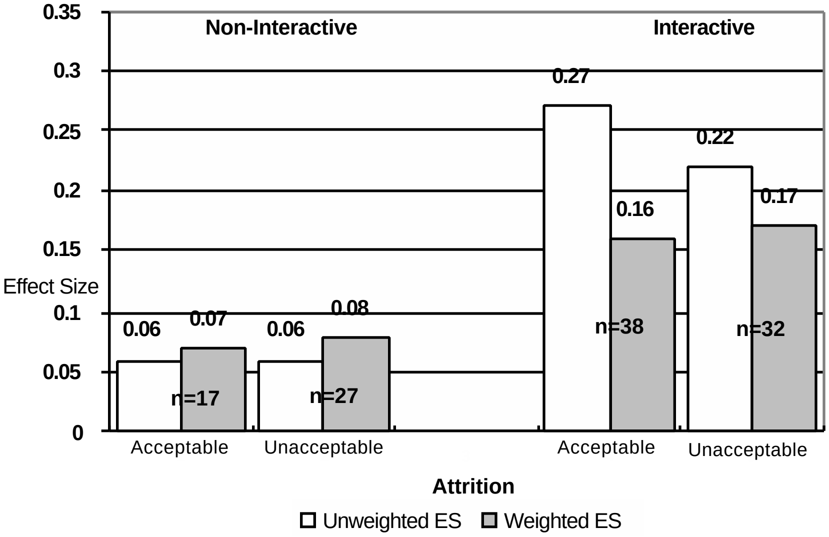
The F change for the type of control group did not reach significance in any of the regression equations, most probably because the mean differences (approximately 0.08) between the health class control and the no-treatment control were not large enough (figure 13). Examining the differences within the noninteractive and interactive programs shows equivalent mean differences between the two different types of control/comparison groups. Also, the two types of control/comparison groups occurred with the same frequency in both the noninteractive and interactive programs.



Content and Process Components by Type of Program

The frequencies for the most prevalent content items and the type of group process are shown for the set of high-quality experimental studies (see table 4). Only three of the six major subgroups are included: K+A, SI, and CLS. These groups contained a sufficient number of programs to allow comparisons of their content and process components. The three remaining major subgroups were not included because of the limited numbers: KO (N = 3), AO (N = 1), and others (N = 2). Because the focus of this section is to identify specific content items associated with a major program type, four drinking/driving programs were also excluded as their content was much different.

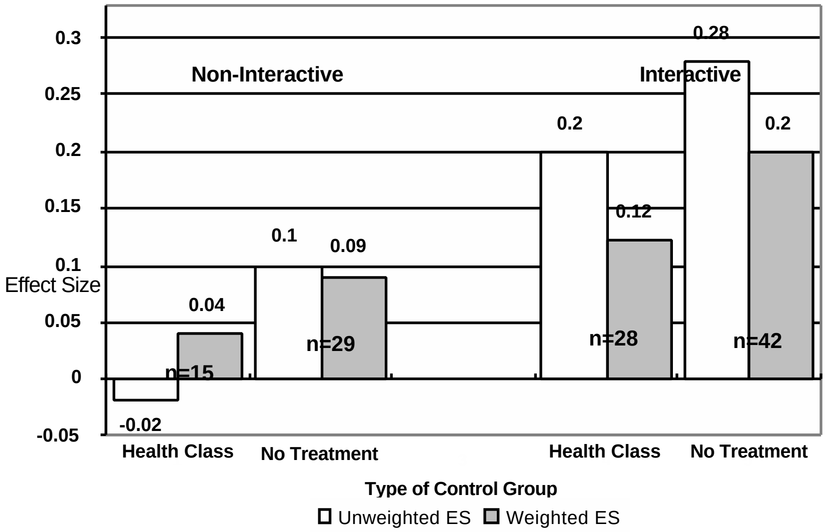
K+A Noninteractive Programs. Within the K+A category, two types of programs were identified and empirically confirmed by cluster analysis procedures. The two subcategories were called values and DARE type programs. Both the values and the DARE type programs include



knowledge about drug effects, a strong emphasis in the affective domain, and also delivery in a noninteractive setting (group B).

Values Programs. Media influences were included in some of the values programs, although none included information about normative expectations. In the affective domain the major emphasis was attitudes and values, followed by insight and self-awareness. None of the values programs included drug refusal skills, although they did emphasize generic decisions/problemsolving skills.

The values programs differed from the DARE type, SI, and CLS programs in the importance placed on insight and self-awareness for the purpose of changing attitudes and values. These components were nonexistent in the DARE type, SI, and CLS programs. Notably, the values programs did not include information about normative expectations or drug refusal skills; these were found in the SI and CLS programs and to some degree in the DARE type programs. Although decision/problemsolving skills were frequently used in the values programs, the focus was intrapersonal, not interpersonal



DARE Type Programs. All five DARE type programs included knowledge about drug effects, media influences, and normative expectations. A different emphasis was placed on the affective content than in the values programs. None of the DARE type programs incorporated self-awareness and insight, but all stressed self-esteem and feelings. Only 40 percent included attitudes and values. Drug refusal skills and decision/problemsolving skills were present in all the DARE type programs, although the amount of time spent on refusal techniques was limited. Another content domain emphasized was generic skills (e.g, assertiveness skills, coping skills).

DARE type program content was closely related to the CLS programs, particularly the inclusion of generic skills, but the DARE type programs did not emphasize drug refusal skills or other interpersonal skills as much as the CLS programs. The DARE type programs placed a greater emphasis on self-esteem than the CLS programs. The most important difference between the DARE type programs and CLS programs was the manner in which the content was delivered. The DARE type programs

TABLE 4. *Contents and process by type of program.*

	Noninteractive		Interactive	
	Values	DARE	SI	CLS
	N = 9	N = 5	N = 13	N = 19
KNOWLEDGE				
Knowledge of drug effects	100%	100%	100%	100%
Media & social influences	44%	100%	69%	63%
Normative expectations	0%	80%	77%	74%
AFFECTIVE				
Self-esteem, feelings	33%	100%	0%	53%
Insight, self-awareness	56%	0%	0%	0%
Attitudes, beliefs, & values	78%	40%	0%	5%
REFUSAL SKILLS				
Drug-related refusal skills	0%	100%	92%*	100%
Public commitment activities	11%	60%	46%	32%
GENERIC SKILLS				
Communication skills	0%	20%	8%	74%
Assertiveness skills	22%	60%	8%	63%
Decisions/problemsolving	78%	100%	8%	95%
Coping skills	11%	60%	8%	74%
Social/dating skills	0%	0%	31%	58%
Goal setting	33%	20%	0%	68%
SAFETY SKILLS				
Skills to protect peers	0%	0%	0%	0%
Drinking/driving safety	11%	0%	0%	0%
PROCESS				
Group A noninteractive	0%	0%	0%	0%
Group B noninteractive	100%	100%	0%	0%
Group C interactive	0%	0%	100%	100%
Group D interactive	0%	0%	0%	0%

KEY: * = Culturally sensitive to Native American population.

used a noninteractive group process versus the interactive group process used by the CLS programs.

SI and CLS Interactive Programs. All the SI and CLS programs include knowledge of drug effects, drug refusal skills, and an interactive setting (group C).

The SI programs were highly focused drug refusal skills programs with only two other content items: media influences and normative expectations. One program did not include refusal skills but had all the other SI components. This modified program was designed to be culturally sensitive for a Native American population. The programs that the Native American adolescents received were atypical and presented classification problems, as was observed in the cluster analyses.

The CLS programs included media influences and normative expectations as frequently as the SI programs. However, as the name implies, CLS programs were more comprehensive. These programs included many generic skills that were not related solely to the use of drugs (i.e., decision/ problemsolving, communication skills, coping skills, goal setting, assertiveness skills, and social skills). In other words, the CLS programs subsumed all the components of the SI programs, added intrapersonal skills, and also included additional nondrug interpersonal skills.

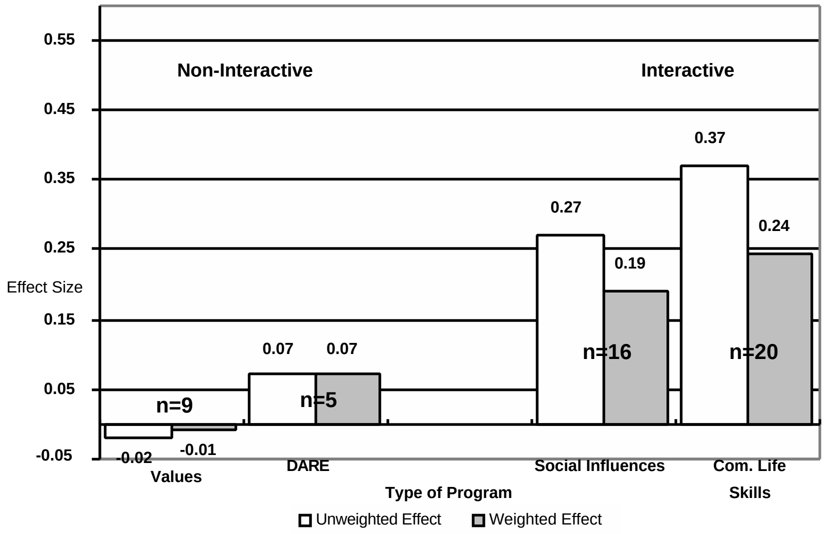
Effect Sizes for Values, DARE Type, SI, and CLS Programs

The magnitude of the UNES and WES increases from left to right in figure 14. The values programs were essentially zero. The DARE type programs had much lower effect sizes than either of the interactive programs. The SI programs were higher than the DARE type programs but lower than the CLS programs, which had the highest effect sizes.

Four of the five DARE type programs were Project DARE program evaluations that had been delivered to sixth graders. To alleviate concerns about being unable to detect program success with sixth graders (i.e., very low use rates), the four Project DARE evaluations were compared only to the sixth grade programs. As Project DARE shares content with both the noninteractive programs and the interactive programs, comparisons were made to both types of programs. The 16 fifth and sixth grade interactive programs effect sizes were much higher (UNES = 0.35, WES = 0.19) than the four Project DARE programs (UNES = 0.07, WES = 0.07) and were also higher than the other nine noninteractive programs (UNES = 0.05, WES = 0.08).

Comparisons Between Interactive Programs: SI Versus CLS

None of the regressions reported in Tobler (1992*a*) were statistically significant for the planned comparisons of SI programs and CLS programs. These results were confusing; the CLS programs effect sizes were consistently



higher than the SI effect sizes. Two possible confounds were discovered: more mental health specialists (the most successful leaders) conducted the CLS programs, and the SI programs were implemented more frequently on a large scale. First, to address the leaders issue, the OLS and WLS regressions were rerun with the regressions detailed in this chapter using only the 37 SI programs contrasted against the 25 CLS programs. Second, the regressions were repeated after eliminating all the programs that used mental health specialists as leaders.

The F change for the type of program (i.e., SI versus CLS) was nonsignificant in the four unweighted regressions (see table 5A) and the four weighted regressions (see table 5B). Additionally, the increment to R^2 for the SI versus CLS contrast accounted for less than 1 percent of the total R^2. Only one covariate, sample size, had a significant F change in all eight regressions. The increment to R^2 for sample size accounted for most of the total R^2. One other covariate—targeted drug—was significant, but only in the two OLS regressions.

TABLE 5A. *UNES and OLS regressions, SI versus CLS programs with and without MHS as leaders.*

	SI		CLS					
114 programs	N	X_{SI}	N	X_{CLS}	X_{CLS}- X_{SI}	R^2 change	F change	Sig. F
SI vs CLS	37	0.18	27	0.37	0.19	0.9%	0.881	0.352
Targeted drug						8.8%	4.458	0.016
Sample size						34.1%	31.026	0.000
56 PROGRAMS								
SI vs CLS	16	0.27	20	0.37	0.10	0.9%	0.451	0.508
Targeted drug						11.7%	3.110	0.060
Sample size						30.9%	15.176	0.000
114 PROGRAMS WITHOUT MHSs								
SI vs CLS	34	0.15	17	0.29	0.14	0.6%	0.426	0.518
Targeted drug						11.2%	4.196	0.021
Sample size						28.1%	19.173	0.000
56 PROGRAMS WITHOUT MHSs								
SI vs CLS	13	0.22	14	0.30	0.08	1.3%	0.392	0.540
Targeted drug						15.0%	2.600	0.999
Sample size						26.9%	9.186	0.006

TABLE 5B. *UNES and WLS regressions, SI versus CLS programs with and without MHS as leaders.*

	SI		CLS					
114 programs	N	X_{SI}	N	X_{CLS}	X_{CLS}- X_{SI}	R^2 change	F change	Sig. F
SI vs CLS	37	0.12	27	0.23	0.11	0.4%	0.355	0.554
Sample size						25.0%	20.020	0.000
56 PROGRAMS								
SI vs CLS	16	0.19	20	0.23	0.04	1.2%	0.490	0.491
Sample size						23.7%	10.562	0.003
114 PROGRAMS WITHOUT MHSs								
SI vs CLS	34	0.12	17	0.20	0.08	1.2%	0.078	0.781
Sample size						17.5%	10.417	0.002
56 PROGRAMS WITHOUT MHSs								
SI vs CLS	13	0.17	14	0.21	0.03	1.0%	0.252	0.622
Sample size						16.8%	5.034	0.034

Even though the effect sizes were lower after removing the mental health specialists, the large effect size differences between the SI and CLS programs remained (see tables 5A and 5B). Both the regressions and effect sizes indicate that the size of the program was the most important covariate when contrasting the SI and CLS programs.

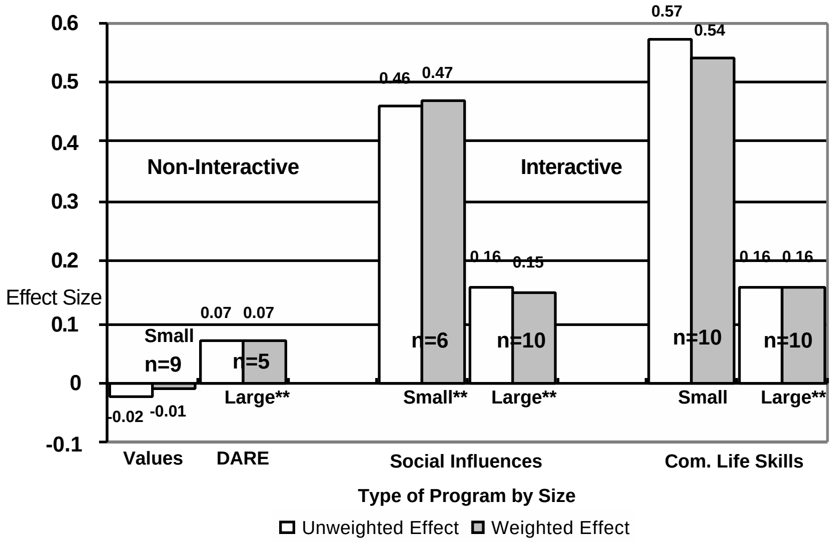
Effect Sizes by Sample Size for the K+A, SI, and CLS Programs

To further examine the relationship of program size to the type of program, the subset of 56 high-quality experimental studies was used as other areas of potential bias had also been eliminated. Only the K+A, SI, and CLS programs had enough cases to be further subdivided by size. Fortuitously, when the K+A programs were subdivided by size, all the small programs were in the subcategory called values and all the remaining large K+A programs were DARE type programs. Before examining the relationships between the six groups of programs (figure 15), it should be noted that the UNES and WES programs for each group were nearly identical. The large differences that existed between the UNES and WES for the other five covariates were not present when divided by the size of the program.

Both of the small interactive (SI and CLS) programs were extremely successful, while the small noninteractive (values) programs were totally ineffective. Also, the large mean differences between the SI programs and the CLS programs were substantially reduced; the small SI programs were nearly as effective as the small CLS programs. More important, the effect sizes for the large SI programs were equal to the large CLS programs. Unfortunately, both the large SI programs and the large CLS programs were only one-third as effective as their counterparts when implemented on a smaller scale. Despite this drop, the large interactive programs were still twice as effective as the large noninteractive DARE type programs.

Effectiveness by Drug Type for Noninteractive and Interactive Programs.

The differential results for the noninteractive and interactive programs by type of drug are presented in figure 16. The interactive programs were equally successful for cigarettes, alcohol, and marijuana; the UNESs ranged from 0.22 to 0.33 and the WESs ranged from 0.15 to 0.39. The effect size for marijuana was slightly higher than that for alcohol, with tobacco use having the lowest effect size of the three most frequently used drugs. Illicit drugs, excluding marijuana, had extremely high effect sizes but these results were based on only six programs.

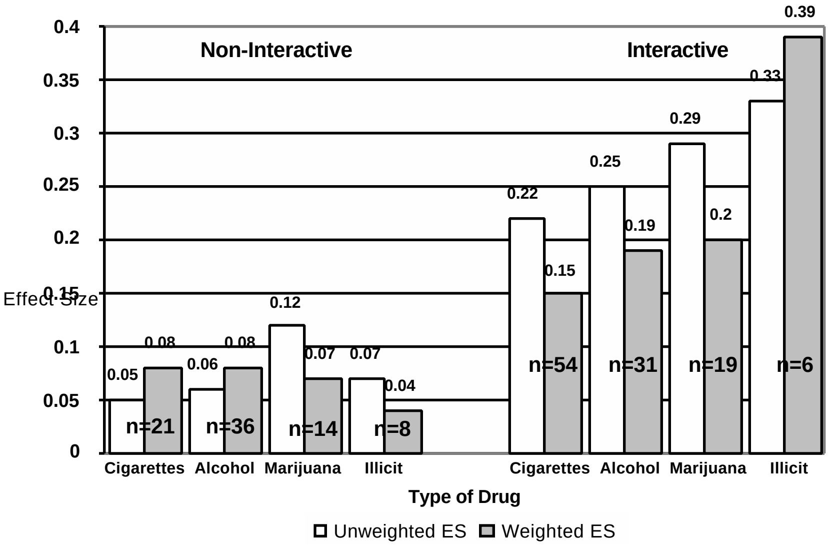


Further examination of the 56 high-quality programs showed effect sizes more closely related to other drugs: 0.19 UNES and 0.18 for the WES. The noniteractive programs were equally unsuccessful with all four types of drugs; the UNES ranged from 0.05 to 0.12, and the WES ranged from 0.04 to 0.08.

DISCUSSION

A Priori Organizational Scheme

Similar to primary research analyses, meta-analysis can be used to investigate relationships or to test specific hypotheses. The 1986 meta-analysis was exploratory; a wide net was cast to include a variety of programs and thereby identify relationships that were developed after a thorough coding of 143 programs. The relationships identified in 1986 laid the groundwork for the development of specific hypotheses, particularly concerning the type of program. Even though the 1993 analysis of type of program was based on two dimensions (content and process), it was similar to, and was a continuation of, the 1986 organizational scheme.



In the 1993 analysis, a program was coded and placed in one of the six subgroups in the predetermined classification scheme. Cluster analyses for 20 content and 4 process items verified the similarity of the programs within the 6 subgroups. The six subgroups were divided into the two major types of programs and were then tested with a priori planned comparisons (Tobler 1992*a*). When testing a specific hypothesis, the direction of inference is opposite of that found in an exploratory meta-analyses. "A hypothesis asserts which treatment is most effective: a review then examines empirical evidence to test the hypothesis" (Light and Pillemer 1984, p. 27).

Descriptive Analyses Confirm Two Types of Programs

The pessimistic reports of drug prevention program research have definitely resulted from the improper combining of two independent sets of programs. When analyzed collectively, the efficacy of drug prevention programming is questionable. Together, the mean effect size was 0.17 and the mode was zero. These results echo the pessimistic conclusions of the traditional literature reviews. However, when separated by the type of program (based on the a priori organizational scheme), two independent effect size distributions were observed. A second distribution for the effect size by sample size also indicated that two independent groups were combined. The noninteractive and interactive programs had similar funnel distributions showing a definite leveling of effect size, albeit at different magnitudes. A substantial difference was observed: the interactive programs had a mean effect size of 0.25 and a mode of 0.25, whereas the noninteractive programs had a mean effect size of 0.06 and a mode of zero. Both distributions verify the need to analyze the two types of programs separately and clearly illustrate the danger in grouping all drug prevention programs into one category. Instead of arriving at the incorrect assumption that nothing works, it can be concluded that although not all drug prevention programs work, the interactive programs were effective.

Group Process

The largest effect size differences were found between the noninteractive and interactive programs. Substantively, the characteristic that specified the difference between the noninteractive and interactive programs was the method used to deliver the program's content (i.e., the group process). Irrespective of the program content, the noninteractive programs did not emphasize interactions between peers as did the more participatory interactive programs. In fact, the delivery method or group process, not the content, was fundamental in defining the two types of programs. The majority of programs had multiple content components (Hansen 1992; Tobler 1993), and these overlapped within and between the two major types of programs (tables 1 and 3). Because the group process was not only an integral part, but perhaps was central, in defining the noninter-active and interactive programs, a brief review follows (see table 2 for greater detail).

Noninteractive Group Process.

The two noninteractive groups, A and B, used classroom dynamics familiar to all teachers. In the least interactive group (A), the leaders delivered a didactic presentation in a manner similar to a math, history, or health class. For the most part, these highly structured classes did not actively involve the students. Group B format was used by the majority of the noninteractive programs. Although a structured lecture format (i.e., passive) was used to present information, these groups also reported that students actively participated in teacher-led discussions. Experiential activities were incorporated, but these activities remained focused on the individual rather than on interactions with others in the group. For example, a values clarification exercise might involve adolescents independently listing their personal values, but the results of the exercise were generally shared only with the group leader in exchanges that excluded group peers.

Interactive Group Process.

Interactive group process skills have been defined by the Office of Substance Abuse Prevention

(1989, p. xiv): "This teaching technique is used to stimulate active participation of all students in the classroom activity, be it discussion, brainstorming session, or practice of new behaviors." Optimally, in group C the interactions included everyone and were both participatory and between peers. Structured small group activities were used to introduce program content and promote the acquisition of skills. This highly structured format was developmentally appropriate for younger adolescents, who bond with their peers as they participate in activities together. The leader keeps the group on track by initiating appropriately timed structured activities. Ideally, all adolescents practiced their newly acquired skills and received corrective feedback in a supportive atmosphere, enabling them to use their new skills in a situation of higher stress (i.e., a real world, drug-related situation).

The second interactive group, group D, was the converse of the traditional classroom. Group D had the least structure and, therefore, was more appropriate for older adolescents. Only three programs reported using this type of group. Even these groups maintained a definite structure and were neither wide-open discussions nor therapy groups. Optimally, the leaders in both interactive groups encouraged everyone to participate, promoted positive and supportive interactions between the adolescents, and assumed an authoritative role only when it was necessary to correct a misconception.

Importance of Sample Size

The success of the interactive programs was not without a caveat: the loss in effectiveness demonstrated by the larger programs was disappointing. This post hoc finding was second in importance only to the a priori hypotheses about the type of program. Although the mean effect size differences between the small programs and large programs were not quite as large as those observed between the two types of programs, the size of the program was also statistically significant in all the regressions. The magnitude of these effect size differences mandates that comparisons be made between similarly sized programs. Ideally, two independent meta- analyses should be conducted; one for the smaller efficacy trials[16] and one for the larger scale effectiveness trials.[17] However, this approach was not possible in the present analysis because of the limited number of studies.

For the small programs in the set of 114 programs, the extraordinary superiority of the 27 interactive programs is evident in figure 6. It is important to note that even when implemented on a small scale under ideal conditions, the noninteractive programs were ineffective. The difference between the small noninteractive and interactive programs was 0.36 for the UNES and 0.33 for the WES. When comparing the large programs, the differences between the noninteractive and interactive programs were much smaller, 0.07 for the UNES and 0.06 for the WES. Still, the large interactive programs were twice as effective the large noninteractive programs (figure 6).

Content

Focus of Noninteractive Programs.

The content of the noninteractive programs was directed towards individuals and their own internal perceptions, and therefore had a primarily intrapersonal focus. Despite variations within the noninteractive programs (KO, AO, and K+A subcategories), the program content maintained an intrapersonal focus. For example, the KO programs stressed the acquisition of factual knowledge about the physical and psychological consequences of drug use. The theoretical assumption was that given sufficient knowledge, the adolescent would develop negative drug attitudes that, in turn, would lead to healthy personal choices. The AO programs assumed that psychological factors place certain persons at risk of use and/or abuse. Various activities focused on building self-esteem and self-awareness, and promoting positive personal feelings with the aim of increasing the individual's intrapersonal competence and social functioning (no information about drugs was provided). The majority of the AO programs included in this meta-analysis trained teachers extensively in use of effective classroom management techniques (Moskowitz et al. 1984) for the purpose of altering the entire classroom milieu.

The K+A programs also had an intrapersonal focus, yet the two subcategories, value programs and DARE type programs, were based on very different theoretical assumptions. The values programs aimed to change the individual's personal attitudes and values about drug use. Therefore, the content included knowledge, decisionmaking skills, problemsolving skills, goal setting, values clarification, and so forth. These programs encouraged the adolescents to make a personal decision to abstain from using drugs based on ethical or moral considerations. The DARE type programs focused on ways to strengthen an individual's intrapersonal functioning to forestall the involvement with drugs (self-esteem building, self-acceptance, feelings of competence), and also included some interpersonal skills to strengthen social functioning.

Developmentally, the intrapersonal focus with its goal of increasing self-esteem may have greater potential in the elementary grades. Elementary students are usually in contained classrooms or with a single teacher for most of the day, allowing the individual attention and recognition necessary for this type of approach. A junior high school teacher, in contrast, can be involved with upwards of 120 adolescents daily (four to five classes), which makes using this approach particularly difficult. Many leaders reported that the K+A programs were hard to implement (Hansen et al. 1988; Schaps et al. 1981).

Focus of Interactive Programs. Interactive programs focus primarily on interpersonal competence, and peer pressure is assumed to be the paramount reason for adolescents' use of drugs. Newcomb and Bentler (1989) identified peer influences as the "most consistent and strongest of all factors, influencing the ‘average' youth" (p. 245). Therefore, two types of peer pressure were central to the interactive programs. First, drug refusal skills were used in all SI and CLS programs to enable the adolescents to skillfully negotiate the refusal of a drug offer and simultaneously remain accepted by their peer group. Second, peer pressure can take another form: "[P]eer influence may result from perception of peer attitudes and behaviors rather than from actual peer behavior" (Beisecker 1991, p. 234). Krohn and colleagues (1982) found that adolescent drug behaviors were determined by the "norm qualities of friends (compared to parents and religion)" and this "is clearly the most predictive variable" (p. 343).

Adolescents usually overestimate the drug use of their friends and other peers. Normative education was used to challenge the adolescent's perceptions. Firsthand, through peer-to-peer interactions, adolescents learn about their acquaintances' drug use or lack thereof. Also, through the leader's input, information was provided about the local school, State, and national levels of adolescent drug use. The presumption was that, as adolescents develop more realistic perceptions, their anxieties related to peer pressure will be reduced and, in turn, their drug use. Although not used as frequently as drug refusal skills, normative education was a component in the majority of the SI and the CLS programs.

Adolescence is a period in which establishing peer relationships takes priority over adult relationships. The peer pressure issues central to the interactive programs are in contrast to the noninteractive programs, which depend upon an ethical decision or personal change of values. It comes as no surprise that the interactive programs, based on peer-to-peer exchanges, were developmentally more appropriate and therefore more effective.

Interrelated Factors: Content, Process, and Size

The type of program was determined from a detailed coding for the nature of the program content and the type of delivery method or group process. The size of the program was identified as an influential third factor in this meta-analysis. Therefore, there exists a three-way interplay between the program content, the group process, and the size of the program. It is important to keep in mind that the focus of the content, in part, determines the method of delivery or group process. Until experimental studies compare identical content delivered with different group processes involving both small scale programs and larger programs, this puzzle will not be resolved. Notwithstanding, unequivocal statements can be made about certain combinations of content, process, and size that were identified in this collection of programs. To eliminate as many forms of bias as possible, only the values, the DARE type, SI, and the CLS programs included in the subset of 56 high-quality studies were used for the following comparisons.[18]

Noninteractive Programs. Comparisons cannot be made between the values programs and the DARE type programs, even though both were K+A programs and therefore used the same group process. The values (small K+A) programs were not implemented on a large scale and conversely, the DARE type (large K+A) programs were not implemented on a small scale. Comparisons cannot be made between the small unsuccessful values programs and the extremely successful SI and CLS programs because both the content and process differed. Only the following can be stated about the values programs: Content based on intrapersonal, ethical, and/or moral decisions that were presented in a noninteractive group did not change drug use behaviors, even when implemented in small programs.

Surprisingly, the large DARE type (K+A) programs appear to be somewhat more effective than the small values (K+A) programs. Statistical testing was not pursued because the two programs were not implemented in similarly sized programs. However, it would be expected that the larger DARE type programs would not do as well as the smaller values programs based on their size alone. Therefore, these results suggest that the values programs were particularly ineffective.

The DARE type programs were implemented in the same size programs as the large SI programs, but, again, comparison of the content or process cannot be made as both were different. What can be stated is that the combination of content and process used in the large SI programs showed statistically significant superiority when compared to the combination of content and process used by the DARE type (large K+A) programs. However, the content of the DARE type programs did show similarities to the CLS programs, although the group process[19] was different. Although both programs had an intrapersonal and interpersonal focus, the focus of the content in the DARE type programs was highly intrapersonal, with less emphasis on interpersonal drug skills. The CLS programs focused primarily on interpersonal skill building and to a lesser degree included a variety of intrapersonal components. The large CLS interactive programs were statistically superior to the large DARE type (K+A) noninteractive programs, indicating that the more interpersonal emphasis used in an interactive group was more effective.

Interactive Programs. Fortunately, the SI and CLS programs were represented as both small and large programs. Of equal importance, both the SI and CLS programs used an interactive group to deliver the content, and therefore the content can be compared. The SI programs stressed varied aspects of the social context that influences drug use and combined this with mastery of drug refusal skills. The more comprehensive CLS programs added many generic skills to the content of the SI programs and, in some cases, included both an intrapersonal and an interpersonal focus. Within the set of small programs, there were no statistically significant differences between the more singularly focused SI programs and the CLS programs, although the CLS programs had slightly higher effect sizes. For the large programs, the SI and CLS programs showed identical effectiveness.

Comprehensiveness of Content. The comprehensiveness of the program content appears to have no impact, as seen in the above comparison of the SI and CLS programs. Before controlling for the size of the program, it appeared that the more comprehensive CLS programs were more effective than the more singular SI programs. The possibility existed that inclusion of more content would increase effectiveness. However, closer examination of the nature of the additional content showed that its focus was expanded in some of the CLS programs to include both an interpersonal and intrapersonal focus (Botvin and Dusenbury 1989). Yet, other CLS programs maintained the interpersonal focus even though using a more comprehensive content (Schnicke and Gilchrist 1984). In the smaller programs, the additional components did produce programs with somewhat higher effect sizes than the more singular SI programs, but this was not statistically significant. When comparing the large programs, the effect sizes for the CLS programs were equal to the large SI programs. For both the small and large programs, statistically significant evidence does not exist to support the greater efficacy of the more comprehensive CLS programs.

Five Remaining Covariates

Targeted Drug.

The interactive programs were consistently much higher than the noninteractive whether the program targeted cigarettes or alcohol or did not target a specific drug. Only the WES for the noninteractive alcohol programs came close to the interactive alcohol programs. For the interactive programs, the generic drug prevention programs were at least three times more effective in preventing cigarette, alcohol, marijuana, and other drug use than the noninteractive programs.

Another question remains. How effective were programs which target a specific drug versus including that drug in a generic drug prevention program? There appears to be a considerable advantage in targeting cigarettes when using an interactive program, particularly if the program was implemented on a small scale. A possible explanation could be implementation problems experienced when small programs involved in efficacy trials were delivered under real world conditions. However, the evidence from the generic programs necessitates a different conjecture. Whether large or small, the generic programs within the interactive programs were approximately three times less effective than the smoking programs in preventing cigarette use. As size was not a factor, possibly the generic programs were less successful because an implicit message of lifetime abstinence was delivered in the smoking programs, while the generic programs, intentionally or unintentionally, may deliver a message of abstinence only until the drugs can be purchased legally.

Targeting alcohol within the noninteractive programs produces results similar to not targeting alcohol (i.e., generic drug programs). Targeting alcohol appeared to decrease program effectiveness for alcohol use when compared to generic programs in the set of 114 programs. However, the slightly lower effect sizes for the alcohol programs were not statistically significant in the subset of 56 high-quality experimental programs.

Leaders. No single type of leader produced mean effect sizes that were statistically significant. Mental health specialists,[20] whose degree training involves the skills necessary to promote active group participation, were the most effective leaders, although not significantly so. However, only 29 percent of the mental health specialists were involved in large scale implementations associated with lower effect sizes. Peer leaders were used by only 21 percent of the programs. More often, the peer leader was a copartner with an adult leader. Peer leadership does not define an interactive program, nor was it a prerequisite for promoting the necessary group interaction, but peer leaders can be helpful in setting the stage and in supportive roles. In fact, it appears that the credentials of the leader may not be the issue as much as whether the leader can facilitate the necessary group interactions. The use of outside professionals may be questionable despite their level of skills, particularly if they are in the building only once a week for the drug prevention program. Also, a larger challenge remains. Can teachers create the atmosphere necessary for a truly interactive group when they have not been trained in the use of group skills, typically do not use the group process to present the course content, and must act as disciplinarians throughout the day?

Type of Control Group. Differential effectiveness was observed for the different types of control groups. The effect sizes were attenuated when the comparison/control group was a standard health class and/or another treatment. The differential between the two types of control groups was identical whether the programs were noninteractive or interactive. Programs that were compared to a no-treatment group reported an UNES about 0.08 higher than those compared to the standard health class. Not surprisingly, this difference was equivalent to the UNES of 0.09 for the KO programs (typical health classes). The two types of control groups appeared with equal frequency within the noninteractive or interactive programs; therefore, this variable does not contribute to the difference in effectiveness between the two types of programs.

Analyses of the type of control group, however, highlight another important issue for drug prevention program research. Since the 1986 Drug Free Schools and Community Act, few true no-treatment groups exist. In 1993, 38 percent of the programs were compared to the standard health class/another treatment control groups, an increase of 12 percent from previous findings in 1986. This trend can be expected to continue; therefore, researchers should include information about the program content and the delivery method used by the standard health class control group. This information was seldom reported and is extremely important. For example, Ary and colleagues (1989) found that the control schools were receiving the same number of sessions (12) as the treatment schools. In actuality, Ary and colleagues' program was compared to another program of equal strength. In the same vein, it is the rare school system that does not have drug prevention activities (e.g., assemblies, drug prevention week). Therefore, evaluations should mention all other drug education activities and the amount of previous exposure to drug prevention programs.

Experimental Design. A pervasive drug prevention research problem is high dropout rates, that is, experimental mortality. This problem is exacerbated by the fact that drug users drop out of programs (even school-based) at higher rates. This was confirmed by 63 percent of the programs in this meta-analysis. In most cases, the dropouts come from highly transient populations. Frequent moves can be indicative of unstable families, which have been correlated to higher drug use rates among adolescents (Ellickson et al. 1988). If analyses indicate no differential dropout from the treatment group and the control group, the internal validity of the experimental design has not been compromised. Unfortunately, only 37 percent of the programs reported this information. When a program is successful with the drug-using population, the high attrition rates could restrict the magnitude of program effectiveness (i.e., users were not present to show decreased use). This meta-analysis showed higher mortality rates were indeed associated with slightly lower effect sizes in the interactive programs, although no differences were observed for the noninteractive programs.

Special Populations. Programs including minority student populations were equally or slightly more effective than those delivered to white populations across two situations: large programs were highly successful when the school had a combined minority population over 50 percent, and small programs designed to be culturally sensitive were equally successful with black, Hispanic, or Native American adolescents.

Intensity

The explanation of no significant findings for the intensity variable was possibly related to the low intensity of both the noninteractive and interactive programs; both were only 10 hours. Sixty-eight percent of all the programs included were low intensity programs with a mean delivery intensity of only 6 hours; only 16 programs offered boosters, and only 4 provided more than 1 year of boosters. Positive behavioral effects were found for the interactive programs at an intensity of 10 hours, in contrast to findings of a national survey of 4,738 students in grades 3 to 12 in which no behavioral effects were observed at the end of 1 year of health education, although positive behavioral effects did appear at the end of 3 years of continuous health education (Health Education Works 1990). Similarly, the School Health Education Evaluation found "‘medium' effects are achievable for general health practices when more than 30 hours of classroom instruction is provided" (Connell et al. 1985, p. 321).

Levels of Use

Although evaluators are increasingly determining program success based on an individual's initial level of drug use, only 35.8 percent (43) of the programs had classified participants by level of drug use. For more than a decade, Goodstadt (1986) has advocated measuring program success based on a participant's prior level of drug use to determine if differential effectiveness existed. A priori, Dielman and colleagues (1990) separated program participants on previous drinking experience. Without separate analyses, Dielman and colleagues found that some effects would have been attenuated and others would have been completely masked. Biglan and colleagues (1987) also illustrated the danger in lumping all subpopulations together. In this case, the nonsignificant findings for the nonsmokers completely overshadowed the highly significant findings for smokers.

The interactive programs were generally successful with smokers, as evidenced by five small but highly effective program outliers. The program outliers were excluded from the regressions, therefore, they will be mentioned here. Two separate SI programs were implemented with impoverished inner-city black males. One program was highly successful with cigarette users (Spitzzeri and Jason 1979). The second found limited effects with cigarette users, but was highly successful with experimental cigarette smokers (Jason et al. 1982). The remaining three program outliers[21] were variations of the subcategory of interactive programs called "others" and were implemented with high school cigarette smokers (Greenberg and Deputat 1978). Logically, these programs did not include refusal skills as these adolescents were smokers who were dealing with cessation issues. The content centered on knowledge of the physical effects and health risks associated with smoking. An age-appropriate, less structured interactive group (group D) was used to convey this information. These outliers suggest that it might be beneficial to separate out the cigarette smokers with a distinctly different type of program, particularly at high school age.

Implementation

Drops in the magnitude of effectiveness experienced by the large programs suggest that factors other than statistical leveling of effect sizes (observed as the number of participants increases) were operating. Implementation factors seem to be a more probable explanation and a crucial mediating factor in determining the success of a program. Was an essential ingredient of the interactive programs missing, that of active involvement and interaction between peers? Ideally, an interactive program incorporates participation by everyone, preferably in small groups. To become proficient in the use of drug refusal skills or other new interpersonal skills, each individual needs a sufficient opportunity to practice before an assumption can be made that the skills can be transferred to actual drug use situations. If implemented in a regular classroom without extra leaders, the likelihood of every adolescent interacting on a regular basis to make this possible remains questionable. Along the same line, Botvin and colleagues (1990) found that some teachers did not include all parts of the program equally, possibly because they felt uncomfortable with certain areas such as the role plays. Other teachers "may not have been convinced that the approach being tested was as effective as teaching factual information about drugs and the adverse consequences of use" (p. 27). Botvin recommended extensive training to convince the teachers of the merits of this type of program and to provide the requisite skills and confidence necessary for implementation of this type of curriculum.

Replication of Findings With a Set of 56 High-Quality Experimental Programs

The quasi-experimental (nonrandomly assigned) programs were eliminated to rule out the possibility of positive bias. Instead of obtaining lower effect sizes, the set of 56 experimental programs had higher effect sizes and an even larger differential between the noninteractive and interactive programs. An alternative explanation for the higher effect sizes may be the more stringent selection criteria used. Nevertheless, even these high-quality programs had flaws in their evaluations or implementations that may have positively or negatively biased the program effect size. Some sources of systematic error were removed in the subset of 56 studies; the remaining flaws, it is hoped, were random. For example, history may have affected some studies while implementation factors may have presented problems in others; but, when enough programs are included, these flaws may be considered random error. "What is systematic error in an individual study may be random error in the context of a meta-analysis" (Shotland and Mark 1987, p. 86).

Perhaps the confusion reported in the literature arises, to a larger degree, from including programs whose success was attenuated or inflated for one or more reasons. To address this question, the entire set of programs, whether randomly or nonrandomly assigned, was subjected to the more stringent selection criteria. The end result was a set of 68 programs, 56 experimental and 12 quasi-experimental. In table 6 the effect sizes are given for the entire set of 114 programs, the set of 68 programs that excluded programs with problems which could bias their results, and the final set of 56 programs that excluded problematic programs and also were randomly assigned. Following removal of problematic programs, there was an increase of 0.07 in the difference in both UNES and WES between the noninteractive and interactive programs. On the second step, removal of the quasi-experimental programs, the effect size difference between the noninteractive and interactive programs increased by 0.04 for the UNES and 0.02 for the WES. Ruling out the other sources of bias was far more important (nearly twice as much) as whether a program was randomly assigned. Additionally, only 34 percent of all the experimental programs were eliminated for other problems. Whereas 67 percent of the quasi-experimental programs had additional problems, a disproportionate number were compromised for reasons other than the lack of random assignment. The inclusion of quasi-experimental programs attenuated the magnitude of success of the interactive programs and made the results more ambiguous. Similarly, Hedges and Olkin (1985) found "that variations in the outcomes of well-controlled studies are considerably easier to model than are variations in the outcomes of poorly controlled studies" (p. 14).

Success Across Drugs and Populations. The success of the interactive programs was equivalent across all types of substances: cigarettes, alcohol, marijuana, and all other illicit drugs. This finding contradicts the reviews that have reported the success of drug prevention programs for cigarettes, yet have failed to report equivalent success for alcohol and other drugs (Botvin 1990; Flay 1985*b*; Moskowitz 1989). It is this author's conjecture that the majority of reviews arrived at these conclusions because of the limited number and variety of programs included in their reviews and, additionally, because all types of prevention programs were lumped together.

TABLE 6. *Mean difference between noninteractive (NI) and interactive (I) UNES and WES without problematic and nonrandomly assigned programs.*

	114 programs			68 without problems			56 experimental without problems		
	X_{NI}	X_I	Diff	X_{NI}	X_I	Diff	X_{NI}	X_I	Diff
n	44	70		21	47		18	38	
UNES	0.06	0.25	0.19	0.02	0.28	0.26	0.02	0.32	0.30
WES	0.08	0.16	0.08	0.04	0.17	0.15	0.04	0.21	0.17

The findings of the 1993 meta-analysis were similar to earlier findings in 1986 which showed the peer programs had equal success across all types of drugs. (In 1986, both the SI and CLS programs were included in one category called peer programs.) The lower effect sizes in 1993 may be the result of all adolescents receiving some form of drug prevention information in the last decade (i.e., media, school assemblies, community activities).

Encouragingly, the interactive programs were successful in schools with predominately minority populations. This also repeats the 1986 findings where peer programs were found equally successful with minority or white populations.

Similar Conclusions across Multiple Statistical Analyses. Remarkably consistent patterns were observed across the numerous and varied statistical procedures. The results reported here were a reanalysis of an earlier report to resolve the problem of extremely complex regression analyses with four separate size groups (i.e., too few programs for the number of parameters). Additionally, interpretations of the regression analyses were augmented with detailed descriptive statistics for the 114 programs and then separately for the noninteractive and interactive programs. Finally, statistics developed specifically for meta- analysis were used to further verify the results of the regression analyses (e.g., homogeneity of effect size, model specification).

Equivalent Success for Five Extremely Large Programs.

The effectiveness of programs implemented on a large scale can be diminished by control groups subject to factors such as mandated drug curriculum, unmotivated teachers, incomplete implementation, and perhaps most important, a limited amount of small group interaction; the list is almost unending. Fortunately, for the sake of comparison, five of the six largest programs in this meta-analysis were SI programs. These five large implementations had a mean sample size of 6,516 tested students and achieved an WES of 0.13. This effect size was equivalent to the effect size for the remaining 32 SI programs (WES = 0.12), which had a mean sample size of 924 tested students. This WES was accomplished in spite of operating under real world conditions. Also, four of the five large SI programs were compared to a standard health class control and may have achieved an effect size of 0.21 had they been compared to a no-treatment control group (i.e., difference of 0.08 between no-treatment and a standard health class control group). The consistency of these results, even though small, provides a robust finding (Flay 1985*a*). In other words, "Two 0.06 results are much stronger evidence against the null than one 0.05; and 10 p's of 0.10 are stronger evidence against the null than 5 p's of 0.05" (Rosenthal 1990, p. 133).

CHALLENGES

The identification of the types of programs that work generates more questions: Why are people still using those programs that don't work, particularly across whole States (Ennett 1993; Klitzner 1988)? Have efforts stopped short of the goal and not made the successful programs available to the general school population in a marketable form that can be placed in the hand of the teachers or principals? Even when educators are informed about recent research and would choose an interactive program, to the author's knowledge (with one exception[22]), program curriculums are not in a form that could be implemented with minimal effort. The schools have only one choice—the noninteractive programs.

A second challenge is whether policymakers can be convinced to shift to the more interactive approaches and do so quickly. With drug education mandated in most States, answers are needed to the following questions. What would be the cost of "changing horses in midstream"? Is the small effect size for these programs worth the cost to the taxpayers? What will the impact be on parents and communities that have enthusiastically and energetically supported a program, only to find the program has minimal or no effect on adolescent drug use? Imagine what could happen if this community enthusiasm were marshaled to support programs that have already demonstrated the ability to prevent, delay, or decrease drug use. Project STAR (Students Taught Awareness and Resistance) in Kansas City (Pentz et al. 1986) and the Minnesota Heart Health Program (Perry et al. 1989) combined community support with an efficacious school-based program. Both achieved a WES of 0.20, nearly double the effect sizes of similar types of school-based programs implemented on a large scale without community involvement.

Third, what is being done to address other antecedents of adolescent drug use besides peer pressure? School-based programs that are offered only once, most often in junior high school, cannot be considered a silver bullet to last throughout adolescence. School hours occupy only a small part of an adolescent's day; therefore, these programs cannot be expected to "counter the range of powerful forces that operate outside the walls of the classroom and school" (Goodstadt 1987, p. 31).

Finally, the paramount question for school boards and administrators is whether they will provide the necessary class time, the extra personnel, and the aggressive teacher training in the use of interactive group process skills. These efforts would restore the operative ingredient that may have been missing from the larger interactive programs: that of active involvement, an opportunity to exchange ideas and discover alternative perspectives, and sufficient practice time to assimilate the new interpersonal skills.

NOTES

1. The reanalysis also included a correction for overrepresentation of some programs in Tobler (1986). Only one effect size per program strategy was reported.

2. In Tobler (1992*a*), 18 nonorthogonal planned comparisons, the result of an extremely fine-tuned coding scheme, were tested with the full set of 114 programs and also for 3 subsets grouped by size. The number of programs in each of the three size groups was less than 40; therefore, these analyses were open to spurious findings and may have lacked power to detect significant findings. However, this was offset by verifying the results using a second regression procedure, weighted structural regression (WSR). WSR was developed to alleviate problems of numerous, correlated predictors and limited sample sizes faced by social scientists (Pruzek and Lepak 1992).

3. The 1986 meta-analysis used modality to refer to the specific type or strategy of a program.

4. A comparison of the 1986 and 1993 terminology as well as a detailed discussion of the content and the delivery process can be found in Tobler (1993).

5. Means and standard deviations were reported in only 10 percent of the studies in Tobler (1986).

6. High-risk youth is defined as an individual who is a school dropout; has become pregnant; is economically disadvantaged; is the child of a drug or alcohol abuser; is a victim of physical, sexual, or psychological abuse; has committed a violent or delinquent act; has experienced mental health problems; has attempted suicide; or has experienced long-term physical pain due to injury [Public Law 99-570, Sec. 4122 (b)(2)].

7. The drug use etiology for these populations necessitates multimodal and markedly different types of prevention programs (Bry 1982; Hawkins et al. 1987; Swisher and Hu 1983; Wall et al. 1981).

8. The two community studies were excluded as they offered a variety of additional support over the 4 years.

9. The content areas termed "extracurricular activities" and "others" occurred very infrequently and were subsequently dropped.

10. The term "mental health specialists" includes counselors, psychologists, psychiatrists, Ph.D.s or the equivalent in human services, or graduate level social workers.

11. Additional information about the increment to R^2, F change, and its significance for the independent variable as well as any covariates that reached significance in the OLS and WLS regressions for 114 programs and the subset of 56 experimental programs can be found in Tobler (1994).

12. Literature surveys were used only to locate the programs. In all cases, the original report was obtained for the meta-analysis.

13. This group of programs belongs in the strategy type that was referred to as "alternative programs" (MOD5) in the meta-analysis of 143 programs (Tobler 1986).

14. Six programs outliers were identified in the regression analyses and removed, which reduced the set of 120 to 114.

15. Regression procedures could not be used because the number of covariates (N = 6) was too large for the number of programs with cigarette outcome measures (N = 38).

16. "Efficacy trials provide tests of whether a technology, treatment, procedure, or program does more good than harm when delivered under optimum conditions" (Flay 1986, p. 451).

17. Effectiveness trials are defined as "Trials to determine the effectiveness of an efficacious and acceptable program under real-world conditions of delivery/implementation" (Flay 1986, p. 459).

18. The KO, AO, and others programs had three or fewer programs in their categories and were not included here.

19. These programs were implemented by police officers who delivered the content with lectures and/or officer-directed discussions which were seldom broken into small groups to provide the necessary interaction for a strong interpersonal focus.

20. Only 3 of the 20 mental health specialists delivered noninteractive programs.

21. As well as being highly positive outliers, the three programs targeted only cigarette smokers and were excluded for not meeting the selection criteria (i.e., a conservative assumption that regular smokers are addicted).

22. Contact Gilbert J. Botvin, Ph.D., Professor and Director, Institute for Prevention Research, New York Hospital Cornell Medical Center, Room KB 201, 411 East 69th Street, New York, NY 10021. Telephone (212) 746-1270.

REFERENCES

Ary, D.; Biglan, A.; Glasgow, R.; Zoref, L.; Black, C.; Ochs, L.; Severson, H.; Kelly, R.; Weissman, W.; Lichtenstein, E.; Brozovsky, P.; and Wirt, F. *School-Based Tobacco Use Prevention Programs: Comparing a Social-Influence Curriculum to "Standard-Care" Curricula.* 1989 Available from Dennis V. Ary, Ph.D., Oregon Research Institute, Willamette Street, Eugene, OR 97401.

Bangert-Drowns, R. Review of developments in meta-analytic method. *Psychol Bull* 99(3):388-399, 1986.

Beisecker, A. Interpersonal approaches to drug abuse prevention. In: Donohew, L.; Sypher, H.; and Bukoski, W., eds. *Persuasive Communication and Drug Abuse Prevention*. Hillsdale, NJ: Lawrence Erlbaum Associates, 1991. pp. 229-238.

Biglan, A.; Glasgow, R.; Ary, D.; Thompson, R.; Severson, H.; Lichtenstein, E.; Weissman, W.; Faller, C.; and Gallison, C. How generalizable are the effects of smoking prevention program? Refusal skills training and parent messsages in a teacher-administered program. *J Behav Med* 10(6):613-628, 1987.

Botvin, G. Substance abuse prevention: Theory, practice and effectiveness. In: Tonry M., and Wilson J., eds. *Drugs and Crime, Crime and Justice.* Vol. No. 11. Chicago: University of Chicago Press, 1990. pp. 461-520.

Botvin, G., and Dusenbury, L. Substance abuse prevention and the promotion of competence. In: Bond, L., and Compas, B., eds. *Primary Prevention and Promotion in the Schools*. Newbury Park, CA: Sage Publications, 1989. pp. 146-178.

Botvin, G.; Baker, E.; Filazolla, A.; and Botvin, E. A cognitive behavioral approach to substance abuse prevention: One year follow up. *Addict Behav* 15:47-63, 1990.

Bry, B. Reducing the incidence of adolescent problems through preventive intervention: One- and five-year follow-up. *Am J Community Psychol* 10(3):265-275, 1982.

Cohen, J., and Cohen, P. *Applied Multiple Regression Correlation Analysis for the Behavioral Sciences.* Hillsdale, NJ: Lawrence Erlbaum Associates, 1983.

Connell, D.; Turner, R.; and Mason, F. Summary of findings of the School Health Education Evaluation: Health promotion effectiveness, implementation, and costs. *J Sch Health* 55(8):316-321, 1985.

Dielman, T.; Shope, J.; and Butchart, A. "Differential Effectiveness of an Elementary School-Based Prevention Program." Paper presented at the University of California, San Diego Extension Seminar, *What Do We Know about School-Based Prevention Strategies? Alcohol, Tobacco and Other Drugs,* San Diego, CA, March 17-20, 1990.

Edwards, G. *Reaching Out.* Barrytown, NY:, Station Hill Press/PULSE, 1972.

Ellickson, P.; Bell, R.; Thomas, M.; Robyn, A.; and Zellman, G. *Designing and Implementing Project ALERT. A Smoking and Drug Prevention Experiment.* Santa Monica, CA: Rand Corporation, 1988.

Ennett, S. How effective is Project DARE? In: Montoya, C.; Ringwalt, C.; Ryan, B.; and Zimmerman, R., eds. *Evaluating School-Linked Prevention Strategies; Alcohol, Tobacco, and Other Drugs*. San Diego, CA: University of California Press, 1993. pp. 51-61.

Flay, B. Psychosocial approaches to smoking prevention: A review of findings. *Health Psychol* 4(5):449-488, 1985*a*.

Flay, B. What we know about the social influences approach to smoking prevention: Review and recommendations. In: Bell, C., and Battjes, R., eds. *Prevention Research: Deterring Drug Abuse among Children and Adolescents.* National Institute on Drug Abuse Research Monograph 63. DHHS Pub. No. (ADM)85-1334. Washington, DC: Supt. of Docs., U.S. Govt. Print. Off., 1985*b*. pp. 67-112.

Flay, B. Efficacy and effectiveness trials (and other phases of research) in the development of health promotion programs. *Prev Med* 15:451-474, 1986.

Giaconia, R., and Hedges, L. Identifying features of effective open education. *Rev Educ Res* 52(4):579-602, 1982.

Gilchrist, L.; Schinke, S.; Trimble, J.; and Cvetkovich, G. Skills enhancement to prevent substance abuse among American Indian adolescents. *Int J Addict* 22 (9):869-879, 1987.

Glass, G.; McGaw, B.; and Smith, M. *Meta-Analysis in Social Research*. Beverly Hills, CA: Sage Publications, 1981.

Goodstadt, M. School based drug education in North America: What is wrong? What can be done? *J Sch Health* 56(7): 278-281, 1986.

Goodstadt, M. Prevention strategies for drug abuse. *Issues Science Technol* 3(2):28-35, 1987.

Goodstadt, M. School-based education research findings: What have we learned? What can be done? In: Rey, K.; Faegre, C.; and Lowery, P., eds. *Prevention Research Findings: 1988*. Office for Substance Abuse Prevention Prevention Monograph 3. DHHS Pub. No. (ADM)91-1615. Washington, DC: Supt. of Docs., U.S. Govt. Print. Off., 1990. pp. 33-45.

Goodstadt, M., and Sheppard, M. Three approaches to alcohol education. *J Stud Alcohol* 44(2):362-380, 1983.

Greenberg, J., and Deputat, Z. Smoking intervention: Comparing three methods in a high school setting. *J Sch Health* 48(1):498-502, 1978.

Hansen, W. School-based substance abuse prevention: A review of the state of the art in curriculum, 1980-1990. *Health Educ Res* 7(3):403-430, 1992.

Hansen, W.; Johnson, D.; Flay, B.; Graham, J.; and Sobel, J. Affective and social influences approaches to the prevention of multiple substance abuse among seventh grade students: Results from Project SMART. *Prev Med* 17:135-154, 1988.

Hansen, W.; Tobler, N.; and Graham, J. Attrition in substance abuse prevention research: A meta-analysis of 85 longitudinally followed cohorts. *Eval Rev* 14(6): 677-685, 1990.

Hawkins, D.; Lishner, D.; Jenson, J.; and Catalano, R. Delinquents and drugs: What the evidence suggests about prevention and treatment programming. In: Brown, B., and Mills, A., eds. *Youth at High Risk for Substance Abuse.* Alcohol, Drug Abuse and Mental Health Administration. DHHS Pub. No. (ADM)87-1537. Washington, DC: Supt. of Docs., U.S. Govt. Print. Off., 1987. pp. 81-131.

Health education works. *Prev Forum*, January 17, 1990.

Hedges, L. Advances in statistical methods for meta-analysis. In: Cordray, D., and Lipsey, M., eds. *Evaluation and Statistical Review Annual.* Vol. 11. Beverly Hills, CA: Sage, 1986. pp. 731-748.

Hedges, L., and Olkin, I. *Statistical Methods for Meta-analysis.* New York: Academic Press, 1985.

Hunter, J., and Schmidt, F. *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings.* Newbury Park, CA: Sage Publications, 1990.

Jason, L.; Mollica, M.; and Ferrone, L. Evaluating an early secondary smoking prevention intervention. *Prev Med* 11(1):96-102, 1982.

Johnston, L.; Bachman, J.; and O'Malley, P. *Drug Use, Drinking, and Smoking: National Survey Results from High School, College, and Young Adult Populations, 1975-1988.* DHHS Pub. No. (ADM)89-1638. Washington, DC: Supt. of Docs., U.S. Govt. Print. Off., 1989.

Johnston, L.; O'Malley, P.; and Bachman, J. *Drug Use among American High School Students, College Students, and Other Young Adults: National Trends 1985.* DHHS Pub. No. (ADM)86-1450. National Institute on Drug Abuse. Washington, DC: Supt. of Docs., U.S. Govt. Print. Off., 1986.

Klitzner, M. *Report to Congress on the Nature and Effectiveness of Federal, State, and Local Drug Prevention/Education Programs, Part 2; An Assessment of the Research on School-Based Prevention Programs.* U.S. Department of Education, Office of Planning, Budget and Evaluation. Washington, DC: Supt. of Docs., U.S. Govt. Print. Off., 1988. pp. 1-47.

Krohn, M.; Akers, R.; Radosevich, M.; and Lanza-Kaduce, L. Norm qualities and adolescent drinking and drug behavior: The effects of norm quality and reference group on using and abusing alcohol and marijuana. *J Drug Issues* 12(4):343-360, 1982.

Light, R., and Pillemer, D. *Summing Up: The Science of Reviewing Research.* Cambridge, MA: Harvard University Press, 1984.

Lipsey, M. Juvenile delinquency treatment: A meta-analytic inquiry into the variability of effects. In: Cook, T.; Cooper, H.; Cordray, D.; Hartmann, H.; Hedges, L.; Light, R.; Louis, T.; and Mosteller, F., eds. *Meta-Analysis for Explanation: A Casebook.* New York: Russell Sage Foundation, 1992. pp. 83-128.

Moskowitz, J.; Malvin, J.; Schaeffer, G.; and Schaps, E. Evaluation of an affective development teacher training program. *J Prim Prev* 4(3):150-161, 1984.

Moskowitz, J. The primary prevention of alcohol problems: A critical review of the research literature. *J Stud Alcohol* 50:54-88, 1989.

Murray, D.; O'Connell, C.; Schmid, L.; and Perry, C. The validity of smoking self-reports by adolescents: A re-examination of the bogus pipeline procedures. *Addict Behav* 12:7-15, 1987.

Newcomb, M., and Bentler, P. Substance use and abuse among children and teenagers. *Am Psychologist* 44(2):242-248, 1989.

Oetting, E., and Beauvais, F. Adolescent drug use: Finding of national and local surveys. *J Consult Clin Psychol* 58(4):385-394, 1990.

Office of Substance Abuse Prevention. *Prevention Plus II: Tools for Creating and Sustaining Drug-Free Communities.* Office of Substance Abuse Prevention. DHHS Pub. No.(ADM)89-1649. Washington, DC: Supt. of Docs., U.S. Govt. Print. Off., 1989.

O'Malley, P.; Bachman, J.; and Johnston, L. Reliability and consistency in self-reports in drug use. *Int J Addict* 18(6):805-824, 1983.

Pechacek, T.; Murray, D.; Luepker, R.; Mittlemark, M.; Johnson, C.; and Shutz, J. Measurement of adolescent smoking behavior: Rationale and methods. *J Behav Med* 7(1):123-140, 1984.

Pentz, M.; Cormack, C.; Flay, B.; Hansen, W.; and Johnson, C.A. Balancing program and research integrity in community drug abuse prevention: Project STAR approach. *J Sch Health* 56(9):389-393, 1986.

Perry, C.; Klepp, K.; and Sillers, C. Community-wide strategies for cardiovascular health: The Minnesota Heart Health Program youth program. *Health Educ Res* 4(1):87-101, 1989.

Pirie, P.; Murray, D.; and Luepker, R. Smoking prevalence in a cohort of adolescents, including absentees, dropouts, and transfers. *Am J Public Health* 78(2):176-178, 1988.

Pruzek, R., and Lepak, G. Weighted structural regression: A broad class of adaptive methods for improving linear prediction. *Multivariate Behav Res* 27(1):95-129, 1992.

Rhodes, J., and Jason, L. The retrospective pretest: An alternative approach in evaluating drug prevention programs. *J Drug Educ* 17(4):345-356, 1987.

Ringwalt, C.; Curtain, T.; and Rosenblum, D. "A First-Year Evaluation of D.A.R.E. in Illinois." Paper presented at the University of California, San Diego Extension Seminar, *What Do We Know about School-Based Prevention Strategies? Alcohol, Tobacco, and Other Drugs.* San Diego, October 18-19, 1990.

Rosenthal, R. *Meta-Analytic Procedures for Social Research.* Applied Social Research Methods Series. Vol. 6. Beverly Hills, CA: Sage Publications, 1986.

Rosenthal, R. An evaluation of procedures and results. In: Wachter, K., and Straf, M., eds. *The Future of Meta-*

Analysis. New York: Russell Sage Foundation, 1990. pp. 123-133.

Schaps, E.; Moskowitz, J.; Condon, J.; and Malvin, J. *An Evaluation of an Innovative Drug Education Program: First Year Results.* Report prepared for the National Institute on Drug Abuse Prevention Branch, 1981.

Schinke, S., and Gilchrist, L. *Life Skills Counseling with Adolescents.* Baltimore: University Park Press, 1984.

Shotland, R., and Mark, M. Improving inferences from multiple methods. In: Mark, M., and Shotland, R., eds. *Multiple Methods in Program Evaluation. New Direction for Program Evaluation.* Vol. 35. San Francisco: Jossey-Bass Inc., 1987. pp. 77-94.

Single, E.; Kandel, D.; and Johnson, B. The reliability and validity of drug use responses in a large scale longitudinal survey. *J Drug Issues* 5:426-443, 1975.

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

Spitzzeri, A., and Jason, L. Prevention and treatment of smoking in school age children. *J Drug Educ* 9(4):315-325, 1979.

SPSS Inc. *SPSS Reference Guide*. 1990. Available from SPSS., Inc., 444 North Michigan Avenue, Chicago, IL 60611.

Swisher, J., and Hu, T. Alternatives to drug abuse: Some are and some are not. In: Glynn, T.; Leukefeld, C.; and Ludford, J., eds. *Preventing Adolescent Drug Abuse: Intervention Strategies*. National Institute on Drug Abuse Research Monograph 47. DHHS Pub. No. (ADM)83-1820. Washington, DC: Supt. of Docs., U.S. Govt. Print. Off., 1983.

Tobler, N. *Measuring Drug Use Differences from Pretest to Posttest: A Probit Change Score*. 1985. Available from the author, Box 246, Sand Lake, NY 12153.

Tobler, N. Meta-analysis of 143 adolescent drug prevention programs: Quantitative outcomes results of program participants compared to a control or comparison group. *J Drug Issues* 16(4):537-567, 1986.

Tobler, N. *Meta-Analysis of Adolescent Drug Prevention Programs: Final Report*. Rockville, MD: National Institute on Drug Abuse, 1992*a*.

Tobler, N. Drug prevention programs can work: Research findings. *J Addict Dis* 11(3):1-28, 1992*b*.

Tobler, N. Updated meta-analysis of adolescent drug prevention programs. In: Montoya, C.; Ringwalt, C.; Ryan, B.; and Zimmerman, R., eds. *Evaluating School-linked*

Prevention Strategies: Alcohol, Tobacco, and Other Drugs. San Diego, CA: UCSD Extension, University of California, 1993. pp. 71-86.

Tobler, N. Meta-analytical issues for prevention intervention research. In: Seitz, L., and Collins, L., eds. *Advances in Data Analysis for Prevention Intervention Research.* National Institute on Drug Abuse Research Monograph 142. NIH Pub. No. 94-3599. Washington, DC: Supt. of Docs., U.S. Govt. Print. Off., 1994. pp. 342-403.

Toseland, R., and Rivas, R. *Introduction to Group Work Practice.* New York: MacMillan, Inc., 1984.

United States Congress. House. *Anti-Drug Abuse Act.* Public Law 99-570, 99th Congress (H.R. 5484), October 27, 1986.

Wall, S.; Hawkins, J.; Lishner, D.; and Fraser, M. *Juvenile Delinquency Prevention. A Compendium of Thirty-Six Program Models*. National Institute of Juvenile Justice and Delinquency Prevention. Washington, DC: Supt. of Docs., U.S. Govt. Print. Off., 1981.

AUTHOR

Nancy S. Tobler, Ph.D.
Research Associate Professor
School of Social Welfare
University at Albany
State University of New York
Albany, NY 12150

Validity of Integrity Tests for Predicting Drug and Alcohol Abuse: A Meta-Analysis

Frank L. Schmidt, Vish Viswesvaran, and Deniz S. Ones

INTRODUCTION

The research described in this chapter used psychometric meta-analysis (Hunter and Schmidt 1990*b*) to examine the validity of integrity tests for predicting drug and alcohol abuse. Integrity tests have previously been found to predict other counterproductive workplace behaviors (e.g., absenteeism, property damage, and violence on the job) (Ones et al. 1993; Ones et al., unpublished observations). All studies located were concurrent in nature. For both drugs and alcohol, integrity tests correlated substantially (0.34 to 0.51) with admissions of abuse in student and employee samples. In samples of job applicants, however, the mean validity was lower (0.21) for drug abuse; validity for applicants was high for alcohol abuse, but only one study (N = 320) was found. All meta-analyses indicated that validity was generalizable. Based on these analyses, the authors conclude that the operational validity of integrity tests for predicting drug and alcohol abuse in the workplace is probably about 0.30. But further research is needed; predictive validity studies conducted on applicants would be particularly useful.

THE PROBLEM OF SUBSTANCE ABUSE

Substance abuse is a major societal problem. Numerous surveys (Johnston et al. 1994; Miller et al. 1983) have found that substance abuse, especially the abuse of alcohol and marijuana, is prevalent. Epidemiological surveys (Simpson et al. 1975) indicate that illicit drug abusers are predominantly young adults.

The relationships between substance abuse, job performance, and other job-related behaviors have been studied. In a large sample study of military personnel, McDaniel (1988) found that individuals who reported using drugs at earlier ages were more likely to be rated as unsuitable for service by their supervisors than a control group who indicated they did not use drugs when younger. In a sample of Navy recruiters, Blank and

Fenton (1989) found individuals testing positive for drugs had more behavioral and performance problems than individuals who tested negative for drugs.

Normand and colleagues (1990) found that postal employees who tested positive for substance abuse were more likely to be absent from work. Further, Winkler and Sheridan (1989) found that employees who entered employee assistance programs for drug addiction treatment were more likely to be absent, had twice the number of worker compensation claims, and used more than twice as many medical benefits as a matched control group. Crouch and colleagues (1989) found that drug use correlated with increased accident and absence rates.

Substance abuse has been found to be related not only to measures such as absenteeism, turnover, accidents, and productivity, but also to behaviors such as stealing on the job, violence, and effort expenditure (i.e., not daydreaming) on the job. In fact, Viswesvaran (1993) found that these various measures are positively correlated and a general factor exists across them, suggesting that the various measures of job performance may be influenced in part by the same underlying construct (presumably a personality dimension).

In addition to the above-mentioned studies that compare drug-using individuals to a matched set of controls on various job performance measures, laboratory studies have also found that substance abuse leads to impairment in performance of various experimental tasks (Herning et al. 1989; Jobs 1989; Streufert et al. 1991; Yesavage et al. 1985). Impairments in information-processing capabilities, decisionmaking, and quickness of reflexes have been found to result from drug or alcohol consumption.

With surveys indicating that abuse of alcohol and other drugs is prevalent in the general population and studies indicating a negative relationship between substance abuse and job performance, employers have tried different strategies to ensure a drug-free workplace. Coworkers, customers, and the general public also have a stake in ensuring a drug-free workplace. The growing concern of employers about drug abuse has resulted in increased testing of both current and prospective employees for drug abuse (Guthrie and Olian 1989).

A survey of the literature indicates that an employer’s choice of strategies in drug testing is mainly based on four considerations: the validity and reliability of the techniques used to detect substance abuse; the legal viability of the techniques; the practicality and cost of employing the techniques; and whether employees accept the use of a technique as justified.

Validity refers to whether the technique is measuring what it purports to measure. Reliability indicates whether the measurements are stable and replicable. Legal viability refers to the employers' concerns about whether the courts and arbitrators will accept the findings of the technique. In fact, studies have shown (Hill and Sinicropi 1987) that courts and arbitrators place considerable weight on the reliability and validity of the technique used in deciding cases involving substance abuse. Thus, the technique’s validity and reliability have an indirect as well as a direct effect on the strategies used by the employers to combat substance abuse.

Employee perceptions of a drug testing program’s acceptability have been widely researched. Negative employee reactions to drug testing, if ignored, may lead to lowered commitment and subsequent reduction in performance (Crouch et al. 1989). Knovsky and Cropanzano (1991) present data indicating that employee reactions to drug testing can be analyzed within an organizational justice framework (Adams 1965; Greenberg 1990). Specifically, Knovsky and Cropanzano (1991) found that perceptions of procedural justice affect reactions to drug testing. Two of the key elements in shaping perceptions of procedural justice are the validity, reliability, and psychometric properties of the testing procedures; and invasions of privacy concerns. Other elements include job characteristics, such as situations when impaired performance results in dangers to others (Stone and Vine 1989); the type of drug used (Murphy et al. 1990); the type of personnel action taken against employees testing positive (Gomez-Mejia and Balkin 1987; Stone and Kotch 1989); the role of explanations (Bies 1987; Bies and Shapiro 1987; Crant and Bateman 1989); the chance to appeal; the availability of advance notice; and whether random drug testing or testing with due cause is implemented. Employee objections could result in union contracts restricting the use of certain techniques for detecting substance abuse. Further, courts and arbitrators are likely to give some weight in their decisions to employee and applicant objections. Thus, employee acceptance has both direct and indirect effects (through legal acceptability) on the strategies used by an employer.

In short, the method’s validity and reliability affect legal defensibility of the procedures and acceptability to test takers, as well as directly affecting the employer’s choice of technique used. Further, validity and reliability affect employer strategies through an effect on legal defensibility and acceptability to test takers. Thus, it is of paramount interest to examine a procedure’s validity and psychometric properties to realize the benefits of drug testing without any loss of employee commitment.

Several approaches have been tried to detect drug abuse. Blood testing, breath analyzers, and urinalysis are some of the common approaches to drug testing and detection. One technique gaining prominence in employment settings is the use of paper-and-pencil preemployment integrity tests to assess a job applicant's predisposition to drug and alcohol abuse. Evidence available to date indicates that applicants do not object to such tests (Stone and Kotch 1989; Stone and Bommer 1990; Stecker and Rosse 1992). To the extent that selection methods can be used to eliminate drug abusers at the point of hire, drug testing programs for employees become less necessary.

INTEGRITY TESTS

Defining Integrity Tests

Integrity tests are designed to measure the predisposition of individuals to engage in counterproductive behaviors on the job. Integrity tests are paper-and-pencil tests, as opposed to other methods such as the polygraph (a physiological method), background investigations, interviews, and reference checks. These tests have been developed for use with applicants and employees (a normal population); hence instruments such as the Minnesota Multiphasic Personality Inventory (MMPI), which were designed for use with mentally ill populations, are not classified as integrity tests, even though some organizations claim to use them for screening out delinquent applicants. Most integrity tests have been initially designed to predict a variety of counterproductive behaviors; only later were they found to predict other criteria such as supervisory ratings of overall performance (Ones et al. 1993).

A Brief History of Integrity Tests

The first paper-and-pencil psychological test to assess the integrity of potential employees, the Personnel Reaction Blank, was developed in 1948 (Gough 1948). It was a derivative of what was then called the Delinquency Scale of the California Psychological Inventory. (This scale was later renamed the Socialization Scale.) In 1951 a second type of test, intended to assess honesty of job applicants, was developed. This test, the Reid Report, was a compilation of questions that seemed to distinguish honest and dishonest individuals during polygraph examinations. Since then several other instruments have been developed and used to select applicants on the basis of integrity. A complete treatise on the history of integrity tests can be found in Ash (1989) and Woolley (1991).

There is relatively little information about which companies use paper-and-pencil integrity tests. According to Sackett and Harris (1985), as many as 5,000 companies may use preemployment integrity tests, assessing about 5 million applicants yearly. A variety of surveys of companies indicate that anywhere between 7 percent to 20 percent of all companies in the United States could be testing for integrity, at least for some jobs (American Society for Personnel Administration 1988; Blocklyn 1988; Bureau of National Affairs, Inc. 1988; O'Bannon et al. 1989). Even by the most conservative estimates, millions of people in the United States either have been or are being tested using integrity tests. There are at least 43 integrity tests in current use. Of these tests, about one-quarter seem to be small operations without much market share; 16 to 19 tests overall seem to serve most of the demand for integrity tests. However, this demand can be expected to increase, because in 1988 the Federal Polygraph Act effectively banned the use of the polygraph in most employment settings.

Over the last 15 years, scientific interest in integrity testing has increased substantially. The publication of a series of literature reviews attests to the interest in this area and its dynamic nature (Guastello and Rieke 1991; Sackett et al. 1989; Sackett and Decker 1979; Sackett and Harris 1984). Recently Sackett and colleagues (1989) and O'Bannon and colleagues (1989) have provided extensive qualitative reviews and critical observations regarding integrity testing. In addition to these reviews, the U.S. Congressional Office of Technology Assessment (OTA 1990) and the American Psychological Association (APA) (Goldberg et al. 1991) have each released papers on integrity tests. The OTA paper (1990) was in part prompted by Congress' regulation of the polygraph. The OTA recommendations were based on a limited number of chosen studies and ignored most of the literature on integrity tests. Compared to the OTA paper, the APA report (Goldberg et al. 1991) was more thorough, objective, and insightful. It provided a generally favorable conclusion regarding the use of paper-and-pencil integrity tests in personnel selection.

Personality Constructs Underlying Integrity Tests

Sackett and colleagues (1989) classify honesty tests into two categories: overt integrity tests and personality-based tests. Overt integrity tests (also known as clear purpose tests) are designed to directly assess attitudes regarding dishonest behaviors. Some overt tests specifically ask about past illegal and dishonest activities as well; for several tests, admissions are not a part of the instrument, but instead are used as the criterion. Overt integrity tests include the London House Personnel Selection Inventory (PSI) (London House, Inc. 1975), Employee Attitude Inventory (EAI) (London House, Inc. 1982), Stanton Survey (Klump 1964), Reid Report (Reid Psychological Systems 1951), Phase II Profile (Lousig-Nont 1987), Milby Profile (Miller and Bradley 1975), and Trustworthiness Attitude Survey (Cormack and Strand 1970). According to Sackett and colleagues (1989), "[T]he underpinnings of all these tests are very similar" (p. 493). Hence, high correlations may be predicted, and are found (Ones 1993), among overt integrity measures.

On the other hand, personality-based measures (also referred to as disguised purpose tests) aim to predict a broad range of counterproductive behaviors at work (e.g., violence on the job, absenteeism, tardiness, drug abuse, theft) via personality traits such as reliability, conscientiousness, adjustment, trustworthiness, and sociability. In other words, these measures have not been developed solely to predict theft or theft-related behaviors. Examples of personality-based measures used in integrity testing include the Personal Outlook Inventory (Science Research Associates 1983), the Personnel Reaction Blank (Gough 1954), the Employment Inventory (Paajanen 1985), and Hogan's Reliability Scale (Hogan 1981). Different test publishers claim that their integrity tests measure different constructs, including responsibility, long-term job commitment, consistency, proneness to violence, moral reasoning, hostility, work ethics, dependability, and energy level (O'Bannon et al. 1989). The similarity of integrity measures raises the question of whether they all measure primarily a single general construct. Detailed descriptions of all the above tests can be found elsewhere (Conoley and Kramer 1989), particularly in the extensive literature reviews (O'Bannon et al. 1989; Sackett et al. 1989; Sackett and Harris 1984).

Using both primary data (N = 1,365) and meta-analytic cumulation, Ones (1993) found that a general factor exists across different integrity tests. Ones (1993) found that the variance common to all integrity tests correlated highest with the personality dimension of conscientiousness, followed by emotional stability (neuroticism) and agreeableness. Based on these comprehensive analyses, researchers can conclude that integrity tests tap into the personality dimensions of conscientiousness, agreeableness, and emotional stability. This finding is significant; researchers now can focus on the theoretical construct underlying the different measures rather than investigating each measure separately as if it were unique. All theoretical propositions and causal explanations are stated in terms of constructs and not measures (Nunnally 1978).

Review of Causal Mechanisms: Why Personality Constructs Underlying Integrity Tests Should Predict Substance Abuse

In the literature, three causal mechanisms have been proposed that explain why personality constructs tapped into by integrity tests should predict substance abuse. First, Barrick and colleagues (1994) found evidence for the hypothesis that highly conscientious individuals set higher (or more difficult) goals for themselves and strive to accomplish them. Barrick and colleagues (1994) argued that individuals who set more difficult goals for themselves exhibit better job performance.

Further, Schmidt and Hunter (1992) noted that highly conscientious individuals can be expected to spend more time on task, which also contributes to better job performance. However, high-level job performance is usually incompatible with substance abuse (McDaniel 1988; Normand et al. 1990). Thus, integrity tests that seem to be assessing conscientiousness (Ones 1993) may also correlate with, and predict, substance abuse.

A second explanation lies in the social impulse control enunciated by Gough (1948). According to this explanation, substance abusers are likely to be individuals who have not learned the social skills necessary to function effectively in society and often have poor impulse control. From this perspective, it could be argued that scores on integrity tests found to correlate with measures of neuroticism (emotional stability) (Ones 1993) should also correlate with measures of substance abuse.

Finally, Zuckerman (1983) and others have posited that individuals differ in their proclivity to seek sensations. Individual differences in sensation seeking are reflected in differing personality measures of extroversion and agreeableness. Integrity tests are correlated with agreeableness (Ones 1993) and therefore may be related to substance abuse.

METHODS

A thorough search was conducted to locate all existing integrity test validities. All published empirical studies were obtained from published reviews of the literature (O'Bannon et al. 1989; Sackett et al. 1989; Sackett and Harris 1984), three other meta-analyses of integrity tests (Harris, undated; McDaniel and Jones 1986, 1988), and a computerized search to locate the most recent studies in psychological and management- related journals.

According to O'Bannon and colleagues (1989), there are 43 integrity tests in use in the United States. All the publishers and authors of the 43 tests were contacted by telephone or in writing requesting validity, reliability, and range restriction information on their tests. Of these, 36 responded with research reports. In addition, the authors identified other integrity tests overlooked by O'Bannon and colleagues (1989); their publishers were also contacted. All unpublished and published technical reports reporting validities, reliabilities, or range-restriction information were obtained from integrity test publishers and authors. Some integrity test authors and test publishers responded to the request for validity information on their test by sending computer printouts that had not been written up as technical reports. These were included in the database.

Still other integrity test publishers responded by sending raw data that had not been analyzed. In some instances, using the information supplied, the authors were able to calculate the phi correlation, and then correct it for dichotomization (Hunter and Schmidt 1990*a*). These corrected correlations were used in the meta-analysis. Sample sizes for these corrected correlations were adjusted to avoid underestimating the sampling error variance. First, the uncorrected correlation and the study sample size were used to estimate the sampling error variance for the observed correlation. This value was corrected for the effects of the dichotomization correction, and this corrected sampling error variance was then used with the uncorrected correlation in the standard sampling error formula to solve for the adjusted sample size, which was entered into the meta-analysis computer program. This process results in the correct estimate of the sampling error variance of the corrected correlation in the meta-analysis. The list of integrity tests contributing criterion-related validity coefficients, reliabilities, or range restriction information to this meta-analysis is presented in table 1.

TABLE 1. *Tests contributing data to the meta-analyses.*

Test Name
Accutrac Evaluation System[a]
Applicant Review[a]
Compuscan[a,c]
Employee Attitude Inventory (London House)[a]
Employee Reliability Inventory[a]
Employment Productivity Index[b]
Hogan Personnel Selection Series (Reliability Scale)[b]

Integrity Interview[a]
Inwald Personality Inventory[b]
Orion Survey[a,c]
P.E.O.P.L.E. Survey[a]
Personnel Decisions Inc. Employment Inventory[b]
Personal Outlook Inventory[b]
Personal Reaction Blank[b]
Personnel Selection Inventory (London House)[a]
Phase II Profile[a]
P.O.S. Preemployment Opinion Survey[a,c]
Preemployment Analysis Questionnaire[a]
Reid Report and Reid Survey[a]
Rely[a]
Safe-R[a,c]
Stanton Survey[a]
True Test[a]
Trustworthiness Attitude Survey; PSC Survey; Drug Attitudes/ Alienation Index[a]
Wilkerson Preemployment Audit[a,c]

NOTE: The list of publishers and authors of these tests are available in O'Bannon et al. 1989.

KEY: a = Overt integrity test; b = personality-based integrity test; c = no validity data were reported, but the test contributed to the statistical artifact distributions.

Some researchers have argued for the exclusion of unpublished studies in all meta-analyses based on misleading and erroneous arguments that such unpublished studies constitute poor quality data. The converse argument maintains that published studies have a positive bias that overstates the results. Taken together, these two arguments lead to scientific nihilism (Hunter and Schmidt 1990*b*). The hypothesis of methodological inadequacy of unpublished studies (in comparison to published studies) has not been established in any research area. In fact, evidence exists in many research areas indicating comparability of findings of published and unpublished studies (Hunter and Schmidt 1990*b*).

Hunter and Schmidt (1990*b*) present a hypothetical example that illustrates how differences between published and unpublished studies examining the effectiveness of psychotherapy could have been due to statistical artifacts. Ones and colleagues (unpublished observations) found that the correlation between the reported validity of integrity tests and the dichotomous variable indicating published versus unpublished studies is negligible. In the literature on the validity of employment tests, impressive evidence has been accumulated indicating that published and unpublished studies do not differ in the validities reported (Hunter and Schmidt 1990*b*). For example, the data used by Pearlman and colleagues (1980) was found to be very similar to the Department of Labor's General Aptitude Test Battery (GATB) database used by Hunter (1983) and other large sample military data sets. The mean validities in Pearlman and colleagues' (1980) database are virtually identical to Ghiselli's (1966) reported medians. Further, the percentage of nonsignif-icant studies in Pearlman and colleagues' (1980) database perfectly matches the percent of nonsignificant published studies reported by Lent and colleagues (1971). Finally, the percentage of observed validities that were nonsignificant at the 0.05 level in Pearlman and colleagues' (1980) database (56.1 percent of the 2,795 observed validities) is consistent with the estimate obtained by Schmidt and colleagues (1976): The average criterion-related validation study has statistical power no greater than 0.50. If selectivity or bias in reporting were operating, many of the nonsignificant validities would have been omitted, and the percent significant should have been higher than 43.9 percent. On the other hand, if unpublished studies were of poorer quality, not meeting the standards of peer review, then there should have been more than 56 percent nonsignificant validities among the unpublished studies. Thus, there is ample evidence arguing for the equivalence of published and unpublished studies. The two databases are often comparable. Therefore, both published and unpublished reports are included.

Data Coded/Extracted From Primary Studies

An identification number was given to each study, and when more than one sample was reported in a study, a sample-within-study identification number was given to each sample within that study. Thus, each record contains a study identification number, a (within study) sample identifi-cation number, the validity coefficient, the sample size, the criterion used, whether the criterion measure was based on self-reports or external records, whether the sample was comprised of students or applicants for a job or current employees, and whether the validity coefficient was based on a predictive or a concurrent validation strategy. Wherever possible, the complexity levels of the jobs included in the analyses and other demographic characteristics were also coded.

Overall, 50 validation studies were located. Of these 50 studies, 24 had used employees as samples, 16 had used student samples, and the remaining 10 studies were based on applicant samples. All 50 studies employed the concurrent validation strategy. Forty-eight of the 50 studies had relied on admissions (self-reports) of substance abuse. There was one study conducted on a sample of 46 employees in a fire department that had used apprehension and conviction for substance abuse as the criterion. The observed validity coefficient in that study was 0.44. One study provided inadequate information as to whether admissions or external measures were employed. The observed validity coefficient in that study was 0.62, and it was based on a sample of 320 job applicants.

The admissions criterion was measured using self-report questionnaires. Measures of admissions of drug abuse included questions on number and type of illegal drugs used, number of times one had become high from drug use, and so forth. Measures of admissions of alcohol abuse included questions on frequency of alcohol intoxication, number of drinks consumed on the job, number of drinks on work breaks and during lunch on work-days, and number of alcohol-related problems. The final score was the sum (sometimes weighted) of such admissions.

Twenty of the 50 studies were conducted in the Midwest while 4 were conducted in the Northwestern region of the United States. Thirteen of the 50 studies were conducted in supermarket or grocery stores or convenience stores or on gas station employees. Seven of the 50 studies were done using security personnel as the sample. One study was conducted in a fire department while another was in a fast-food chain. Twenty studies focused on alcohol consumption while the remaining 30 used drug abuse as the criterion.

Given this set of validity coefficients, only two potential moderators could be tested: sample type (students, employees, and applicants) and criterion type (drug abuse versus alcohol abuse).

Intercoder agreement in summarizing or extracting information from the primary studies is a concern in meta-analyses. Haring and colleagues (1981) presented empirical data indicating that intercoder agreement in meta-analyses is a function of the judgmental nature of the items coded. Haring and colleagues' review of meta-analyses found that eight of the nine items lowest in coder agreement were judgments (e.g., the quality of the study) as opposed to calculation-based variables (e.g., effect sizes, number of subjects). Jackson

(1980) and Hattie and Hansford (1982, 1984) also provided data indicating that problems of intercoder agreement in meta-analyses are negligible for coding computation-based numerical variables. Finally, Whetzel and McDaniel (1988) found no evidence of any coder disagreements in validity generalization databases. The intercoder agreement in the present research was over 85 percent for all categories coded. Disagreements between the two coders were resolved through discussion.

Psychometric Meta-Analyses

Data from the sources described in the previous section were cumulated by the methods of psychometric meta-analyses (Hunter and Schmidt 1990*b*). Depending on the availability of information in the primary studies, the meta-analysis can either correct the observed correlations for the effects of statistical artifacts and cumulate the individually corrected correlations, use artifact distributions to correct the observed distribution of correlations, or use a combination of individual corrections and artifact distributions.

Because the degree of split for dichotomization was given in the research reports, it was possible to correct the correlations individually for the attenuating effects of dichotomization (Hunter and Schmidt 1990*a*). But to correct for the effects of artifacts such as unreliability and range restriction, where the available information was sporadic, recourse was made to the use of artifact distributions. That is, a mixed meta-analysis was employed. In the first step, the correlations were corrected individually for the effects of dichotomization. In the second step, the partially corrected distribution obtained from the first step was corrected for sampling error, unreliability, and range restriction using artifact distributions (Hunter and Schmidt 1990*b*).

In using artifact distributions for correcting two or more artifacts, one has the option to use either the interactive procedure (which corrects the observed correlations for the effects of the various statistical artifacts simultaneously), or the noninteractive procedure (which sequentially corrects the observed correlation for the effects of the statistical artifacts). Recent computer simulation studies (e.g., Law et al. 1994; Schmidt et al. 1993) have shown that among the methods of psychometric meta-analyses, the interactive procedure used with certain refinements (e.g., nonlinear range restriction and mean observed correlation in the sampling error formula) is the most accurate one.

The use of the mean observed correlation in the sampling error formula provides a more accurate estimate of the sampling error variance (Hunter and Schmidt 1994). The sampling error variance formula for the correlation requires knowledge of the population correlation. In individual studies, the observed correlation is taken as an estimate of the population value because nothing better is available. But meta-analysts can be more precise by using the mean observed correlation across studies. This value is a better estimate of the population correlation than the individual observed correlation, which is strongly affected by sampling error unless sample sizes are large.

The second refinement involves the use of a nonlinear range-restriction correction formula in estimating the standard deviation (SD) of true validities. In artifact distribution-based meta-analyses, the mean and SD of the residual distribution (the distribution of observed correlations expected when sample sizes are infinite and reliability and range-restriction values are held constant across studies at their mean values) are corrected for the mean value of the artifacts. This procedure is accurate when the artifact corrections are linear (e.g., reliability corrections) because the correction is the same for every value of the correlation in the residual distribution. But the correction for range restriction is not linear; it is smaller for large correlations and larger for smaller correlations. This results in an overestimation of the true SD when the linear approximation is used. Computer simulation studies have shown that a new, nonlinear correction procedure is more accurate (Law et al. 1994). That new procedure was used in this study. More details of these refinements can be found in Schmidt and colleagues (1993), where examples are also provided to illustrate application of the refinements.

In correcting for unreliability in the measures, the use of the correct form of reliability coefficient requires the specification of the nature of the error of measurement in the research domain of interest (Hunter and Schmidt 1990*b*). Several sets of artifact distributions were compiled: one distribution for the reliability of the integrity tests, one distribution for the reliability of the criterion variables, and one distribution of range restriction values. Descriptive information on the artifact distributions is provided in table 2.

TABLE 2. *Descriptive information on statistical artifact distributions used to correct validities.*

	N of	Mean	SD	Mean of	SD of

	values			the square roots of reliabilities	the square roots of reliabilities
Integrity test reliabilities	124	0.81	0.11	0.90	0.06
Criterion reliabilities	13	0.84	0.13	0.94	0.07
Range restriction values[c]	79	0.81	0.19	--	--

KEY: c = The ratio of the selected group standard deviation to the referent group standard deviation (s/S).

A total of 124 integrity test reliability values was obtained from the published literature and the test publishers. Of the 124, 68 were alpha coefficients (55 percent) and 47 were test-retest reliabilities over periods of time ranging from 1 to 1,825 days (mean = 111.4 days; SD = 379.7 days). The mean of the coefficient alphas was 0.81 (SD = 0.10) and the mean of the test-retest reliabilities was 0.85 (SD = 0.10). There were nine reliabilities reported with no statement of the type of reliability. The ideal estimate of test reliability for purposes of this meta-analysis is coefficient alpha or the equivalent. However, test-retest reliability estimates usually provide reasonably close approximations to alpha coefficients. In this case the means of the two reliability types were similar. The overall mean of the predictor reliability artifact distribution was 0.81 and the SD was 0.11. The mean of the square roots of predictor reliabilities was 0.90 with an SD of 0.06.

No correction for predictor unreliability was applied to the mean true validity because the interest was in estimating the operational validities of integrity tests for selection purposes. However, the observed variance of validities was corrected for variation in predictor unreliabilities in addition to variation in criterion unreliabilities, range restriction values, and sampling error. For comparison purposes, the authors provide the percent variance due to sampling error alone in the results.

To estimate the reliability of the criterion measure, the authors reviewed the literature on delinquency. Viswesvaran and colleagues (1992) meta-analyzed correlations between admissions and external measures of delinquency; the mean correlation was found to be 0.50. That study compiled a reliability distribution for questionnaires measuring admissions of delinquent acts. This distribution consisted of 13 values of coefficient alpha. The average of the reliability distribution was 0.84 and the SD was 0.13. The average of the square roots of the reliability estimates was 0.94 and the SD was 0.07. This distribution was used in the present study for admissions of alcohol and drug abuse.

Because integrity tests are used to screen applicants, the validity calculated using an employee sample may be affected by restriction in range. A distribution of range restriction values was constructed from the studies contributing to the database. There were 75 studies which reported both the SD in the study sample and the applicant group SD. The range restriction ratio was calculated as the ratio of study to reference group standard deviations (s/S). In four studies, correlations were reported for both the applicant and the employee groups. From these four studies, range restriction ratios were calculated by taking the ratio of the two correlations reported and solving for the range restriction value using the standard range restriction formula (case II formula, Thorndike 1949). Overall there were 79 range restriction values included in the artifact distribution. The mean ratio of the restricted sample SD to the unrestricted sample SD was 0.81 and the SD was 0.19; these figures indicate that there is considerably less range restriction in this research domain than is the case for cognitive ability (Alexander et al. 1989). Thus, range restriction corrections were much smaller in present research than in meta-analyses in the abilities domain. No range restriction corrections were made for student samples.

The parameters of interest estimated from a meta-analysis are the true validity, the SD of the true validity, and the 90 percent credibility value. From the observed distribution of validities, the authors estimated the distribution of true validities. There are four substantive inferences of interest here. First, the authors want to know the average validity coefficient across situations. This is captured in the mean true validity. Second, the authors want to know whether the validity coefficient will be positive across situations. To answer this question, the authors examined the 90 percent credibility value. The 90 percent credibility value indicates that in 90 percent of the situations, the validity coefficient will be higher than this value. As such, if the 90 percent credibility value is positive, one can conclude that the instrument has a validity coefficient that is positive in over 90 percent of the situations. That is, validity generalizes across situations.

The third substantive question involves an examination of the SD of true validities to examine the extent to which the validity varies across situations. In a meta-analysis, if the 90 percent credibility value is greater than zero but there is a sizable variance in the validities after corrections, it can be concluded that validities are positive across situations (i.e., validity generalizes), although the actual magnitude may vary across settings. However, the remaining variability may also be due to uncorrected statistical artifacts as well as methodological differences between studies. A final possibility is truly situationally specific test validities and/or the operation of moderator variables. In sum, the 90 percent credibility value is used to judge whether the validities are positive across situations (i.e., validity generalizes), whereas the estimated SD of true score validities is used to assess whether the estimated true validity is constant across situations.

Finally, to test for the moderating influence of a hypothesized moderator, the validity coefficients are grouped into subsets based on the hypothesized moderator. Psychometric meta-analyses are then conducted on each subset. If the hypothesized moderator exists, it will be reflected in the following findings: the mean true validity computed for each subset will vary across the subsets, and will vary from the mean true validity computed with the entire set of validities across subsets; and the average SD of true validities in the subsets will be lower than the overall SD. The above two results are interrelated as the group means and variances in the analysis of variance (ANOVA) paradigm, and together they test the extent of the moderating influence of the hypothesized moderator.

TABLE 3. *Meta-analyses of the validity of integrity tests for predicting substance (alcohol and drug) abuse.*

Analyses categories	Total N	K	r_{mean}	SD_r	s_{res}	$	SD$	% Var. S.E.	% Var. Total	90% CV
All samples	25,594	50	0.20	0.1175	0.0984	0.26	0.14	13.1	29.9	0.10
Employee samples	1,131	24	0.28	0.1290	0.0000	0.36	0.00	100.0	100.0	0.36
Applicant samples	22,091	10	0.17	0.0710	0.0538	0.22	0.07	08.5	42.5	0.13
Student samples	2,372	16	0.45	0.1440	0.1266	0.48	0.14	20.8	28.0	0.31

KEY: K = number of correlations; r_{mean} = mean observed correlation; SD_r = observed standard deviation; s_{res} = residual standard deviation; **$** = true validity; SD$ = true score standard deviation; % Var. S.E. = % variance due to sampling error; % Var. Total = % variance due to all corrected statistical artifacts; 90% CV = lower 90% credibility value.

RESULTS AND DISCUSSION

The results of the psychometric meta-analyses of integrity test validities for predicting overall substance abuse (alcohol and drug together) are presented in table 3.

Based on all 50 samples, the mean true validity is 0.26. Further, the 90 percent credibility value of 0.10 implies that the true validity will be greater than 0.10 in more than 90 percent of the situations. These values are based on a total sample size of 25,594. The SD of the true score validities is low (0.14), which suggests that perhaps alcohol and drug abuse can be conceptualized as manifestations of the same phenomenon of substance or chemical abuse. That is, one might hypothesize that the same personality characteristics might underlie both alcohol and drug abuse.

The separate mean true validities for student, employee, and applicant populations are also provided in table 3. In a selection setting, the focal population of interest is the applicant population. Many researchers have argued (see Ones et al. 1993 for a summary) that conscious and/or unconscious response distortion will affect integrity test validities. In taking these tests, applicants have the greatest incentive for response distortion, followed by employees and students in that order. That is, to the extent integrity test validities are affected by response distortion, true validities based on applicant samples should be lower than true validities based on employee samples, which in turn should be lower than the true validities computed on student samples.

The results reported in table 3 confirm this expected gradient. Although response distortion seems to attenuate the validity of integrity tests, its effects do not destroy validity. Even in the applicant population the true validity was 0.22 and the 90 percent credibility value was 0.13. Although this level of validity is moderate, these values suggest that the use of integrity tests in employment selection will translate into reduced levels of substance abuse in the workplace.

It is of interest to note that most of the sample consisted of applicants (about 90 percent). This is significant because applicants to jobs are the focus of interest. However, it would have been better if the applicant validities had been predictive in nature. The reader will recall that all validities in this meta-analysis are concurrent. The criterion for applicants was admissions of drug and/or alcohol abuse made at the time they were applicants. Use of this same criterion measure taken later (after participants had been on the job for some time) would have given a better indication of predictive validity. Since there may be less response distortion on the admissions criterion measure in predictive studies, predictive validity estimates might be higher than the 0.22 obtained here. (The authors return to this point later.)

Next, the authors analyzed the results of integrity tests for predicting alcohol abuse alone. The results are summarized in table 4.

The overall estimated true validity across 20 samples involving 1,402 individuals is 0.45 and the 90 percent credibility value is 0.29. The corresponding values in the employee samples were 0.34 and 0.34,

TABLE 4. *Meta-analyses of the validity of integrity tests for predicting alcohol abuse.*

Analyses categories	Total N	K	r_{mean}	SD_r	s_{res}	$	SD$	% Var. S.E.	% Var. Total	90% CV
All samples	1,402	20	0.35	0.1638	0.0966	0.45	0.14	41.2	63.0	0.29
Employee samples	644	16	0.27	0.1128	0	0.34	0	100.0	100.0	0.34
Applicant samples	320	1	0.62	--	--	--	--	--	--	--
Student samples	438	3	0.29	0.0125	0	0.31	0	100.0	100.0	0.31

NOTE: K = number of correlations; r_{mean} = mean observed correlation; SD_r = observed standard deviation; s_{res} = residual standard deviation; **$** = true validity; SD$ = true score standard deviation; % Var. S.E. = % variance due to sampling error; % Var. Total = % variance due to all corrected statistical artifacts; 90% CV = lower 90% credibility value.

respectively. All the observed variation in validities computed on employee samples was attributable to statistical and measurement artifacts. In the student samples, the mean true validity is 0.31 and the 90 percent credibility value is 0.31 (again, all the observed variation was explained by variations in statistical artifacts across the samples). There was only one study that used an applicant sample; in that study the observed validity coefficient was 0.62. Studies using employee samples and studies using student samples had similar levels of validity, implying that response distortion is not a serious problem in employee samples for the criterion of alcohol abuse. However, the key question is the extent to which there is response distortion among applicants; the data here are too thin to really answer this question.

The results for the integrity test validities for the criterion of drug abuse alone are summarized in table 5.

Across student, employee, and applicant populations there were 30 studies based on 24,192 individuals. Across these 30 studies, the overall true validity was 0.25 and the 90 percent credibility value was 0.10. The true validity was highest in student samples and lowest in applicant samples, indicating that response distortion may affect the operational validities of integrity tests for predicting the criterion of drug abuse. However, the same caveats apply here as in the case of alcohol abuse (table 4). Specifically, with admissions as the criterion measure, concurrent studies done on applicants may underestimate predictive validity computed on applicants. Concurrent studies done on applicants using admissions may strongly lend themselves to response distortion on the criterion measure, which in turn would bias validity estimates downward. Applicants for jobs have strong incentive to minimize

TABLE 5. *Meta-analyses of the validity of integrity tests for predicting drug abuse.*

Analyses categories	Total N	K	r_{mean}	SD_r	s_{res}	$	SD$	% Var. S.E.	% Var. Total	90% CV
All samples	24,192	30	0.19	0.1075	0.0909	0.25	0.13	10.0	28.4	0.10
Employee samples	487	8	0.30	0.1468	0.0561	0.38	0.08	64.5	85.4	0.29
Applicant samples	21,771	9	0.16	0.0456	0.0097	0.21	0	18.9	95.5	0.29
Student samples	1,934	13	0.48	0.1444	0.1280	0.51	0.15	19.3	21.5	0.34

KEY: K = number of correlations; r_{mean} = mean observed correlation; SD_r = observed standard deviation; s_{res} = residual standard deviation; $ = true validity; SD$ = true score standard deviation; % Var. S.E. = % variance due to sampling error; % Var. Total = % variance due to all corrected statistical artifacts; 90% CV = lower 90% credibility value.

admissions of previous illegal drug use. Present employees already have jobs, and in addition are usually told their responses will be used for research purposes only. So present employees have much less incentive for response distortion on the criterion. In contrast, response distortion on the predictor (test) does not bias estimates of operational predictive validity, because it reflects the reality that will hold when the test is used in hiring applicants. That is, real applicants will display some response distortion.

Given this likely downward bias in the mean true validity derived from concurrent studies done on applicants, the actual operational validity of integrity tests for predicting drug abuse is probably somewhere between the value of 0.21 and the value of 0.38 obtained from concurrent studies of incumbent employees. For prediction of alcohol abuse, the value corresponding to this 0.38 is 0.34. (No meta-analytic estimate of the value for applicant concurrent validity was possible for the criterion of alcohol abuse.) Hence, the operational validity of integrity tests for predicting the two types of substance abuse may be very similar. The authors would speculate that in both cases operational validity is around 0.30, a value large enough to produce practically significant reductions in substance abuse on the job if integrity tests are used in hiring.

Some limitations of the present study need to be pointed out. First, a fully hierarchical moderator analysis (Hunter and Schmidt 1990*b*) was not possible. In fact, even the main effects of some moderators could not be tested. For example, the authors could not compare the results of predictive and concurrent studies because there were no predictive studies. Also, there was only one study that used a criterion measure other than admissions of drug and/or alcohol abuse. Second, the number of existing studies was small enough in certain analyses to raise concerns about the stability of the estimates. Third, the type of study most relevant to answering questions about operational validity—predictive studies conducted on applicants—was absent from this research literature.

Any meta-analysis of test validities is limited by the number and type of available validation studies with particular criterion-predictor combinations. This has implications for second-order sampling error in meta-analyses (Hunter and Schmidt 1990*b*). But even with this limitation, a meta-analytic review based on a sound theoretical framework provides a better basis for conclusions than other approaches to understanding research findings, including the traditional narrative review. However, in this area, more research is needed. Predictive validity studies conducted on applicants would be particularly useful.

REFERENCES

Adams, J.S. Inequity in social exchange. In: Berkowitz, L., ed. *Advances in Experimental and Social Psychology*. New York: Academic Press, 1965.

Alexander, R.A.; Carson, K.P.; Alliger, G.; and Cronshaw, S.F. Empirical distributions of range restricted SD_x in validity studies. *J Appl Psychol* 74:253-258, 1989.

American Society for Personnel Administration. Most employers test new job candidates, ASPA survey shows. *Resource*, no. 6, 1988.

Ash, P. *The Construct of Employee Theft Proneness*. Rosemont, IL: SRA/London House, Inc., 1989.

Barrick, M.R.; Mount, M.K.; and Strauss, J.P. Conscientiousness and performance of sales representatives: A test of the mediating effects of goal setting. *J Appl Psychol* 79:272-280, 1994.

Bies, R.J. The predicament of injustice: The management of moral outrage. In: Cummings, L.L., and Staws, B.M., eds. *Research in Organizational Behavior*. Vol. IX. Greenwich, CT: JAI Press, 1987. pp. 289-319.

Bies, R.J., and Shapiro, D.L. Interactional fairness judgments: The influence of causal accounts. *Soc Justice Res* 1:199-218, 1987.

Blank, D.L., and Fenton, J.W. Early employment testing for marijuana: Demographic and employee retention patterns. In: Gust, S.W., and Walsy, J.M., eds. *Drugs in the Workplace: Research and Evaluation Data*. National Institute on Drug Abuse Research Monograph 91. DHHS Pub. No. (ADM)89-1612. Washington, DC: Supt. of Docs., U.S. Govt. Print. Off., 1989. pp. 139-150.

Blocklyn, P.L. Consensus: Preemployment testing. *Personnel* 2:66-68, 1988.

Bureau of National Affairs, Inc. *Recruiting and Selection Procedures, Personnel Policy Forum Survey No. 146*. Washington, DC: 1988.

Conoley, J.C., and Kramer, J.J., eds. *The Tenth Mental Measurements Yearbook*. Lincoln, NE: Buros Institute of Mental Measurements, 1989.

Cormack, R.W., and Strand, A.L. *Trustworthiness Attitude Survey*. Oakbrook, IL: Personnel Systems Corporation, 1970.

Crant, J.M., and Bateman, T.S. A model of employee responses to drug testing programs. *Employees Responsibilities and Rights J* 2:173-190, 1989.

Crouch, D.J.; Webb, D.O.; Peterson, L.V.; Butler, P.F.; and Rollins, D.E. A critical evaluation of the Utah Power and Light Company's substance abuse program: Absenteeism, accidents and costs. In: Gust, S.W., and Walsh, J.M., eds. *Drugs in the Workplace: Research and Evaluation Data*. National Institute on Drug Abuse Research Monograph 91. DHHS Pub. No. (ADM)89-1612. Washington, DC: Supt. of Docs., U.S. Govt. Print. Off., 1989. pp. 169-193.

Ghiselli, E.E. *The Validity of Occupational Aptitude Tests*. New York: Wiley, 1966.

Goldberg, L.R.; Grenier, J.R.; Guion, R.M.; Sechrest, L.B.; and Wing, H. *Questionnaires Used in the Prediction of Trustworthiness in Pre- employment Selection Decisions*. An APA task force report. Washington, DC: American Psychological Association, 1991.

Gomez-Mejia, L.R., and Balkin, D.B. Dimensions and characteristics of personnel manager perceptions of effective drug-testing programs. *Personnel Psychol* 40:745-763, 1987.

Gough, H.G. A sociological theory of psychopathy. *Am J Sociol* 53:359-366, 1948.

Gough, H.G. *Personnel Reaction Blank*. Palo Alto, CA: Consulting Psychologists Press, 1954.

Greenberg, J. Organizational justice: Yesterday, today, and tomorrow. *J Management* 16:399-432, 1990.

Guastello, S.J., and Rieke, M.L. A review and critique of honesty test research. *Behav Sci Law* 9:501-523, 1991.

Guthrie, J.P., and Olian, J.D. "Drug and Alcohol Testing Programs: The Influence of Organizational Context and Objectives." Paper presented at the Fourth Annual Conference of the Society of Industrial/ Organizational Psychology, Boston, April 1989.

Haring, M.J.; Okun, M.A.; Stock, W.A.; Miller, W.; Kinney, C.; and Ceurvorst, W.R. "Reliability Issues in Meta-Analyses." Paper presented at the annual meeting of the American Education Research Association, Los Angeles, April 1981.

Harris, W.G. *An Investigation of the Stanton Survey Using a Validity Generalization Model*. Charlotte, NC: Stanton Corporation, undated.

Hattie, J.A., and Hansford, B.C. Self-measures and achievement: Comparing a traditional review of literature with a meta-analysis. *Aust J Educ* 26:71-75, 1982.

Hattie, J.A., and Hansford, B.C. Meta-analysis: A reflection on problems. *Aust J Educ* 36:239-254, 1984.

Herning, R.L.; Glover, B.J.; Koeppel, B.S.; and Jaffe, J.H. "Cocaine and Workplace Performance: Inferences from Clinical Studies." Paper presented at the National Institute on Drug Abuse Conference on Drugs in the Workplace: Research and Evaluation Data, Bethesda, MD, September 1989.

Hill, M.F., Jr., and Sinicropi, A.V. *Evidence in Arbitration*. Washington, DC: BNA Books, 1987.

Hogan, R. *Hogan Personality Inventory*. Minneapolis: National Computer Systems Inc., 1981.

Hunter, J.E. *Test Validation for 12,000 Jobs: An Application of Job Classification and Validity Generalization Analysis to the General Aptitude Test Battery (GATB)*. Test Research Report No. 45. Washington, DC: U.S. Employment Service, U.S. Department of Labor, 1983.

Hunter, J.E., and Schmidt, F.L. Dichotomization of continuous variables: The implications for meta-analysis. *J Appl Psychol* 75:334-349, 1990*a*.

Hunter, J.E., and Schmidt, F.L. *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. Newbury Park, CA: Sage, 1990*b*.

Hunter, J.E., and Schmidt, F.L. The estimation of sampling error variance in the meta-analysis of correlations: Use of _r in the homogenous case. *J Appl Psychol* 78:171-177, 1994.

Jackson, G.B. Methods for integrative reviews. *Rev Educ Res* 50:438-460, 1980.

Jobs, S. "Impact of Moderate Alcohol Consumption on Business Decision Making." Paper presented at the National Institute on Drug Abuse Conference on Drugs in the Workplace: Research and Evaluation Data, Bethesda, MD, September, 1989.

Johnston, L.D.; O'Malley, P.M.; and Bachman, J.G. *National Survey Results on Drug Use from the Monitoring the Future Study, 1975-1993*. Vol. II. *College Students and Young Adults*. NIH Pub. No. 94-3810. Rockville, MD: National Institute on Drug Abuse, 1994.

Klump, C.S. *Stanton Survey*. Charlotte, NC: Stanton Corporation, 1964.

Knovsky, M.A., and Cropanzano, R. Perceived fairness of employee drug testing as a predictor of employee attitudes and job performance. *J Appl Psychol* 76:698-707, 1991.

Law, K.S.; Schmidt, F.L.; and Hunter, J.E. Nonlinearity of range restriction corrections in meta-analysis: A test of an improved procedure. *J Appl Psychol* 79:425-438, 1994.

Lent, R.H.; Aurbach, H.A.; and Levin, L.S. Predictors, criteria, and significant results. *Personnel Psychol* 24:519-533, 1971.

London House, Inc. *Personnel Selection Inventory*. Park Ridge, IL: London House Press, 1975.

London House, Inc. *The Employee Attitude Inventory*. Park Ridge, IL: London House Press, 1982.

Lousig-Nont, G.M. *Phase II Profile*. Las Vegas, NV: Self-published, 1987.

McDaniel, M.A. Does pre-employment drug use predict on-the-job suitability? *Personnel Psychol* 41:717-729, 1988.

McDaniel, M.A., and Jones, J.W. A meta-analysis of the employee attitude inventory theft scales. *J Bus Psychol* 2:327-345, 1986.

McDaniel, M.A., and Jones, J.W. Predicting employee theft: A quantitative review of the validity of a standardized measure of dishonesty. *J Bus Psychol* 2:327-345, 1988.

Miller, J.F., and Bradley, P. *Milby Profile.* Minneapolis: Milby Systems Inc., 1975.

Miller, J.D.; Cisin, I.H.; Gardner-Keaton, H.; Harrell, A.V.; Wirtz, P.W.; Abelson, H.I.; and Fishburne, P.M. *National Surveys on Drug Abuse: Main Findings*. Rockville, MD: National Institute on Drug Abuse, 1983.

Murphy, K.R.; Thornton, G.C.; and Reynolds, D.H. College students' attitudes toward employee drug testing procedures. *Personnel Psychol* 43:615-631, 1990.

Normand, J.; Salyards, S.D.; and Mahony, J.J. An evaluation of preemployment drug testing. *J Appl Psychol* 75:629-639, 1990.

Nunnally, J.C. *Psychometric Theory*. New York: McGraw Hill, 1978.

O'Bannon, R.M.; Goldinger, L.A.; and Appleby, G.S. *Honesty and Integrity Testing*. Atlanta, GA: Applied Information Resources, 1989.

Ones, D.S. "The Construct Validity of Integrity Tests." Ph.D. diss., University of Iowa, 1993.

Ones, D.S.; Viswesvaran, C.; and Schmidt, F.L. Meta-analysis of integrity test validities: Findings and implications for personnel selection and theories of job performance. *J Appl Psychol* 78:4, 1993.

Paajanen, G. *PDI Employment Inventory*. Minneapolis: Personnel Decisions, Inc., 1985.

Paajanen, G.E. "The Prediction of Counterproductive Behavior by Individual and Organizational Variables." Ph.D. diss., University of Minnesota, 1987.

Pearlman, K.; Schmidt, F.L.; and Hunter, J.E. Validity generalization results for tests used to predict job proficiency and training success in clerical occupations. *J Appl Psychol* 65:373-406, 1980.

Reid Psychological Systems. *Reid Report*. Chicago, IL: Reid Psychological Systems, 1951.

Sackett, P.R., and Decker, P.J. Detection of deception in the employment context: A review and critical analysis. *Personnel Psychol* 32:487-506, 1979.

Sackett, P.R., and Harris, M.M. Honesty testing for personnel selection: A review and critique. *Personnel Psychol* 37:221-246, 1984.

Sackett, P.R., and Harris, M.M. Honesty testing for personnel selection: A review and critique. In: Bernardin, H.J., and Bownas, D.A., eds. *Personality Assessment in Organizations*. New York: Praeger, 1985.

Sackett, P.R.; Brusis, L.R.; and Callahan, C. Integrity testing for personnel selection: An update. *Personnel Psychol* 42:491-529, 1989.

Schmidt, F.L., and Hunter, J.E. Development of a causal model of processes determining job performance. *Curr Dir Psychol Sci* 1:89-92, 1992.

Schmidt, F.L.; Hunter, J.E.; and Urry, J.E. Statistical power in criterion-related validation studies. *J Appl Psychol* 61:473-485, 1976.

Schmidt, F.L.; Law, K.; Hunter, J.E.; Rothstein, H.R.; Pearlman, K.; and McDaniel, M. Refinements in validation generalization methods: Implications for the situational specificity hypothesis. *J Appl Psychol* 78:3-12, 1993.

Science Research Associates. *Personal Outlook Inventory*. Parkridge, IL: Science Research Associates, 1983.

Simpson, D.D.; Curtis, B.; and Butler, M.C. Description of drug users in treatment: 1971-1972 DARP admissions. *Am J Drug Alcohol Abuse* 2(1):15-28, 1975.

Stecker, M., and Rosse, J. "Attitudes Toward Random Drug Testing in the Aviation Industry." Paper presented at the annual meeting of the Society for Industrial and Organizational Psychology, Montreal, April 1992.

Stone, D.L., and Bommer, W. "Effects of Drug Testing Selection Method and Justification Provided for the Test on Reactions to Drug Testing." Paper presented at the meeting of the Academy of Management, San Francisco, August 1990.

Stone, D.L., and Kotch, D.A. Individuals' attitudes toward organizational drug testing policies and practices. *J Appl Psychol* 74:518-521, 1989.

Stone, D.L., and Vine, P.L. "Some Procedural Determinants of Attitudes Toward Drug Testing." Paper presented at the Fourth Annual Conference of the Society of Industrial/Organizational Psychology, Boston, April 1989.

Streufert, S.; Pogash, R.; Campbell, L.; Kantner, A.; Gingrich, D.; and Jacques, C.H. "Effects of Alcohol Consumption upon Performance of White Collar Tasks." Paper presented at the National Institute on Drug Abuse Conference on Drug Abuse Research and Practice: An Alliance for the 21st Century, Washington, DC, January 1991.

Thorndike, R.L. *Personnel Selection*. New York: Wiley, 1949.

U.S. Congressional Office of Technology Assessment. *The Use of Integrity Testing for Pre-employment Screening*. (OTA-SET-442). Washington, DC: Supt. of Docs., U.S. Govt. Print. Off., 1990.

Viswesvaran, C. "Modeling Job Performance: Is There a General Factor?" Ph.D. diss., University of Iowa, 1993.

Viswesvaran, C.; Ones, D.S.; and Schmidt, F.L. Appropriateness of self-reported criteria for integrity test validation: Evidence from beyond. In: Schmidt, F.L., ed. *Criterion-related Validity, Construct Validity, and Legality of Integrity Tests*. Proceedings of the symposium conducted at the seventh annual meeting of the Society of Industrial and Organizational Psychologists, Montreal, Canada, May 1992.

Whetzel, D.L., and McDaniel, M.A. Reliability of validity generalization data bases. *Psychol Rep* 63:131-134, 1988.

Winkler, H., and Sheridan, J. "An Examination of Behavior Related to Drug Use in Georgia Power Company." Paper presented at the National Institute on Drug Abuse Conference on Drugs in the Workplace: Research and Evaluation Data, Bethesda, MD, September 1989.

Woolley, R. "An Examination of the Construct and Criterion-related Validity of Overt and Personality Oriented Predictors of Counter-productivity." Masters thesis, University of British Columbia, Vancouver, 1991.

Yesavage, J.A.; Leirer, V.O.; Denari, M.; and Holister, L.E. Carry-over effects of marijuana intoxication on aircraft pilot performance: A preliminary report. *Am J Psychiatry* 142:1325-1329, 1985.

Zuckerman, M., ed. *Biological Basis of Sensation-Seeking, Impulsivity, and Anxiety*. Hillsdale, NJ: Erlbaum Press, 1983.

AUTHORS

Frank L. Schmidt, Ph.D.
Professor

Vish Viswesvaran, Ph.D.
Professor

Deniz S. Ones, Ph.D.
Professor

College of Business Administration
University of Iowa
Iowa City, IA 52242

Meta-Analysis and Models of Substance Abuse Prevention

Betsy Jane Becker

INTRODUCTION

The idea of synthesizing available information about treatment efficacy or the strength of relationships among variables is not new. Procedures for combining such evidence date to the 1930s (Fisher 1932) and have been widely applied in the social sciences since Glass' introduction of meta-analysis in the 1970s (Glass 1976; Glass et al. 1981).

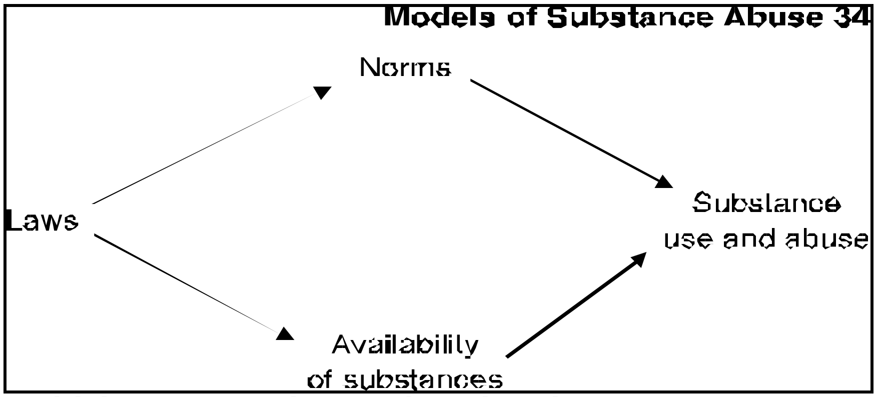
Recently reviewers in a number of disciplines have realized that research synthesis techniques can be applied in reviews of issues more complex than those previously studied. Meta-analysis has been criticized for attending only to main effects (Cook and Leviton 1980) and ignoring the important roles of mediating and moderating variables. Applications of meta-analytic techniques to complex processes (Becker 1992*b*; Premack and Hunter 1988), as well as methodological developments (Becker 1992*a*, 1992*c*), show that this oversimplification need not occur (see also Cook et al. 1992, p. 341).

This chapter introduces research synthesis methods for the analysis of complex processes and outlines how they can be applied in the study of the literature on substance abuse prevention. In particular, the chapter describes a model for the roles of risk and protective factors in substance abuse prevention, based on the review of Hawkins and colleagues (1992). The author next discusses how evidence about models could be gathered and examined in a quantitative synthesis of the literature on this topic, and describes key issues that arise in the application of this approach. A brief example of data analysis for a four-variable model is also presented. The chapter concludes with a discussion of how a model-based synthesis of risk and protective factors could be used in the design and analysis of substance abuse prevention programs.

MODEL-DRIVEN META-ANALYSIS

Model-driven meta-analysis refers to the quantitative synthesis of evidence pertaining to a model of the interrelationships among a set of constructs or variables. Often such models are illustrated using flowcharts or path diagrams. Flay and Petraitis (1991) showed two very detailed models of behavior that have served as theoretical frameworks for drug use behavior. Flay’s model focused primarily on the psychological antecedents of drug use, whereas Elliott and colleagues (1985) outlined a broader sociological model for delinquent behavior.

Figure 1 shows a simple diagram of the roles of three broad social context factors influencing substance abuse (variables are drawn from Hawkins et al. 1992). Models can show direct influences, such as the relationships of norms and availability to substance use and abuse shown in figure 1. Indirect relationships (mediated by other variables) can also be shown. Laws are depicted in figure 1 as having two indirect influences on abuse.



Model-driven meta-analysis is inherently multivariate. In contrast to narrative reviews and more limited syntheses of bivariate relationships, model-based meta-analysis can provide quantitative evidence about interactive effects of relevant variables. This should be particularly useful in a review of evidence on drug abuse, since "[T]here is little evidence available regarding the relative importance and interactions of various risk factors in the etiology of drug abuse" (Hawkins et al. 1992, p. 65). Similarly, Flay and Petraitis noted that, despite many reviews of correlates of drug use, "[T]here is no information about the relationships among the correlates" (1991, p. 82). Under certain assumptions, it may be possible to examine a model through meta-analysis that yields information about interactions not tested in any primary research study. Those assumptions are described more fully below.

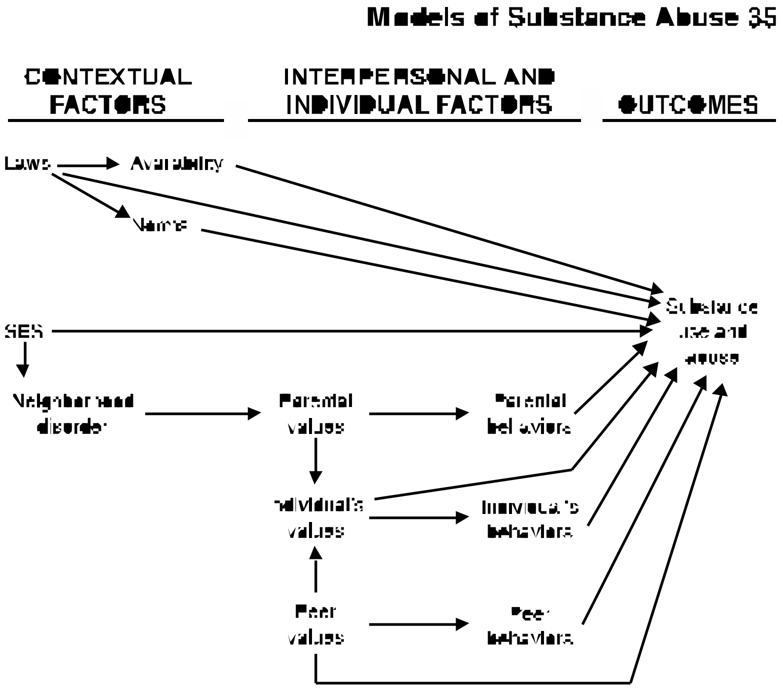
The models examined in a model-driven synthesis may arise empirically or be derived from theory. The theoretical-empirical distinction is rarely clear cut. Empirical research arises from implicit models of theory, and theory is often modified or even "discovered" by empirical work. An empirical model shows relationships that have been examined in primary research. This chapter describes an empirically derived model based on the narrative review of Hawkins and colleagues (1992). However, several authors (including Hawkins and colleagues) have noted the importance of a theoretical model or "conceptual framework for evaluating the content of substance abuse prevention curricula" (Hansen 1992, p. 408). Flay and Petraitis (1991) described 12 ways that theory is important in the area of substance abuse.

Theoretically derived models provide a context in which to assess the existence and the strength of evidence about a proposed model. Some parts of a theoretical model (e.g., hypothesized relationships) may be well studied, whereas others may never have been studied. These less studied (or unstudied) aspects of a model may be appropriate domains for further research. Clearly, it will be difficult or impossible to conduct a compre-hensive model-driven quantitative synthesis of a process if the bulk of the relationships proposed by the model have not been studied. Data requirements are discussed below.

Finally, model-driven meta-analyses can provide reviewers and policy-makers with information about processes that can help in practical decisions and program design. For instance, a review of the process of substance abuse may identify influences or combinations of influences that could be targeted in a substance abuse prevention program. Derivation of an empirical model may even allow the reviewer to test particular ideas about program features.

MODELS OF THE ROLES OF RISK FACTORS IN SUBSTANCE ABUSE

Models of the roles of risk and protective factors in substance abuse are implicit in the narrative review by Hawkins and colleagues (1992). Figure 2 shows one possible model that incorporates contextual factors and many of the individual and interpersonal factors described in the review.[1] The model shown in figure 2 has 11 broad predictors of substance use and abuse outcome for a total of 12 components. Table 1 lists those components.



Five components represent contextual factors, while the rest are interpersonal (parent and peer) and individual factors. The outcome itself is broadly defined, and leads to a good example of how such process models can be further delineated. For example, one could refine the model in figure 2 by focusing on drug abuse or on alcohol abuse. Some

TABLE 1. *Components in the model of substance abuse.*

Components	Examples
Laws Norms Availability Neighborhood disorder Socioeconomic status (SES)	
Peer values	Advocacy of drugs
Peer behavior	Drug and alcohol use, aggression, acceptance of individual
Parental values	Permissiveness towards drugs, educational aspirations for children
Parental behavior	Drug and alcohol use, hostility, marital dissolution, family conflict
Individual values	Attachment to parents, liking of school, educational expectations
Individual behavior	Delinquent behaviors, aggression, school performance, intellectual ability
Substance use and abuse	

predictive factors may be more relevant for one outcome than another; factors that are irrelevant to a particular outcome could be omitted from the refined model for that outcome.

The model in figure 2 shows 19 paths or connections between components. Both direct and indirect influences are outlined. Another way of refining the model is to change the paths shown in the model. For instance, all three "values" components have both direct and indirect connections to substance use and abuse. A different model might remove the three direct paths and show only indirect influences (i.e., those moderated by relevant variables). Moreover, this model does not show parent, peer, and individual behaviors. Such relationships may be important, but they are not direct paths to the outcome.

The model in figure 2 is certainly not the only possible model of the process described by Hawkins and colleagues. It is not an exact representation, since some of the component factors are very broad. However, it illustrates one process model that could be examined in a model-driven synthesis, and is based on empirical evidence.

EVIDENCE IN A MODEL-DRIVEN META-ANALYSIS

Existence of Research

The model in figure 2, having been derived from a narrative review of existing work, is empirical. Hawkins and colleagues cite evidence about many of the relationships shown in figure 2. Table 2 shows counts of the etiological studies reviewed by Hawkins and colleagues for each of the paths or components listed in table 1 and depicted in figure 2. Table 2 is an example of the first kind of evidence provided by a model-driven synthesis: existence of research on particular relationships.

These totals are based on each relationship described by Hawkins and colleagues (1992) and categorized according to the two components listed in table 1 that best matched the interrelated variables. Studies that examined several risk factors were included for each relationship studied. Original primary research was not consulted for this coding; decisions were made on the basis of the brief descriptions in Hawkins and colleagues' (1992) report. Different classification decisions could have been reached either with more information about the studies or by a coder more familiar with the literature on substance abuse.

Table 2 also includes studies that examined relationships on paths not depicted in figure 2; these counts are underlined. Five direct relationships not shown in figure 2 were examined in studies cited by Hawkins and colleagues (1992). Additionally, six entries represent relationships (denoted by asterisks) described as potentially important by Hawkins and colleagues and shown in figure 2, but not examined by any etiological studies in their table of results. Thus 6 of 19 paths, or nearly one-third of the paths in figure 2, are apparently unstudied. If this model is truly representative of the process of substance abuse development, research is needed to understand these paths in the model.

Several trends are apparent in table 2. First, the bulk of the studies mentioned in Hawkins and colleagues' (1992) table 1 looked at direct relationships of predictor variables to the substance use/abuse outcome. Of the 192 relationships counted, 149 (78 percent) involved substance use or abuse. Also, nearly half of the use/abuse relationships involved individual factors as predictors (i.e., the individual's own values and behaviors). Parental factors were mentioned next most frequently; 37 studies (roughly 25 percent) of use/abuse outcome examined parental values and behaviors as predictors. Finally, of the 43 instances in which the relationship did not involve the focal use/abuse outcome, over 90 percent (39 instances) were relationships in which the individual's behaviors (other than use/abuse) were the outcome.

Many of the possible entries in table 2 are simply empty. These empty positions represent paths that neither appear in figure 2 nor are mentioned by Hawkins and colleagues. As noted above, alternative models might include those other paths, and it is likely that studies not reviewed by Hawkins and colleagues (1992) included examination of those paths.

In an actual model-driven meta-analysis, thorough searches would be conducted to identify studies relevant to all paths in the model or models. Searches for model-driven meta-analyses often involve more extensive keyword lists and search strategies than more traditional meta-analyses or narrative reviews (Becker 1992*b*).

Analysis of Existing Data

Table 2 shows counts of studies that examined relationships relevant to the proposed model of substance abuse in figure 2. Many of the studies included in these counts probably presented their results in terms of indexes of association. In a quantitative synthesis of the evidence concerning the substance abuse model, the reviewer would retrieve and analyze these measures of association. Analyses of those measures provide the second type of evidence in a model-driven meta-analysis: evidence about strengths of relationships. These analyses are discussed in the following section.

ISSUES IN THE SYNTHESIS OF DATA

Cooper (1989) outlined five stages in the research synthesis process: problem formulation, data collection, data evaluation, data analysis, and reporting of results. Both problem formulation and data collection have been briefly discussed above. Problem formulation deals primarily with selecting or deriving a model or models to study. Data collection (gathering of studies) in a model-driven synthesis is likely to be more extensive than that for a traditional quantitative review, as mentioned above, because of the multivariate nature of model-driven syntheses.

Data Evaluation

Cooper’s third stage, data evaluation, involves retrieving study outcomes and coding study features such as study quality and characteristics of samples, measures, and, possibly, treatments. Coding study quality is at least as important in model-driven synthesis as it is in a more traditional meta-analysis. Also in a model-driven synthesis, the reviewer must code information relevant to the models being studied. For example, the studies reviewed by Hawkins and colleagues (1992) were classified according to the paths in the hypothetical model. This step would be crucial in a more extensive review because incorrect or careless classification could prompt critics to argue that dissimilar studies (apples and oranges) had been combined.

Between-Studies Differences.

Because meta-analyses have often been criticized for overlooking important between-studies variables, coding these variables is critical. Differences in samples (e.g., age or SES of subjects), in the nature and duration of treatments given, and in study quality can all lead to variation in results.

Variation in outcomes (e.g., the strength of relationship of particular predictors with substance abuse) sometimes can be explained by a small number of between-studies variables (that is, study-level covariates). Then fixed-effects models may apply, and the relevant study features may be moderator variables for one or several paths in a model.

In other cases, between-studies variation may not be accounted for even after many study characteristics have been examined. In these cases, random-effects models may be applied. Essentially, the reviewer expects some uncertainty or amount of variation across studies or uncertainty in the strengths of relationships studied. One assumes that different populations (or more precisely, populations with different correlation structures) may have been examined in different primary studies. The object is to estimate variability or uncertainty in the population correlations and to incorporate those estimates into further analyses of the data. The distinction between fixed and random models has both conceptual and statistical subtleties (for more information on fixed- and random-effects models see Hedges 1994). Both fixed- and random-effects approaches are available for the synthesis of model-based data (Becker 1992*c*).

Data Retrieval.

At the data evaluation stage, the reviewer retrieves correlational (associational) data from the primary research. While correlation indices such as Pearson's r, Spearman's rho, and the phi coefficient are often reported, many studies yield more complex data. Regression analyses, canonical correlation analysis, and path analyses may provide data on relationships of interest, but their results are not as easily synthesized as zero-order correlations. The object of data retrieval is to retrieve the same index of association from each study (for each relationship) or to convert the indices that are retrieved into values that are comparable across studies. Often the correlation is the most useful index (i.e., most easily made comparable).

Specific illustrations are easily found in the literature on substance abuse. Extensive research by Brook and colleagues (1983, 1986) has examined the correlates of adolescent drug use. However, though zero-order correlations of many predictors to the drug use outcome are presented in some studies (e.g., Brook and colleagues 1986), intercorrelations among the predictors are not given directly but are incorporated into canonical correlation analyses. Another format for presentation of information on drug use correlates is found in Brook and colleagues (1983). Drug users were first categorized by level of drug use, then mean values for each of the correlates were reported for each group of users.

An additional data retrieval issue concerns the measurement of the substance use and abuse outcome. If a study measures substance use as a dichotomy, typical measures of association that assume bivariate normality of both variables (e.g., Pearson's r) are inappropriate. However, it may be possible to convert more appropriate measures for association (given this dichotomy) into indices of the correlation between continuous variables that might underly the dichotomy. McDermott (1984) examined the associations among parental drug use (measured as use versus nonuse), attitude toward adolescent drug use (categorized as permissiveness versus disapproval), and adolescent drug use (also measured dichotomously); three 2 x 2 tables presented the categorical results. In a more complex analysis of a dichotomous alcohol use outcome, Barnes and Welte (1986) used discriminant analysis to relate more than 10 potential predictors to alcohol use. Indices from studies with dichotomous outcomes will also differ in their statistical properties from those based on continuous outcomes, such as level or amount of substance use. At present, the methodology for synthesizing model-based results for dichotomous outcomes has not been developed.

When different studies report results of analyses of different statistical models (i.e., models that control for different factors), they provide information about different partial relationships. Thus the slope for peer drug use from a regression of SES, parental drug use, and peer drug use on child’s use of drugs is not comparable to the slope for peer drug use when SES is the only other predictor. Analyses of structural models and regression models often pose this difficulty. Hansen and coauthors (1987) examined an elaborate hierarchical model of drug use using structural equation analyses. Their extensive results included reports of many path coefficients and residual correlations, but no zero-order correlations. Many other similar examples exist, suggesting that research on how to handle indices of partial relationships may be an area for further inquiry. Combining estimates of different parameters (e.g., of relations under different model specifications) is not sensible and is likely to yield inconsistent results in many circumstances. Combining zero-order indices avoids this confusion.

Data Analysis

For simplicity, temporarily assume that a reasonable number of studies have been gathered that examine all or parts of a proposed model. Further, assume that the studies provide zero-order correlation indices for the relationships studied. Several questions can then be posed about the relationships under study.

Procedures for analyzing correlational data in model-driven meta-analysis are described elsewhere (Becker 1992*c*; Becker and Schram 1994). The methods require that zero-order correlations be presented in each primary research study, or that they be retrievable from other study indices.

These methods enable the reviewer to ask, first, whether all studies show the same pattern of interrelationships among the variables in a correlation matrix (here, among the 12 components in the model). Then the reviewer can estimate a common correlation matrix (if studies appear similar) or a pooled matrix that accounts for between-studies variation in the correlation of values. Finally, either of these average matrices can be used to estimate standardized regression models showing the relative importance of the different predictors as well as intercorrelations among them. The reviewer can then piece together, from an entire literature, models similar to path-analytic (causal) models derived in single studies. The potential of these procedures to elucidate the nature of complex processes is tremendous. The approach has both strengths and weaknesses, however, as described below.

Availability of Data.

As described in the section on data evaluation, obtaining zero-order correlations or measures of association is necessary to apply currently available model-based synthesis methods. However, many studies do not present complete correlation matrices or indices of zero-order relationships.

Indices of partial relationships present problems of comparability, as discussed above. Missing (unreported) or nonexistent data cause statistical problems in estimation of average correlation matrices across studies (Becker 1992*a*). Data are considered missing, for instance, when a researcher reports correlations for a set of predictors with a substance use/abuse outcome but does not report intercorrelations among the predictors that also appear in the model under study.

If a proposed path in a model has not been studied, an average correlation for that path cannot be estimated. This may lead to a misspecified model if the omitted predictor is crucial. Estimated effects for studied variables may be biased if an important variable is omitted from the model. Such model misspecification can lead to incorrect conclusions, but may be difficult to avoid when using existing primary research. This problem highlights the importance of thoroughly searching for and collecting relevant studies.

Between-Studies Differences. Between-studies differences in study features as well as in the nature and extent of reported data may also present problems in a model-driven synthesis. Consider a very simple illustration by returning to the model in figure 1. Suppose that the search had identified 50 studies relevant to the four paths in figure 1, but that half of the studies examined adolescent drug use and half studied adults. Further suppose that these two groups of users are known to differ dramatically in many ways. If all of the studies of adolescents had examined the relationships of laws to norms and norms to use/abuse, and studies of adults had examined the remaining paths, it would not be possible to generalize about the entire model from the studies. Usually, the situation is not so clearly confounded as in this illustration.

If the reviewer is willing to apply a random-effects conceptualization to the model, however, some conclusions can be drawn. This is equivalent to arguing that, although the particular groups studied may represent different populations (e.g., of user types), there exists a "population of user populations" that is of interest. The task then is to determine how different the patterns of relationships appear to be in the populations being considered.

Artifactual Variation. Another source of between-studies differences in results that poses a problem in meta-analysis is artifactual variation. This can include such influences as differential reliability of measures (even if identical constructs have been studied) and restriction of range. For instance, results based on samples drawn from a single population can differ if one sample is unselected and the other is composed of high scorers (e.g., selected on the basis of an employment selection test or other similar instrument).

Corrections for both unreliability and range restriction are readily available for a series of single (bivariate) correlations (e.g., Schmidt and Hunter 1994). However, until recently the effects of applying these corrections to correlation matrices had not been studied. Schram (1995) examined a variety of methods for correcting correlation matrices for attenuation due to unreliability, and found that the familiar univariate correction performed well. Schram also derived a large-sample variance-covariance estimator for the corrected matrix that incorporates uncertainty due to the estimation of both the correlations and the reliability coefficients. While the reviewer may not have access to complete information about artifacts, it is important to acknowledge that artifactual variation can lead to variation in observed results.

Causality. When models are used in the planning of substance abuse programming, there is an implicit assumption that manipulation of relevant predictor values can lead to changes in substance abuse. Essentially, program planners are looking for potential causal relationships. Strong inferences of causality require both temporal precedence of the cause relative to the effect and elimination of other competing explanations of change in the outcome. Cook (1990, 1991) has written extensively about causality in meta-analysis and program evaluation.

An Example

To illustrate the possibilities for quantitative synthesis of correlational data, an example is presented of a synthesis of results from three samples. These three samples all arise from a single study by Mills and Noyes (1984). This is an overly simplistic example that avoids issues such as differential unreliability, comparability of constructs, and range restriction that might arise in a more realistic example.

The three samples are of 8th, 10th, and 12th graders from Maryland public schools. Four "use" variables are examined, two of which will be treated as predictors (smoking and use of alcohol) and two as outcomes (use of marijuana and cocaine). Methods used are described in Becker (1992*c*) and Becker and Schram (1994).

Table 3 shows the upper halves of the correlation matrices for the three grades. Each sample provides six correlation values. The first task is to ask whether the three sets of correlations arise from a single population. If so, a single pooled correlation matrix can adequately represent relationships in all the samples.

Example. The test of whether a single population correlation matrix applies to the three grades is a chi-square test with (3-1) x 6 = 12 degrees of freedom. For the data in table 3 the value is 25.64, which is significant at $p < 0.025$. The results to not appear to be completely consistent with the model of a single underlying population correlation structure. Thus a random-effects model can be adopted and an average correlation matrix can be estimated.

Estimating Variation in Population Correlations. In order to incorporate the uncertainty or variation in correlation strength that results from having samples from several populations with different correlation structures, the variances (and covariances) among the population correlations must first be estimated. Becker and Schram (1994) describe how the estimation and maximization (EM) algorithm can be applied to obtain these variance component estimates. For each relationship, an estimate is obtained of the variation in the population values of correlations representing that relationship.

For example, let $r_{SA(i)}$ represent the correlation between levels of smoking and alcohol use in study i. Then $\$_{SA(i)}$ is the corresponding population

TABLE 3. *Correlation matrices from Mills and Noyes (1984).*

	Smoking	Alcohol	Marijuana	Cocaine
		Grade 8 (N = 672)		
Smoking	1.00	0.48	0.50	0.20
Alcohol		1.00	0.50	0.24
Marijuana			1.00	0.37
Cocaine				1.00
		Grade 10 (N = 691)		
Smoking	1.00	0.43	0.54	0.25
Alcohol		1.00	0.52	0.27
Marijuana			1.00	0.43
Cocaine				1.00
		Grade 12 (N = 589)		
Smoking	1.00	0.34	0.38	0.23
Alcohol		1.00	0.43	0.20
Marijuana			1.00	0.42
Cocaine				1.00

correlation. The model for a single correlation value under random effects shows that

$$r_{SA(i)} = \$_{SA(i)} + e_{SA(i)} ,$$

and thus variation (uncertainty) in the sample values of r_{SA} incorporates variation in the $\$_{SA(i)}$ values and sampling variance. The variance component for the smoking/alcohol use correlations is an estimate of variation among the $\$_{SA(i)}$ values. Similarly, covariances are estimated among the population correlations. If $\$_{SM(i)}$ represents the correlation of smoking with marijuana use in population i, the covariance component for $\$_{SA}$ and $\$_{SM}$ would be Cov ($\$_{SA(i)}$, $\$_{SM(i)}$) across populations.

Example. For the data from Mills and Noyes (1984), the EM algorithm produced a variance covariance matrix (denoted T) for the six correlation indices of

$$\begin{array}{c|cccccc|} & \$_{SA} & \$_{SM} & \$_{SC} & \$_{AM} & \$_{AC} & \$_{MC} \\ \hline \$_{SA} & .0027 & .0028 & -.0003 & .0016 & .0010 & -.0007 \\ \$_{SM} & .0028 & .0038 & .0001 & .0022 & .0015 & -.0003 \\ \$_{SC} & -.0003 & .0001 & .0003 & .0000 & .0001 & .0003 \\ \$_{AM} & .0016 & .0022 & .0000 & .0013 & .0009 & -.0002 \\ \$_{AC} & .0010 & .0015 & .0001 & .0009 & .0007 & -.0000 \\ \$_{MC} & -.0007 & -.0003 & .0003 & -.0002 & -.0000 & .0004 \end{array} = T.$$

Variances are shown on the diagonal, and the covariances are the off-diagonal elements of T. The standard deviations of the six populations of correlation values are 0.052, 0.062, 0.017, 0.036, 0.025, and 0.020. The second correlation, representing the relationship between smoking and marijuana use, shows the most variation. A standard deviation of 0.062 would correspond to a normal distribution ranging roughly from -0.20 to 0.20, if centered on zero. Even this is not a broad range for correlations.

Estimating the Average Correlation Matrix. Once an estimate of variation in the population correlations has been obtained, it can be incorporated in the estimation of an average correlation matrix. The estimate of the mean correlation matrix is obtained via generalized least squares (GLS) estimation (Becker 1992*c*). The GLS estimates can be obtained under fixed- and random-effects models. Covariation among the several correlations from each sample is accounted for in both cases. In the random-effects model, variation and covariation in population effects are also incorporated into the uncertainty of the estimates.

The random-effects GLS estimate of the mean correlation matrix for the three samples is

	Smoking	Alcohol	Marijuana	Cocaine
Smoking	1.00	0.42	0.48	0.23
Alcohol use	0.42	1.00	0.49	0.24
Marijuana use	0.48	0.49	1.00	0.41
Cocaine use	0.23	0.24	0.41	1.00

.

Comparing this estimate with the original data matrices in table 3 shows that the values of the sample correlations of smoking with alcohol use (r$_{SA}$) and with marijuana use (r$_{SM}$) indeed vary more about these means than the other correlation values, as suggested by the variance components in T above.

The variance-covariance matrix for the set of six average correlations is

	r_$_{SA}$	r_$_{SM}$	r_$_{SC}$	r_$_{AM}$	r_$_{AC}$	r_$_{MC}$
r_$_{SA}$	0.0012	0.0010	-0.0000	0.0007	0.0004	-0.0002
r_$_{SM}$	0.0010	0.0016	0.0002	0.0008	0.0006	-0.0001
r_$_{SC}$	-0.0000	0.0002	0.0005	0.0001	0.0002	0.0003
r_$_{AM}$	0.0007	0.0008	0.0001	0.0007	0.0004	-0.0000
r_$_{AC}$	0.0004	0.0006	0.0002	0.0004	0.0007	0.0002
r_$_{MC}$	-0.0002	-0.0001	0.0003	-0.0000	0.0002	0.0005

.

As could be expected from the amount of variation in the population values, the averages of the first two correlations, r_$_{SA}$ and r_$_{SM}$, show the most uncertainty, with standard errors of 0.035 and 0.040, respectively. These are still quite small relative to the magnitudes of the average correlations, however, which are both about 0.40.

Estimating Linear Models. Once an estimate of a mean correlation matrix has been obtained, it can be used to estimate a variety of predictive models for the intercorrelated variables. Here two standardized regression models are estimated. The first incorporates smoking and alcohol use as predictors of marijuana use.

The estimated model for this regression (based on the random-effects mean correlations and their variance, given above) is

^M = 0.33 S + 0.35 A,

where ^M represents a predicted standardized (z) score on the marijuana use scale, S is a z score for level of smoking, and A is a z score for level of alcohol use. The slopes, their standard errors, and tests of the hypothesis = 0 for each slope are given in table 4. Both slopes differ significantly from zero at very stringent levels. It is also possible to test whether the two slopes (say, $_S$ and $_A$) are equal, using their variances and the estimated covariance between b_S and b_A. The test of H_0: $_S =$ $_A$ uses the statistic:

which has a standard normal distribution when H_0 is true. Since _ z _ œ 1.96, it is not significant at the = 0.05 level. It can be concluded that both level of smoking and level of alcohol use are significant, and equally strong, predictors of marijuana use.

The second model examines smoking, alcohol use, and marijuana use as predictors of cocaine use. The estimated model is

^C = 0.03 S + 0.05 A + 0.37 M,

where ^C is a z score for level of cocaine use and S, A, and M are as described above. Table 4 shows that the only significant predictor in this

TABLE 4. *Standardized regressions showing contributions to substance abuse.*

	Outcome					
	Marijuana use			Cocaine use		
Predictor	b	SE(b)	z	b	SE(b)	z
Smoking	0.33	0.028	11.79*	0.03	0.039	0.82
Alcohol use	0.35	0.033	10.61*	0.05	0.049	0.92
Marijuana use	---	---	---	0.37	0.048	7.73*

NOTE: All predictors and outcomes represent standardized scores on the four "use" variables.

KEY: * = significant slope coefficients.

model is (standardized) level of marijuana use. Levels of smoking and alcohol use do not predict level of cocaine use for these 8th through 12th graders.

Display of Regression Results. Figure 3 shows the results of the standardized regression analyses displayed on a flow diagram similar to those in figures 1 and 2. The slopes are entered on the paths in the model, and significant slopes are starred. This model shows that across the three grades there are direct relationships between smoking and marijuana use, alcohol and marijuana use, and marijuana use and cocaine use. The effects of smoking and alcohol use on cocaine use are only indirect (i.e., mediated by level of marijuana use). According to the tests in table 4, the two paths (from S to C and from A to C), representing direct effects of smoking and alcohol use on cocaine use, could be eliminated.

Summary of Example. This example indicates the possibilities for analyses when results from multiple studies (here samples) are combined using techniques for model-driven quantitative synthesis. Tests of homogeneity (consistency) of results indicate whether fixed- or random-effects models are most appropriate. Average correlation matrices can be inspected for their own intrinsic value or used to obtain estimates of the

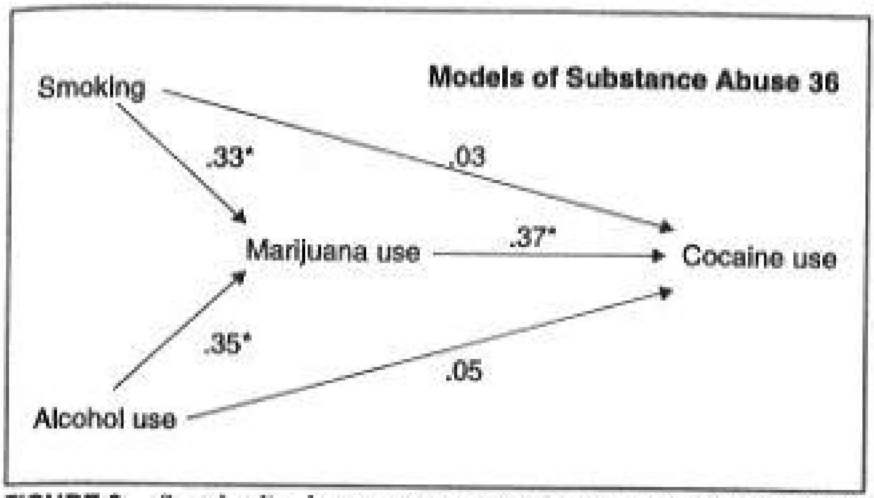


FIGURE 3. *Standardized regression results with smoking and alcohol use as predictors of marijuana and cocaine use.*

simultaneous relationships of several predictors to each outcome. These analyses can then inform the reviewer about the plausibility of a variety of models of relationships among variables.

USING MODELS FOR PROGRAM PLANNING

Most researchers and practitioners dealing with substance abuse prevention use both theory and empirical research in program planning. Reviews of school-based abuse curriculums (e.g., Hansen 1992) and other prevention programs (Tobler 1986, 1992) emphasize these ideas. Flay and Petraitis (1991) also discuss the importance of theory for program planning.

Hansen (1992) devoted nearly one-fifth of a review to the conceptual underpinnings of curriculum content for school-based programs, and described "the building block theoretical concepts used by researchers" and the "theoretical or quasi-theoretical assumptions about the means by which [program] components affect behavior" (1992, p. 408). Hansen's framework "provides a description of programmatic approach linked to mediating process" (1992, p. 408). A quantitative model-driven meta-analysis can provide an empirical assessment of proposed models such as those described by Hansen.

Tobler's two reviews (1986, 1992) also describe mediating processes underlying program strategies (or modalities). Table 1 in each article describes the assumptions of five program strategies. For instance, peer programs assume that "peer pressure can impact attitudes and behaviors" (1992, p. 6). In the model shown in figure 2, these assumptions could refer to the peer values, individual values, individual behavior predictors, and the substance use/abuse outcome. The premises underlying Tobler's two types of peer programs involve different beliefs about the kinds of individual responses (behaviors) that can inhibit drug use. Model-based meta-analyses with sufficient data can support detailed comparisons of those program types, or of their assumptions. Such comparisons may aid in the refinement of existing program designs or the development of new programs that incorporate strategies that seem to work better in combination than in isolation.

Status Studies Versus Intervention Studies. One question the reviewer must address in conducting a model-based meta-analysis is whether to include both intervention studies and "status studies" in which no manipulation of variables is attempted. If both are included, it will be important to examine differences between the results of the two kinds of studies. The presence of an intervention could attenuate relationships seen in a one-group status study (of the same relationship) by making subjects appear to be more similar on the manipulated variable than they would naturally be. Alternately, if an intervention differentiates subjects (e.g., by making them more variable on coping skills, self-esteem, or knowledge of the effects of drug use), a study of that intervention may show a stronger relationship of the manipulated variable to substance use and abuse than a status study. Thus, intervention studies may not present the same view of the potential effectiveness of intervention strategies as status studies.

Comparisons of Program Models. To be most useful for program planning, a model-driven meta-analysis should contrast and compare different process models. Does a model that includes components for both peer and parental behaviors explain more variation in substance use and abuse than one dealing with peers only? Is attention to the individual’s values necessary to understand levels of substance use/abuse? These questions imply different process models and different program designs.

Perhaps parental behaviors explain considerable variation in child drug abuse, but securing parental interest and participation in a substance abuse prevention program may be both difficult and costly. With a model-driven synthesis, such practical questions about program design can be weighed in light of concrete evidence about differences in process models.

Decisions about which models can or should be compared can be based on theory or on the need to make specific decisions about program components.[2]

CONCLUSION

This chapter illustrates the potential of model-driven quantitative synthesis for exploring and testing models of the influences on substance abuse, and for providing information for substance abuse prevention program planning. The application of these ideas in a thorough empirical review of the literature provides an exciting possibility for future work.

NOTES

1. Physiological factors have been omitted from this model. Other parts of the model are greatly simplified by creating very broad categories (e.g., "behaviors" and "values"). Other more differentiated models (e.g., specifying and separating particular behaviors) are possible.

2. Clearly, if the collection of studies for the meta-analysis does not include data on the models of interest, such comparisons will be impossible. The above discussion assumes that sufficient data are, in fact, available.

REFERENCES

Barnes, G.M., and Welte, J.W. Patterns and predictors of alcohol use among 7-12th grade students in New York State. *J Stud Alcohol* 47(1):53-62, 1986.

Becker, B.J. "Complications and Corrections in Constructing Synthetic Regressions." Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1992*a*.

Becker, B.J. Models of science achievement: Factors affecting male and female performance in school science. In: Cook, T.D.; Cooper, H.; 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*b*.

Becker, B.J. Using results from replicated studies to estimate linear models. *J Educ Stat* 17:341-362, 1992*c*.

Becker, B.J., and Schram, C.M. Examining explanatory models through research synthesis. In: Cooper, H., and

Hedges, L.V., eds. *The Handbook of Research Synthesis*. New York: Russell Sage Foundation, 1994.

Brook, J.S.; Whiteman, M.; and Gordon, A.S. Stages of drug use in adolescence: Personality, peer, and family correlates. *Dev Psychol* 19:269-277, 1983.

Brook, J.S.; Whiteman, M.; Gordon, A.S.; and Cohen, A.S. Some models and mechanisms for explaining the impact of maternal and adolescent characteristics on adolescent stage of drug use. *Dev Psychol* 22(4):460-467, 1986.

Cook, T.D. Meta-analysis: Its potential for causal description and causal explanation within program evaluation. In: Albrecht, G., and Otto, H., eds. *Social Prevention and the Social Sciences: Theoretical Controversies, Research Problems, and Evaluation Strategies*. New York: Walter de Gruyter, 1991.

Cook, T.D. The generalization of causal connections: Multiple theories in search of clear practice. In: Sechrest, L.; Perrin, E.; and Bunker, J., eds. *Research Methodology: Strengthening Causal Interpretations of Nonexperimental Data.* Conference proceedings, U.S. Dept. of Health and Human Services, May 1990.

Cook, T.D., and Leviton, L.C. Reviewing the literature: A comparison of traditional methods with meta-analysis. *J Pers* 48:449-472, 1980.

Cook, T.D.; Cooper, H.; 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.

Cooper, H.M. *Integrating Research: A Guide for Literature Reviews.* Beverly Hills, CA: Sage, 1989.

Elliott, D.S.; Huizinga, D.; and Ageton, S.S. *Explaining Delinquency and Drug Use.* Beverly Hills, CA: Sage Publications, 1985.

Fisher, R.A. *Statistical Methods for Research Workers.* 4th ed. London: Oliver and Boyd, 1932.

Flay, B.R., and Petraitis, J. Methodological issues in drug use prevention research: Theoretical foundations. In: Leukefeld, C.G., and Bukoski, W.J., eds. *Drug Abuse Prevention Intervention Research: Methodological Issues.* National Institute on Drug Abuse Research Monograph 107. DHHS Pub. No. (ADM)91-1761. Washington, DC: Supt. of Docs., U.S. Govt. Print. Off., 1991.

Glass, G.V. Primary, secondary, and meta-analysis of research. *Educ Res* 5(10):3-8, 1976.

Glass, G.V.; McGaw, B.; and Smith, M.L. *Meta-Analysis of Social Research.* Beverly Hills, CA: Sage, 1981.

Hansen, W.B. School-based substance abuse prevention: A review of the state of the art in curriculum, 1980-1990. *Health Educ Res* 7(3):403-430, 1992.

Hansen, W.B.; Graham, J.W.; Sobel, J.L.; Shelton, D.R.; Flay, B.R.; and Johnson, C.A. The consistency of peer and parent influences on tobacco, alcohol, and marijuana use among young adolescents. *J Behav Med* 10(6):559-579, 1987.

Hawkins, J.D.; Catalano, R.F.; and Miller, J.Y. Risk and protective factors for alcohol and other drug problems in adolescence and early adulthood: Implications for substance abuse prevention. *Psychol Bull* 112:64-105, 1992.

Hedges, L.V. Statistical considerations in problem formulation. In: Cooper, H., and Hedges, L.V., eds. *The Handbook of Research Synthesis.* New York: Russell Sage Foundation, 1994.

McDermott, D. The relationship of parental drug use and parents' attitude concerning adolescent drug use to adolescent drug use. *Adolescence* 73:89-97, 1984.

Mills, C.J., and Noyes, H.L. Patterns and correlates of initial and subsequent drug use among adolescents. *J Consult Clin Psychol* 52(2):231-243, 1984.

Premack, S.L., and Hunter, J.E. Individual unionization decisions. *Psychol Bull* 103:223-234, 1988.

Schmidt, F.L., and Hunter, J.E. Correcting for sources of artifactual variation across studies. In: Cooper, H., and Hedges, L.V., eds. *The Handbook of Research Synthesis*. New York: Russell Sage Foundation, 1994.

Schram, C.M. "A Comparison of Methods for Correcting Multivariate Data for Attenuation, with Application to Synthesizing Correlation Matrices." Ph.D. diss., Michigan State University, 1995.

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 Issues* 16(4):537-567, 1986.

Tobler, N.S. Drug prevention programs can work: Research findings. *J Addict Dis* 11(3):1-28, 1992.

AUTHOR

Betsy Jane Becker, Ph.D.
Professor
Michigan State University
456 Erickson Hall
East Lansing, MI 48824-1034

Realities of the Effect Size Calculation Process: Considerations for Beginning Meta-Analysts

Patricia D. Perry

INTRODUCTION

This brief, nontechnical chapter is intended for individuals who have not conducted a meta-analysis, but may be considering doing so. The goal of the chapter is to alert potential meta-analysts to unanticipated hazards they may encounter during the effect size calculation process. It is hoped that awareness of these hazards may assist individuals to assess the extent to which these hazards apply to their particular field of study, to implement plans to minimize the effect of these hazards, and to document the extent to which these hazards were encountered.

Meta-analysis has become a well-accepted method of conducting a quantitative literature review. The intuitive appeal of meta-analysis comes from the belief that findings from multiple studies may provide a more stable and meaningful measure of the magnitude of a treatment effect than results from a single study. Findings from individual studies can be aggregated if their quantitative results are transformed into a standardized difference between a treatment and a control group (i.e., an effect size). While the concept of an effect size is straightforward, the meta-analyst must rely on data provided by other researchers to calculate an effect size estimate. The quality and quantity of those data can vary as a function of the subject matter for the meta-analysis, editorial practices of specific journals, and individual study methodologies.

Since the ability to calculate an effect size is dependent on the way in which the primary researchers conducted and reported their respective studies, the ideal data from which to calculate an effect size may not be available in all studies. Alternatively, the data may be reported in such detail that the meta-analyst has several options available for calculating an effect size estimate. These two complementary issues and their corresponding hazards are the focus of this chapter. The implications of these issues to the planning and implementation of a meta-analysis are also described.

LACK OF SPECIFIC INFORMATION FROM WHICH TO ESTIMATE AN EFFECT SIZE

Ideally, the posttest means and standard deviations (SDs) on a given outcome measure for a treatment and a control group, along with their corresponding sample sizes, are included within individual research reports. However, frequently one or more of these summary statistics is missing for a particular outcome measure, and a meta-analyst must use alternative data to calculate an effect size estimate. Several examples follow.

No Information About Nonsignificant Results

Outcomes with nonsignificant (NS) findings are frequently described within the narrative of the report, or designated as "NS" within a table. A common practice in meta-analysis is to record an effect size of zero for a nonsignificant outcome. However, the magnitude of nonsignificance could impact the aggregated effect size for that particular outcome measure within a meta-analysis. For example, suppose effect sizes from three studies are used to determine the average effect for a given outcome. Study A and study B reported nonsignificant findings, while study C reported data that yielded an effect size estimate of 0.20. The average effect size from these studies would be 0.07 (assuming all other variables to be equal among the studies). By contrast, suppose that studies A and B reported specific means and SDs for each outcome, even though the findings were nonsignificant within their respective studies. If studies A, B, and C had effect size estimates of 0.07, 0.05, and 0.20 respectively, the average effect size would be 0.11 (e.g., 0.32/3). This simplistic example is intended to demonstrate that estimating values, when precise values are missing, can affect the meta-analytic finding. (See Hedges and Olkin 1985 for a discussion on the consequences of observing only significant effect sizes.)

Rounded Probability Levels

When effect sizes are calculated from levels of significance (i.e., probability of type I error), more precise effect size estimates can be obtained if the actual probability level is reported. An effect size calculated from a significance level of $p < 0.05$ is different from one in which the probability level was reported as p =

0.35. The magnitude of difference in the effect size estimate will depend on the degrees of freedom, but in principle, the more specific probability level will produce a better estimate of the population effect size than an estimate calculated from a rounded value.

Calculating Effect Size Estimates From Values Within a Graph

Data tables are often one of the best sources of information for a meta-analysis. By contrast, data contained within a graph are usually imprecise. The purpose of a graph is to show a trend (or trends) within the data. When graphs are small, the individual data points that produce the trend are difficult to quantify, and the meta-analyst is forced to rely on a best guess for the actual data points that comprise the graph. It can be a frustrating experience to ponder a graph, knowing that the information is before one’s eyes, yet be unable to reproduce the exact value obtained by the original researcher.

Sample Size Not Reported

The sample sizes of the treatment and control groups are used to calculate the inverse of an effect size variance, which, in turn, can be used to weight the effect size estimates according to their respective sample sizes (Hedges 1986). When precise sample sizes are missing from individual reports, they can be obtained by contacting the original researcher, estimated by simple division of the total sample size by the number of groups in the study, or estimated from the degrees of freedom. It is surprising how many research reports state a total sample size but do not report the sample size for the treatment and comparison groups, or for the subgroups within the treatment and comparison groups (e.g., males and females).

The four examples previously described were encountered by the author of this chapter during the effect size calculation process in Tobler’s meta-analysis of adolescent drug prevention programs (Tobler 1993). Table 1 lists the proportion of studies in which all the data were available to calculate an effect size estimate for a single drug use outcome measure[1] (i.e., no estimated values were used) and the proportion of studies in which at least one component of the effect size was estimated.

TABLE 1. *Sources of estimated values for effect size calculations from Tobler's 1993 meta-analysis of 120 studies.*

Source of estimated value	Proportion of studies (N)
Finding reported as nonsignificant	16% (19)
Rounded probability level	8% (10)
Outcome value obtained from a graph	10% (12)
Other (misc.)	3% (4)
None; all values reported	63% (75)

TABLE 2. *Source of derivation of sample size from Tobler's 1993 meta- analysis of 120 studies.*

Source of sample size derivation	Proportion of studies (N)
Sources of estimated sample size	
Total sample size divided by the number of groups	19% (23)
Estimated from degrees of freedom	9% (11)
Sources of exact sample size	
No estimation; all sample sizes given	57% (68)
Primary researcher contacted	15% (18)

lists the proportion of studies in which the sample size was estimated (e.g., by degrees of freedom or by dividing the total sample size by the number of groups) and the proportion of studies for which the actual sample size was known. Data from tables 1 and 2 indicate that estimating a value required for effect size calculation was common. While it is certainly preferable to have actual values for effect size calculation, those values simply may not be available from the individual studies that comprise the meta-analysis.

SURPLUS OF INFORMATION FROM WHICH TO CALCULATE AN EFFECT SIZE ESTIMATE

While the lack of information from which to calculate an effect size can be frustrating for a meta-analyst, inclusive information presents a different set of issues and options that deserve consideration. The meta-analytic literature contains multiple references regarding effect size calculation methods (Hedges and Olkin 1985; Hunter and Schmidt 1990; Glass et al. 1981; Lund 1988; McGaw and Glass 1980; Rosenthal 1991; Seifert 1991; Thomas 1986). Indeed, it is the variety of methods available to the meta-analyst that complicates selection of effect size calculation methods.

For example, within Tobler's meta-analysis, 15 different effect size calculation methods were utilized depending on the summary statistics reported within the individual research studies. These methods included:

- Raw posttest means and standard deviations;
- Between groups independent t-test;
- Two-group F statistic;
- One-way analysis of variance (ANOVA) with three groups (omnibus F statistic);
- Chi-square between two groups;
- Correlated t from gain scores;
- Dependent t (matched pairs);
- Raw gain score;
- Level of statistical significance (probability);
- Repeated measure ANOVA;
- Proportions;
- Probit transformation of percentage (Tobler 1989);
- Probit change scores;
- Probit transformation of posttest percentage rates; and
- Regression coefficients.

The extent to which each of these methods was used depended on their frequency of use within the individual studies of interest. Given the multiple options for effect size calculation available to the meta-analyst, frequently more than one method could be used to calculate a specific effect size from a given study. The meta-analyst should be sure that all of the effect size estimates are estimating the same parameter.

An example of this issue is when an F statistic, derived from analysis of covariance, is used to calculate an effect size from an individual study. The F statistic resulting from analysis of covariance incorporates prior information in the final analysis. In other words, the F statistic represents the difference between two groups adjusting for some preexisting differences between groups. For example, the effectiveness of a specific teaching strategy might be evaluated by giving a pretest and a posttest. The individual researcher may want to control for differences in the outcome that may have been present at the pretest. The pretest would be considered a covariate and variance in the outcome attributable to the pretest would be accounted for in the summary statistic. An effect size calculated from an F statistic derived from analysis of covariance would be a more precise estimate of the population effect size (e.g., effective- ness of the teaching strategy) when compared to an effect size calculated from an F statistic derived from the more simple ANOVA (in which differences in outcome due to preexisting differences were not considered).

When an analysis of covariance F statistic is reported, the meta-analyst has the option of selecting two methods for calculating an effect size estimate. One method would produce a more accurate estimate of the population effect size for that particular study (i.e., by using the covariate-adjusted F statistic to calculate an effect size estimate), while the other method would modify the F statistic to estimate differences between the groups without such adjustment (Smith et al. 1980). Thus, two different effect sizes could be computed for a specific outcome, each one representing a different effect size concept.

The previous example was included to demonstrate that calculating an effect size estimate may not be a simple, straightforward procedure. There are many options available to the meta-analyst for calculating an effect size, and decisions about the effect size calculation process can affect the final meta-analytic findings. When the subject for meta-analysis contains a set of studies in which the degree of effectiveness of the treatment is computed variously, the meta-analyst must consider the extent to which comparable effect sizes can be derived from this set of studies.

IMPLICATIONS FOR PLANNING AND IMPLEMENTING A META-ANALYSIS

The planning and implementation of a meta-analysis is a sophisticated task. There is a great deal of technical information that must be understood regarding sampling error, sampling bias, aggregation of effect sizes, and so forth (see Cook et al. 1992; Cooper and Hedges 1994; Hunter and Schmidt 1990). It is imperative that the beginning meta-analyst be familiar with the different meta-analytic methods available, and proceed to select a method that is compatible with the study objective and the particular field from which the literature will be reviewed. In order to do this, the meta-analyst must be familiar with two sets of literature: the field of study for the meta-analysis, and the meta-analytic literature.

Several steps can be taken during the planning and implementation of the meta-analysis to monitor the extent to which the hazards described in this chapter may be present. First, the meta-analyst should develop a set of inclusion criteria to determine whether an individual study will or will not be in the analysis. While many factors need to be considered in developing the set of inclusion criteria (e.g., date of publication, type of outcome measurement, type of research design), it is important that the type of primary data available from which to calculate an effect size be incorporated into this set of criteria. This process requires that the meta-analyst be cognizant of potential effect size calculation methods early in the literature review process. Indeed, inability to calculate an effect size is a reasonable criterion for excluding a study from the meta-analysis.

Second, a meta-analyst must decide how restrictive the inclusion criteria will be. For example, must all studies have the exact sample sizes reported, or will an estimate of the sample size be acceptable? The criteria for inclusion should be stated so other meta-analysts and consumers of the meta-analysis will be informed regarding potential sources of error variance in the effect size estimate.

Third, documentation of effect sizes calculation methods should be built into the meta-analytic process (if more than one procedure was utilized). It is important to document how the effect size estimates were calculated so that the method of calculation can be examined vis-a-vis outcome. For example, if better effect size estimates resulted from one effect size calculation method, the results could be examined to determine whether the effect size calculation method itself introduced an artifact that affected the meta-analytic results. Fourth, consideration should be given to calculating and recording an effect size using alternative methods. Using the analysis of covariance example cited earlier, a meta-analyst could calculate a covariate and a noncovariate adjusted effect size estimate. This would enable the meta-analyst to conduct a general meta-analysis (with all studies represented) and a subanalysis of studies that reported a covariate-adjusted summary statistic. The degree of concordance between the two analyses could be informative regarding the strength of treatment effect.

Finally, the meta-analyst is in a unique position to monitor the methodological state of the art for a particular field of interest. If an abundance of studies within the literature lack scientific rigor, the meta-analyst is well placed to discuss such issues. The meta-analyst can also document the extent to which reporting practices lack specificity. For example, data within tables 1 and 2 suggest that the primary research for Tobler’s (1993) meta-analysis contained meaningful reporting deficiencies that affected the meta-analytic process. One can only speculate the extent to which the results of Tobler’s meta-analysis might have differed if more precise information had been available. Recommendations for improvement in research methodology and editorial practices are an important outcome of a meta-analytic investigation.

CONCLUSION

This chapter was designed to alert the beginning meta-analyst to a few potential hazards that could be encountered during the effect size calculation process. The extent to which these hazards will be experienced depends on the subject and scope of the meta-analysis. Indeed, one way of minimizing these hazards is to limit the set of studies in the meta-analysis by creating strict inclusion criteria. However, when the purpose of the meta-analysis is broad in scope, excluding studies from the meta-analysis defeats that goal. There is often a tenuous balance between creating inclusion criteria that enhance the validity of the meta-analysis (a factor that tends to limit the number of studies included) while maintaining the goal of a comprehensive review (a factor that supports including numerous studies).

The beginning meta-analyst has much to consider. The meta-analyst must not only be familiar with the field of study, but also possess sufficient competence in statistical analysis to recognize and address the unique issues that arise from quantitatively combining individual research reports (which can vary in almost every aspect of research design). The intuitive attraction of conducting a meta-analysis (i.e., the attempt to summarize the literature from a quantitative perspective) must be attenuated by an appreciation for the complexity of the process by which the meta-analytic findings are generated. Without such an appreciation and a willingness to conduct an indepth study of the statistical procedures associated with calculating and aggregating effect size estimates, the results of the meta-analysis are likely to be spurious and uninterpretable.

NOTES

1. Many studies had more than one drug use outcome measure. The data in table 1 are derived from a single outcome measure in each study and do not represent the entire set of effect sizes in Tobler's meta-analysis.

REFERENCES

Cook, T.; Cooper, H.; Cordray, D.; Hartmann, H.; Hedges, L.; Light, R.; Louis, T; and Mosteller, F., eds. *Meta-Analysis for Explanation: A Casebook.* New York: Russell Sage Foundation, 1992.

Cooper, H., and Hedges, L., eds. *The Handbook of Research Synthesis.* New York: Russell Sage Foundation, 1994.

Glass, G.V.; McGaw, B.; and Smith, M.L. *Meta-Analysis in Social Research.* Beverly Hills, CA: Sage Publications, 1981.

Hedges, L.V. Advances in statistic methods for meta-analysis. In: Cordray, D.S., and Lipsey, M.W., eds. *Evaluation Studies Review Annual: Vol. 11*. Newbury Park, CA: Sage, 1986. pp. 731-748.

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

Hunter, J.E., and Schmidt, F.L. *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. Newbury Park, CA: Sage, 1990.

Lund, T. Some metrical issues with meta-analysis of therapy effects. *Scand J Psychol* 29:1-8, 1988.

McGaw, B., and Glass, G.V. Choice of metric for effect size in meta-analysis. *Am Educ Res J* 17(3):325-337, 1980.

Rosenthal, R. Meta-analysis: A review. *Psychosom Med* 53:247-272, 1991.

Seifert, T.L. Determining effect sizes in various experimental designs. *Educ Psychol Meas* 51:341-347, 1991.

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

Thomas, H. Effect size standard errors for the non-normal non-identically distributed case. *J Educ Stat* 11:293-303, 1986.

Tobler, N.S. *Measuring Drug Use Differences from Pretest and Posttest Change Scores.* Unpublished manuscript. Available from author, Box 246, San Lake, NY 12153. 1989.

Tobler, N.S. "Updated Meta-Analysis of Adolescent Drug Prevention Programs." Paper presented at the conference on Evaluating School-linked Prevention Strategies: Alcohol, Tobacco, and Other Drugs, University of California, San Diego, March 17-20, 1993.

ACKNOWLEDGMENTS

This chapter was prepared with support from the National Institute on Drug Abuse grant #R01DA05505-2.

AUTHOR

Patricia D. Perry, Ph.D.
Independent Consultant
P.O. Box 9545
Niskayuna, NY 12309

Issues and Challenges in Coding Interventions for Meta-Analysis of Prevention Research

Elizabeth C. Devine

INTRODUCTION

Meta-analysis is the statistical analysis of a large collection of results from individual studies for the purpose of integrating findings (Glass 1976). In other words, it is a quantitative review of existing research in a substantive area, involving multiple tests of a common hypothesis. When applied to intervention research, meta-analysis can be helpful in determining whether multiple tests of an intervention yield effects on an outcome construct of interest that are similar in direction and magnitude. Although the concept and some of the statistics used in meta-analysis date from the 1930s (Hedges and Olkin 1985), tremendous strides in the acceptance and popularity of meta-analysis have occurred in the last 15 years (Chalmers 1991; Myers 1991).

Many of the challenges facing meta-analysts arise from the fact that they are restricted to investigating what has been studied previously in primary research. Unless meta-analysts obtain additional information from individual primary researchers, they are further limited by the information provided in the research reports. Meta-analysts lack the control that primary researchers have to specify the population to be studied, the interventions to be tested, and the outcome measures to be used. In addition, it is rare to find exact replications within a body of research. Even studies of the same basic hypothesis may have noteworthy differences in sampling and operationalization of the intervention and outcome constructs.

Faced with what can aptly be described as "lumpy data," the meta-analyst must make many decisions, and possibly revise those decisions, as the extent and limitations of the existing data are discovered. Like primary researchers, meta-analysts must make many judgments that help to determine the final product. These include what to study, the source of data to use, the final selection criteria for the review, what to measure and how to operationalize the constructs of interest, who should collect and code the data, what analyses to perform, and how to report results. In both meta-analysis and primary research, the research decisions to be made far outnumber the calculations to be performed. The prevalence of choices, judgments, and compromises commonly made in meta-analysis (Nurius and Yeaton 1987), as well as their influence on the outcome of the meta-analysis (Wanous et al. 1989), have been discussed to a limited extent in the meta-analysis literature.

Purpose

The primary purpose of this chapter is to discuss the issues and challenges involved in one of the major judgment areas in meta-analysis, that of coding interventions. The focus is on interventions typically found in field research, such as the evaluation of drug abuse prevention programs and health care-related interventions.

BACKGROUND

Presby (1978), an early critic of meta-analysis, noted that combining overly broad categories of interventions can obscure important differences between treatments. Similar cautions can be raised for creating overly broad categories of subjects or outcomes. With this in mind, and without the control to insure that there are sufficient numbers of studies in all of the potential subcategories of interest, the meta-analyst must determine the selection criteria, the coding categories, and the grouping of studies for analysis. The decisions made should be based on the populations and constructs that are the target of intended generalization. There are no simple, "canned" programs for making these choices; many different decisions are possible. In fact, many different constellations of decisions may be justified and yield useful results, assuming that they are based on the current knowledge in the field and are consistent with the objective of the review. For example, reviews of the same general content area may look quite different depending on whether the primary purpose of the review is to inform professional practice, to test theory, or to influence policy.

In the early literature on meta-analysis, combining studies with multiple differences in a single analysis was often referred to as the "apples and oranges" problem (Glass et al. 1981). At one extreme, critics saw meta-analysis as hopeless mishmash. At the other extreme, proponents saw it as a way to learn about constructs that include both apples and oranges (e.g., fruit salad). In other words, meta-analysis may provide a way to identify whether certain phenomena are stubbornly replicable. That is, do they occur across many studies despite minor differences in subjects, settings, measures, or interventions? However, as the meta-analyst strives for general conclusions about phenomena, there is a nagging question that must never be far from the meta-analyst's mind: How many differences (e.g., in subjects, outcomes, or interventions) can be tolerated before the analyses obscure meaning rather than inform?

Over the years there have been dramatic changes in the way differences in outcomes have been treated in meta-analyses. In the very early years of meta-analysis, it was not uncommon to see all the effect size values from all dependent variables combined in a single unweighted mean. The outcome would be termed something global like "well-being." Meta-analysis has come a long way from those beginnings. For example, in recent years there is general consensus that only one effect size value should represent a study in any analysis. This helps to ensure the statistical independence of the data. It is also much more widely accepted that only effect size values from measures of the same construct should be aggregated (Hedges and Olkin 1985).

There have not been such dramatic changes in the way meta-analysts treat differences among subject characteristics. However, for several reasons, aggregating across studies of subjects with different characteristics often presents somewhat fewer conceptual or practical problems. First, researchers are accustomed to subjects with different characteristics being included in a single study. Second, subject characteristics often are better reported and thus are easier to code than treatment characteristics. Third, if the studies in the meta-analysis include relevant information on the subgroups of interest to the meta-analyst, it is fairly easy to disaggregate studies according to subject characteristics and determine if the pattern of results is consistent across relevant subpopulations of interest.

There has been less discussion of, and there is less consensus about, coding and aggregating interventions. In any topical area there are many potential consumers of a meta-analysis and many different purposes for doing a review of existing research. Even among high-quality meta-analyses, it should not be surprising if meta-analyses of similar topics vary widely in the coding and aggregating of interventions. Some of the issues meta-analysts must deal with as they decide about coding interventions are discussed below.

CODING INTERVENTIONS

There are at least four major areas to be considered related to coding interventions. To facilitate discussion, these areas are presented in a linear fashion. However, in the meta-analyst's reality, they present themselves like a bowl of jelly that jiggles all over no matter where it's touched.

The first decisions relate to what should be coded about experimental interventions. These are followed closely by decisions about when and by whom experimental interventions should be coded. Third are the decisions about how to minimize bias in the coding of experimental interventions. And finally, since the essence of each experimental intervention is defined by the ways in which it is different from the control condition, decisions must be made about what to code about the control condition.

What To Code About the Experimental Intervention

With regard to the experimental intervention itself, the meta-analyst must decide what and how much information about the experimental intervention is useful to code, and how to categorize and aggregate experimental interventions.

There are no simple answers to these questions. In addition to considering the purpose of the meta-analysis, it is essential to consider the analyses that are planned, the size of the research base of studies meeting the selection criteria, and the variability among interventions that have been tested in the literature under review.

What To Code. When one is summarizing and analyzing the results obtained from a group of studies that are construct replications (Lykken 1968) rather than exact replications, the actual content of interventions will most likely have been operationalized in many different ways. This may be particularly true of interventions such as drug abuse prevention, counseling, or patient education that are proposed to ameliorate specific (and often complex) social, interpersonal, or health-related problems. The variability of content typically included in drug abuse prevention programs is illustrated by the classifications of curriculum content developed and used by Hansen (1992) in a review of 45 drug abuse prevention curriculums. Hansen identified 12 content areas called domains of content. These areas are: information about drugs, resistance skills training, decisionmaking skills, pledges or personal commitments not to use drugs, values clarification, norm setting, creating alternatives, stress management, self-esteem building, life-skills training, goal setting, and peer problemsolving skills. Each of those 12 domains of content can be operationalized in many different ways. Such differences may be clinically or theoretically important.

In addition to coding the actual content of the intervention, it is often useful to code information about the manner of experimental treatment delivery and the context in which the intervention is tested. This includes information such as:

- Who delivered the content (e.g., a teacher, a counselor, a peer);

- What was the format of the program (e.g., lecture, discussion, role play);

- How long was the program;

- What substances were the focus of the program;

- What was the goal of the program (e.g., delay onset of drug use, abstinence, decreased high-risk use); and

- Who is the audience (all regular students, volunteers for a drug education program, residents in a juvenile detention center).

Depending on the purpose of the review, it may be desirable to code treatments so that various operationalizations of each domain of content can be examined for differences in treatment effectiveness. However, depending on the sample of studies included in the meta-analysis, analyzing individual treatment components may or may not be possible. Hansen (1992) noted two problems that may limit the ability of meta-analysts to determine the effects of some specific components of drug abuse prevention interventions. First, most of the drug abuse prevention interventions contained elements of content from multiple domains of content. This may be advantageous from a clinical perspective when the goal is to determine if certain programs are effective in field settings. However, from theory or policy perspectives it is often desirable to identify the maximally efficacious components of a prevention program to understand causal mechanisms or to refine and streamline an intervention. This will require the testing of specific components of the intervention, often using factorial-type designs or including within-study comparisons of alternate treatments. There must be multiple tests of individual treatment components in the extant research before meta-analysis will be useful in summarizing the effects of those treatment components. Nonetheless, if only multidimensional intervention programs have been tested, it is important that their effects be summarized as well.

The second problem is related to reporting weaknesses that are often found in primary research (Orwin and Cordray 1985). Hansen's (1992) analyses were limited by serious deficiencies in the documentation of intervention characteristics. For example, time on task within a multidimensional treatment was not reported consistently. Tobler (this volume) noted that the descriptions of interventions were variable, with the content of drug abuse prevention programs being better reported than the manner of treatment delivery. Sometimes it is possible to obtain missing information from primary researchers. Tobler (1994) reported contacting primary authors when information about the intervention program was missing or ambiguous. Although contacting primary researchers is time consuming and not frequently done, it should be considered if vital information is missing from the research report.

In addition to coding the characteristics of the experimental intervention that are absolutely essential for the planned analyses, it can also be helpful to abstract or code detailed treatment characteristics. Having this data readily available allows the reviewer to provide a thick description of the interventions included in the review. It also enables the reviewer to determine whether interventions grouped together are, in fact, quite variable in substantial ways. Such differences could be used to explain results if outcomes are not homogeneous.

There is a downside, of course. Detailed coding of the experimental intervention requires additional time and resources that may be hard to justify if the resulting data play only a minor role in the review. It also may turn out to be an inefficient use of resources if weaknesses are so prevalent that specific characteristics (e.g., time on task within a multidimensional treatment) are reported in too few studies to allow them to be used as meaningful descriptors or potential moderator variables.

How To Categorize and Aggregate Interventions. In addition to coding specific characteristics of the experimental intervention, it is usually necessary to group experimental interventions into meaningful categories to enhance interpretability and facilitate data analysis. A review of several meta-analyses on drug abuse prevention programs reveals different approaches to establishing categories, different reporting about the processes involved and the information used to develop intervention categories, and rather different numbers of categories of interventions developed.

Some approaches have been theory driven, like the four-factor classification of prevention program orientation used by Rundall and Bruvold (1988; Bruvold 1993). Various empirical approaches to developing categories of interventions have been used. Tobler (1986) reported analyzing major themes reported by researchers and proposed five functional-content categories. In later work, Tobler (this volume) grouped programs empirically based on both content matter and the manner of program delivery. Bangert-Drowns (1988) also categorized programs functionally. However, that categorization scheme included only three types of interventions: knowledge only, affective education only, and mixed. Hansen (1992), on the other hand, identified and coded each program according to "building block theoretical concepts." The pattern of occurrence of these concepts within programs was used to develop a provisional conceptual framework containing six distinct types of program content. In other substantive areas, such as psychology, education, and health care, theoretical (Shadish 1992), empirical (Devine and Cook 1983, 1986; Fernandez and Turk 1989), and multidimensional scaling (Smith and Glass 1977) approaches have been used to classify interventions. All of these approaches can be useful for addressing certain research questions. Just as one study does not answer all research questions on a topic, neither can a single meta-analysis. It should, however, be up to the experts in drug abuse prevention to judge the usefulness and relevance of the categories developed by the various meta-analysts of the drug abuse prevention research.

There are many sources of relevant information for the meta-analyst to use while developing categories of interventions. These include the research base itself; relevant theoretical, descriptive, and social policy literatures; and practice-derived knowledge. Among these sources, the studies under review provide the major limiting factor for the meta-analyst. The meta-analyst is working with an existing data set (the studies on the topic that have been competed and can be retrieved). If there is not a sufficient number of studies that included a specific version of the experimental intervention, then the effect of that specific type of treatment can not be examined. However, other approaches are possible; even if there is not a sufficient number of studies in each of several types of norm-setting interventions, it may be appropriate and useful to group together somewhat different versions of the intervention into a more general category (e.g., of norm-setting interventions).

Two aspects of coding interventions are essential for the meta-analyst to keep in mind. First, whatever the process and the source of data used to develop categories of interventions, readers of the review should be well informed about how the categorization scheme was developed and the characteristics of intervention programs that were aggregated into specific categories. Without this type of information, the consumers cannot judge the validity of the conclusions drawn about the interventions. Second, it is essential to remember that between-study contrasts of the relative effectiveness of different types of experimental interventions (also called review-generated evidence by Cooper (1989)) have some inherent weaknesses. For example, caution must be used when one interprets differences in average effect size values from different subsets of studies. The subsets of studies may differ in many ways other than the fact that intervention "x" was the experimental treatment in one group of studies and intervention "y" was the experimental treatment in another group of studies. Differences in historic period, sample, setting, and/or operationalization of the outcome measure could be the reason for any observed differences in treatment effectiveness; therefore, between-study contrasts should be viewed as descriptive, and not as a basis for causal inference. However, between-study contrasts can provide a good basis for designing future research. Within-study contrasts of intervention "x" and intervention "y" from well-designed and executed studies can be used to examine causal relationships about the relative effectiveness of different types of experimental interventions. Within-study contrasts involve two or more experimental interventions being compared directly in the same study, or multiple experimental interventions being compared with a control group from the same study. By their very nature, within-study contrasts hold constant most, if not all, of the source of differences outlined above that plague between-study contrasts.

Timing and Personnel Involved in Coding Interventions

In addition to deciding what to code about interventions and how to group interventions for analysis, the meta-analyst must decide who will code this data and when it will be coded. It is fairly typical for meta-analyses to involve the use of an established coding form with specific directions. Nonetheless, the need for knowledgeable, well-trained coders with adequate background in the substantive area is essential (Wachter 1988). In developing a coding form for studies, it is often possible to adapt certain sections of the coding form from other meta-analyses. For example, from a practical point of view, there are a fairly limited number of ways to record such things as publication source, publication date, or the number of subjects in the treatment group. In spite of the ambiguities often present in research reports, it is fairly easy to train individuals who have graduate research experience in the content area under review to code many characteristics of the study, the subjects, and the outcomes, with good interrater reliability.

Developing coding categories for interventions, on the other hand, requires much more specific knowledge about the substantive area under review. If coding categories for interventions are to be useful, they must be fine tuned to the purpose of the meta-analysis, the content area under review, and the typical reporting practices in the studies being reviewed.

Coding interventions requires extensive substantive knowledge about the phenomenon of interest. Given the typical length of most published research reports, it is probably not surprising that the details provided about interventions are often sketchy. When the description of the experimental intervention is very brief or vaguely worded, the coders are required to make many judgments about the appropriate categorization of various treatment characteristics. Coders who are well grounded in the substantive area will be better able to identify when coding categories need to be modified to capture the essence of the intervention, or to determine when reporting weaknesses or the use of outdated terms or operationalization of treatments is a factor in differences between interventions. Examples of the foregoing are provided by Hansen (1992) in a discussion of the substantial changes in how the constructs norm setting and alternatives training have been used in the drug abuse prevention intervention literature over the last two decades.

Minimizing Bias

Given the many judgments coders must make to categorize interventions, minimizing coder bias is a major concern (Cordray 1990*a*, 1990*b*). Experimenter expectancies (Rosenthal 1966), or in this case coder expectancies, are the main threat to the accuracy of coding interventions. In this context, coder expectancies refer to knowledge or beliefs on the part of the coder that adversely affect the integrity of coding decisions. The issue here is not fraud or malevolent misrepresentation of treatment characteristics. Those can be problems, of course, but they require different remedies that are not addressed in this chapter. Inaccuracies arising from coder expectancies are much more subtle. If one wants to minimize this bias, its sources need to be recognized and steps taken to reduce its effect.

There are two main sources of experimenter expectancies. The first arises from close affiliation, on the part of the principal investigator or the coders, with certain studies or types of treatments under review. Although the author has stressed how important it is for the meta-analyst and the coders to be very familiar with relevant literatures, theories, and practice, such knowledge can be a source of bias. If the individuals making coding decisions are too closely affiliated with (and biased toward) particular studies or types of treatments, or if the data coders presume the principal investigator wants a certain outcome no matter what, their ability to code studies accurately may be restricted. Studies coinciding with a reviewer's beliefs may be evaluated much more favorably than those that do not, an effect that Mahoney (1977) called the "confirmatory bias."

The second potential source of bias comes from the research reports themselves (Cooper 1989). Among almost any large group of studies there will be variability in the prestige of the author(s), in the source of funding for the study, and in the prestige of source of the research report (e.g., a major journal in a field or an unpublished thesis). There will also be variability in the writing ability of the author(s), the direction and magnitude of treatment effects, and whether statistically significant results were obtained. Differences like these can create halo or shadow effects that may influence the decisions coders make about the experimental intervention (Peters and Ceci 1982).

There is no perfect solution to this problem. However, at a minimum the following checks and balances are recommended to improve the accuracy of coding studies.

- Prior to coding studies the research team should thoroughly review the coding categories and the coding directions for clarity, relevance, and comprehensiveness;

- Intercoder agreement should be examined, and the training of coders and refining of directions should continue until established;

- Coders should take frequent breaks to minimize errors arising from fatigue and boredom;

- Outside readers (or research team members), ideally with different theoretical or professional perspectives, can serve as supplemental coders to consider particularly ambiguous research reports or to help resolve conflicts when intercoder agreement is not achieved; and

- Coders should be made aware of the potential sources of bias so that they can be critical of their own decisions.

Another approach to minimizing coder bias that is gaining in popularity is to blind the coders to certain nonpertinent sections of the research report while coding other sections of the study (Chalmers et al. 1988; Devine 1992; Devine and Cook 1983; Sacks et al. 1987). For example, to minimize the effect of information such as the direction or magnitude of treatment effect or the title of the journal when coding interventions, steps could be taken to black out or cut out all irrelevant and potentially biasing information. Ideally, one would want the coders to have available only the sections of the research report related to the type of information being coded. This information can be photocopied, with irrelevant information blacked out, and then labeled with only an identification number. Depending on purpose of the review and the variables being examined, when coders are determining the content and nature of the experimental and control interventions, it might be helpful for the them to have a photocopy of the introduction to the study, the review of literature, relevant parts of the methods sections, and any related papers by the same author for reference. These various sources of information might be critical for coding the experimental and control interventions. For example, coders may be better able to determine the implications of certain program descriptors if they are aware of the theoretical underpinnings upon which the study is based. It is also important for coders to review related papers by the same author, if they include greater detail about the intervention than the primary published report of the study.

There is another advantage in abstracting detailed information about the experimental intervention. One must be concerned with coder expectancies at the time of initial coding of studies and also during data analysis. If the effect size values obtained from a group of studies are heterogeneous (i.e., statistical testing suggests there is an interaction between type of intervention and magnitude of treatment effect), then often studies are partitioned to test for homogeneity within more refined subsets of interventions (Hedges and Becker 1986). It is desirable to avoid knowledge about the direction and magnitude of effect size values when making decisions about how to partition studies. If detailed descriptions of the interventions are available in a form clearly separated from effect size information, then it is easier for the meta-analyst to make less biased decisions about how to reaggregate studies into subgroups for analysis.

While blinding coders to certain information can reduce bias, one must make sure that other problems are not created by artificially dividing studies into so many pieces for coding. Not all apparently credible studies provide credible results. For example, upon close inspection the reviewer may find that there was a poor fit between the program goals and the outcomes measured; a program could have been well designed but not faithfully implemented, or there could have been a fatal flaw arising from failed random assignment or considerable treatment diffusion across research design levels. Given the variability encountered in research reports, such information might appear in the methods section, in a footnote, or at the very end of the discussion section. Thus it is important for the research report to be read in its entirety by a research team member. This way, important information is less likely to be missed.

Coding Control Conditions

In order to fully interpret the content, manner, and context of the experimental intervention, it is essential to know something about the control condition. Control group data can be derived from many sources. Meta-analysis selection criteria usually specify that only studies with certain types of control conditions will be included in the review. In addition to the variability of control groups that arises from the manner of assignment to treatment condition, there is variability in control groups arising from the extent to which there is overlap between the experimental and control interventions. Control group interventions can include "no treatment," the "usual" treatment for someone in their situation, a placebo treatment, or an alternate treatment. There also can be noteworthy variation within each of these categories of control treatment. For example, among no-treatment control groups there can be varying degrees of treatment delivered by others in the environment.

In the drug abuse prevention area, while many forces in an adolescent's life foster drug abuse, many other forces are at work to prevent drug abuse. In addition to the efforts of parents and teachers, there may be relevant programming in the mass media (including video arcade games) and community-based or church-related efforts. If such efforts decrease drug use in the target population, they decrease the base rate of drug use and make it much more difficult for the effectiveness of the intervention to be demonstrated. Because these less formal interventions are rarely documented and vary within communities over time, it is difficult for the reviewer to account for their influence.

Placebo and usual care treatments also can vary in the degree of overlap with experimental interventions. In the health care literature and probably in other literatures as well, the actual content of usual care is rarely documented and so the degree of overlap is difficult to assess. The actual content of placebo interventions is usually better documented. Variability among placebo interventions in the degree of overlap with the experimental intervention has been shown to be related to the magnitude of treatment effect (Devine and Cook 1983). In social situations it is difficult to create placebo treatments that are both credible and as inert as the prototypical sugar pill. With very brief interventions, it often is possible to create an attention-type control treatment that provides equivalent time with an interested researcher or professional but contains irrelevant content. Creating credible placebo treatments for longer experimental interventions is much more difficult. Placebo treatments containing content (e.g., conflict resolution skills training) that is likely to affect an outcome of interest (e.g., drug use) are closer to alternate treatments, and it is better to treat them as such. Alternate interventions as control treatments provide special challenges as well as advantages to the meta-analyst. While they may provide excellent theory-relevant tests of causal mechanisms or the relative effectiveness of different treatments, they address very different hypotheses than contrasts with no-treatment control groups. Although it is problematic to combine tests of different hypotheses into a single analysis, contrasts between alternate treatments should not be disregarded. They are particularly valuable as a source of within-study contrasts from which one can get appropriate data for testing the relative effectiveness of different types of treatments.

Reporting weaknesses about control conditions is a major problem. While primary researchers usually report factors that affect the equivalence between experimental and control groups (e.g., the manner of assignment to treatment condition), the actual experiences of the control group are often not as well reported. This makes it particularly challenging to code the content of control treatments and to conduct fine-grained analyses that account for differences in type of control group.

CONCLUSIONS

Coding experimental and control interventions is a critical step in the meta-analysis of intervention research. This information helps to define the specific constructs involved in the independent variable of the hypothesis tested. Special challenges exist when the effects of multidimensional treatments are the focus of interest or when important information is not detailed in the written report. There are no simple solutions. However, two guiding principles can help the meta-analyst and the consumers of a meta-analysis. First, as with all forms of research, the procedures and protocols guiding decisions should be explicit enough to allow critique and replication. And second, treatments that are aggregated should be similar enough to make combining their outcomes meaningful to likely consumers of the review.

Special actions may be needed to help the meta-analyst overcome the problems associated with working with an existing and limited set of data. In order to protect the integrity of coding, steps should be taken to reduce the likelihood of experimenter expectancies. It may be necessary for the meta-analyst to contact the primary researchers to obtain needed information. Special caution is always needed to appropriately interpret between-study contrasts.

Prospective meta-analysts and consumers of meta-analysis are cautioned to be modest in their expectations. Meta-analysis is limited by the extent, quality, reporting detail, and specific operationalizations tested in the existing research on the topic of interest. All forms of research review have these same limitations. The two major functions of most research reviews are to summarize what is known and to foster the further development of knowledge in an area through recommendations about future research topics and practices. To the extent that meta-analyses are more explicit, comprehensive, and critical than other forms of research reviews, they make unique contributions to the building of knowledge.

REFERENCES

Bangert-Drowns, R.L. The effects of school-based substance abuse education—a meta-analysis. *J Drug Educ* 18(3):243-264, 1988.

Bruvold, W.H. A meta-analysis of adolescent smoking prevention programs. *Am J Public Health* 83(6):872-880, 1993.

Chalmers, T.C. Problems induced by meta-analyses. *Stat Med* 10:971-980, 1991.

Chalmers, T.C.; Berrier, J.; Sacks, H.S.; Levin, H.; Reitman, D.; Nagalingam, R.; and Sacks, H.S. Meta-analysis of randomized control trials as a method of estimating rare complications of non-steroidal anti-inflammatory drug therapy. *Aliment Pharmacol Ther* 2(1):9-26, 1988.

Cooper, H.M. *A Guide for Literature Reviews*. 2nd ed. Newbury Park, CA: Sage Publications, 1989.

Cordray, D.S. Meta-analysis: An assessment from the policy perspective. In: Wachter, K., and Straf, M., eds. *The Future of Meta-Analysis*. New York: Russell Sage Foundation, 1990*a*. pp. 99-119.

Cordray, D.S. Strengthening causal interpretations of nonexperimental data: The role of meta-analysis. In: Sechrest, L.; Perrin, E.; and Bunder, J., eds. *Research Methodology: Strengthening Causal Interpretations of Nonexperimental Data*. Washington, DC: U.S. Department of Health and Human Services Public Health Service, Agency for Health Care Policy and Research, 1990*b*. pp. 151-172.

Devine, E.C. Effects of psychoeducational interventions: A meta-analytic analysis of studies with surgical patients. *Diss Abstr Int* 44:3356B, 1984.

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.; 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.

Devine, E.C., and Cook, T.D. A meta-analytic analysis of effects of psychoeducational interventions on length of post-surgical hospital stay. *Nurs Res* 32:267-274, 1983.

Devine, E.C., and Cook, T.D. Clinical and cost-saving effects of psychoeducational interventions with surgical patients: A meta-analysis. *Res Nurs Health* 9:89-105, 1986.

Fernandez, E., and Turk, D.C. The utility of cognitive coping strategies for altering pain perception: A meta-analysis. *Pain* 38:123-135, 1989.

Glass, G.V. Primary, secondary, and meta-analysis of research. *Educ Res* 5:3-8, 1976.

Glass, G.V.; McGaw, B.; and Smith, M.L. *Meta-analysis in Social Research*. Beverly Hills, CA: Sage Publications, 1981.

Hansen, W.B. School-based substance abuse prevention: A review of the state of the art in curriculum, 1980-1990. *Health Educ Res* 7(3):403-430, 1992.

Hedges, L.V., and Becker, B.J. Statistical methods in the meta-analysis of research on gender differences. In: Hyde, J.S., and Linn, M.C., eds. *The Psychology of Gender*. Baltimore: Johns Hopkins University Press, 1986. pp. 14-50.

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

Lykken, D.T. Statistical significance in psychological research. *Psychol Bull* 70:151-159, 1968.

Mahoney, M. Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cogn Ther Res* 1:161-175, 1977.

Myers, D.G. Union is strength: A consumer's view of meta-analysis. *Pers Soc Psychol Bull* 17(3):265-266, 1991.

Nurius, P.S., and Yeaton, W.H. Research synthesis reviews: An illustrated critique of "hidden" judgments, choices, and compromises. *Clin Psychol Rev* 7(6):695-714, 1987.

Orwin, R.G., and Cordray, D.S. Effect of deficient reporting on meta-analysis: A conceptual framework and reanalysis. *Psychol Bull* 97:134-147, 1985.

Peters, D., and Ceci, S. Peer-review practices of psychological journals: The fate of published articles, submitted again. *Behav Brain Sci* 5:187-255, 1982.

Presby, S. Overly broad categories obscure important differences between therapies. *Am Psychol* 33(5):514-515, 1978.

Rosenthal, R. *Experimenter Effects in Behavioral Research*. New York: Appleton-Century-Crofts, 1966.

Rundall, T.G., and Bruvold, W.H. A meta-analysis of school-based smoking and alcohol use prevention programs. *Health Educ Q* 15(3):317-334, 1988.

Sacks, H.S.; Berrier, J.; Reitman, D.; Ancona-Berk, V.A.; and Chalmers, T.C. Meta-analysis of randomized control trials. *N Engl J Med* 316:450-455, 1987.

Shadish, W.R., Jr. Do family and marital psychotherapies change what people do?: A meta-analysis of behavioral outcomes. In:

Cook, T.D.; Cooper, H.; 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.

Smith, M.L., and Glass G.V Meta-analysis of psychotherapy outcome studies. *Am Psychol* 32: 752-760, 1977.

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 Issues* 16(4):537-567, 1986.

Tobler, N.S. Drug prevention programs can work: Research findings. *J Addict Dis* 11(3):1-27, 1992.

Tobler, N.S. Meta-analytic issues for prevention intervention research. In: Collins, L.M., and Seitz, L.A., eds. *Advances in Data Analysis for Prevention Intervention Research*. National Institute on Drug Abuse Research Monograph 142. NIH Pub. No. 94-3639. Washington, DC: Supt. of Docs., U.S. Govt. Print. Off., 1994.

Wachter, K.W. Disturbed by meta-analysis. *Science* 1407-1408, 1988.

Wanous, J.P.; Sullivan, S.E.; and Malinak, J. The role of judgment calls in meta-analysis. *J Appl Psychol* 74(2):259-264, 1989.

ACKNOWLEDGMENTS

The preparation of this manuscript was supported in part by a grant from the National Institutes of Health, National Institute of Nursing Research R01 NR1539-05.

AUTHOR

Elizabeth C. Devine, Ph.D., FAAN
Associate Dean for Research
Associate Professor
School of Nursing
University of Wisconsin-Milwaukee
P.O. Box 414
Milwaukee, WI 53217

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}(   ) = 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 (Qr = 657.43, df = 17, p < .001; Qe = 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.

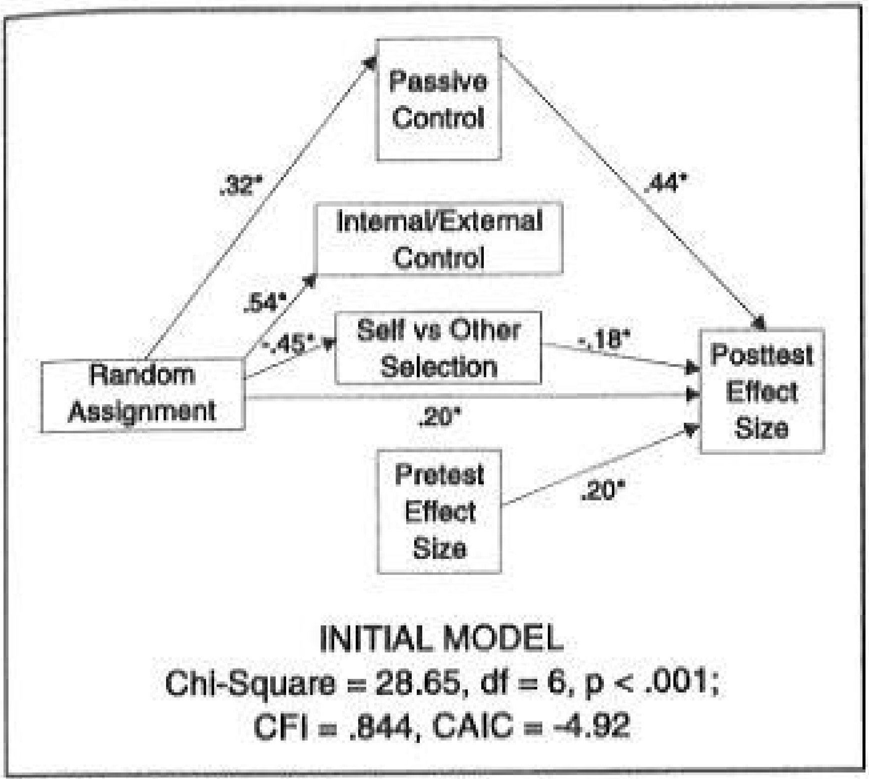


FIGURE 1. *Initial mediational model of assignment-outcome relationship.*

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.

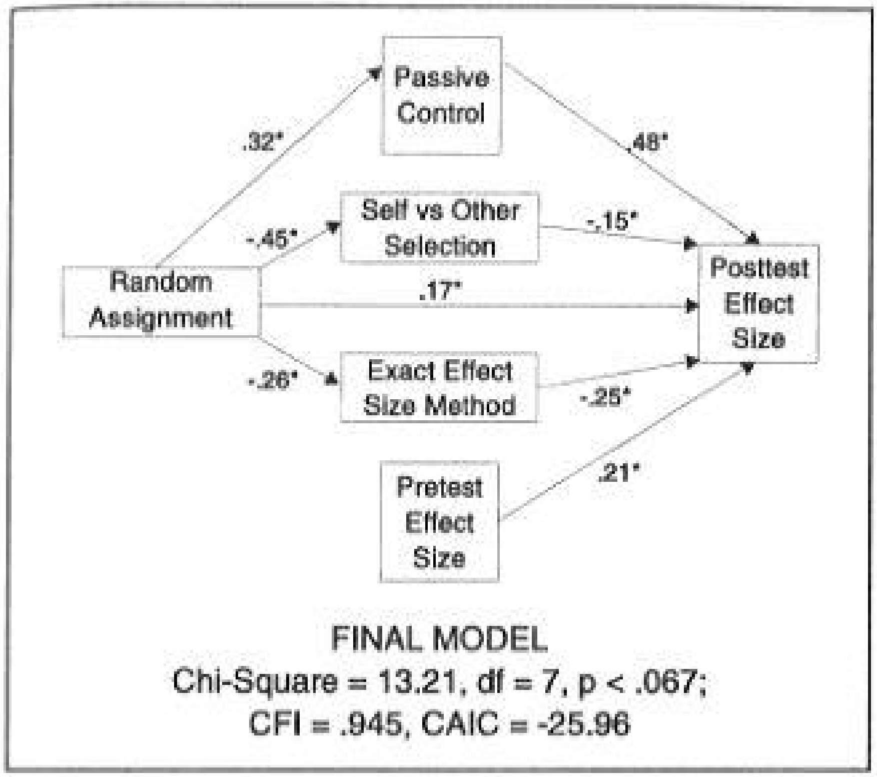


FIGURE 2. *Final mediational model of assignment-outcome relationship.*

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

Drawing Generalized Causal Inferences Based on Meta-Analysis

Georg E. Matt

INTRODUCTION

Research syntheses are more and more used to inform decisionmakers about the effects of a particular policy or of different policy options. For instance, do substance abuse prevention programs in junior high schools reduce drug use in high school? Are random drug tests more effective than drug education programs in reducing drug use? Are social influence prevention programs more effective with boys than with girls?

In the language of Cook and Campbell (1979) and others, these questions involve causal relationships of two kinds: bivariate causal relationships and causal moderator relationships. In bivariate causal relationships, one is examining whether deliberately manipulating one entity (e.g., introducing a prevention program) will lead to variability in another entity (e.g., onset of drug use). In causal moderator relationships, one is interested in identifying variables that modify the magnitude or sign of a causal relationship (e.g., in the presence of peer counselors, prevention programs are more effective than in their absence).

Meta-analyses seek to draw conclusions about populations, classes, or universes of variables. This is different from primary studies in which, for instance, researchers examine the causal effects of a particular drug education curriculum in a particular school with students in a particular grade. Instead, meta-analyses seek to draw conclusions regarding a universe of persons (e.g., students in grades 4 to 12), a universe of interventions (e.g., substance abuse prevention programs), a universe of outcomes (e.g., drug use), a universe of settings (e.g., schools), and a universe of times (e.g., 1980's). Thus, meta-analyses are concerned with generalized causal relationships. This chapter deals with specific threats to the validity of meta-analyses, examining generalized bivariate causal and causal moderator relationships.

As Campbell originally coined the term, "validity threats" refer to situations and issues in research practice that may lead to erroneous conclusions about a causal relationship. However, unlike the validity threats identified by Campbell and Stanley (1963) and Cook and Campbell (1979), this chapter is not concerned with validity threats in primary studies. Because research synthesis relys on the evidence generated from many different studies, the issue is the total bias across studies rather than bias in a single primary study. Thus, the validity threats discussed in this chapter refer to issues in conducting a research synthesis that may lead to erroneous conclusions about a generalized causal relationship.

Drawing generalized causal inferences in meta-analysis involves three major steps. First, research synthesists need to establish that there is an association between the class of interventions and the class of outcomes. In other words, there has to be evidence that the intervention effect across studies is reliably different from zero. Second, research synthesists have to defend the argument that the relationship examined across studies is causal. Phrased differently, they have to rule out that factors other than the treatments as implemented were responsible for the observed change in the outcomes. Third, given the specific instances of interventions, outcomes, persons, settings, and times included in a review, research synthesists have to clarify the universes of interventions, outcomes, populations, settings, and times about which one can draw inferences. The following paragraphs discuss validity threats that research synthesists may encounter at each of these three steps of generalized causal inference. The research reviews by Bangert-Drowns (1988), Hansen (1992), and Tobler (1986, 1992) are used to provide examples of validity threats and to indicate ways for coping with them.

THREATS TO INFERENCES ABOUT THE EXISTENCE OF A RELATIONSHIP: IS THERE AN ASSOCIATION BETWEEN TREATMENT AND OUTCOME CLASSES?

The first group of validity threats deals with issues that may lead a research synthesist to draw erroneous conclusions about the existence of a relationship between a class of independent variables (i.e., interventions) and a class of dependent variables (i.e., outcomes). In the language of statistical hypothesis testing, these threats may lead to type 1 or type 2 errors because of deficiencies in either the primary studies or the meta-analytic review process. Because research syntheses are concerned with generalized relationships, a single threat in a single study is not likely to jeopardize meta-analytic conclusions in any meaningful way. More critical is whether the same source of bias operates across all or most of the studies being reviewed and whether different sources of bias fail to cancel each other out across studies. This may then lead to a predominant direction of bias, inflating or deflating estimates of a relationship. See table 1 for a list of threats to valid inferences about the existence of a relationship in a meta-analysis.

Unreliability in Primary Studies

Unreliability in implementing or measuring variables contributes random error to the within-group variability of a primary study, thereby attenuating effect size estimates not only within such a study but also when studies are aggregated meta-analytically.

In the context of drug prevention programs, reliability issues include the measurement of outcome variables such as drug knowledge, attitudes toward drugs, actual drug use, and the fidelity with which prevention programs were implemented. To deal with this issue, correction formulas have been suggested to adjust effect estimates and their standard errors

TABLE 1. *Threats to inferences about the existence of a relationship between treatment and outcome classes.*

(1) Unreliability in primary studies
(2) Restriction of range in primary studies
(3) Missing effect sizes in primary studies
(4) Unreliability of codings in meta-analyses
(5) Capitalizing on chance in meta-analyses
(6) Bias in transforming effect sizes
(7) Lack of statistical independence among effect sizes
(8) Failure to weight study level effect sizes proportional to their precision
(9) Underjustified use of fixed- or random-effects models
(10) Lack of statistical power

(Hunter and Schmidt 1990; Rosenthal 1984). However, Tobler (1986) found that program implementation and the reliability of outcome measures are often poorly documented in primary studies, making comprehensive attempts to correct for attenuation unfeasible. Nevertheless, attenuation corrections are sometimes useful to make the degree of attenuation constant across studies and to better understand the magnitude of effects if interventions were consistently implemented and outcomes measured without error.

Restriction of Range in Primary Studies

When the range of an outcome measure is restricted in a primary study, all correlation coefficients involving this measure are attenuated. Range restrictions may influence other effect size measures differently. For instance, the selection of homogeneous subgroups, blocking, and matching reduce both within-group variability and range. Everything else being equal, this decreases the denominator of the effect size estimated, thereby increasing the magnitude of effect sizes. When such design characteristics operate, Kulik and Kulik (1986) refer to the resulting effect sizes as operative rather than interpretable. Aggregating such operative effect sizes may yield a predominant bias across studies.

In research syntheses of prevention programs, restricted ranges can occur if primary studies involve extreme groups or homogeneous subgroups from a larger population. Effect estimates based on these studies may overestimate program effects in populations with larger variances. Correction formulas can be applied to adjust effect size estimates (Hunter and Schmidt 1990) if valid estimates of population variances are available.

Missing Effect Sizes in Primary Studies

Researchers sometimes provide an incomplete report of findings because of page limitations in journals, the particular emphasis of a research paper, unexpected results, or poor measurement. This reporting practice may bias effect estimates in meta-analyses if researchers in primary studies fail to report, for instance, statistically nonsignificant findings or statistically significant findings in an unexpected direction.

Selective reporting in primary studies is a pervasive issue in many meta-analyses. To prevent possible biases, it is always desirable to code the most complete documents and to contact study authors to obtain information not available in research reports (Premack and Hunter 1988; Shadish 1992). If this strategy is not feasible, there is a need to consider imputation strategies (Little and Rubin 1987; Rubin 1987) and to explore how missing effect sizes may have influenced effect estimates in a meta-analysis.

Unreliability of Codings in Meta-Analyses

All the data synthesized in a meta-analysis are collected through a coding process susceptible to human error. Thus, meta-analyses contribute sources of unreliability in addition to those in primary studies. Unreali-ability in the coding process adds error variation to the observations, increasing estimates of standard error and attenuating correlations among effect size estimates and study characteristics. Strategies for controlling and reducing error in codings include comprehensive coder training, pilot testing, and reliability assessments (Cooper 1989).

Capitalizing on Chance in Meta-Analyses

There are three major ways in which meta-analyses may capitalize on chance. First, a publication bias may exist such that studies with statistically significant findings in support of a study's hypotheses are more likely to be submitted for publication. If this is the case, the studies published in the behavioral and social sciences are likely to be a biased sample of all the studies actually carried out (Greenwald 1975; Rosenthal 1979). A second way meta-analysts may capitalize on chance is in extracting effect sizes within studies. Research reports frequently present more than one estimate, especially when there are multiple outcome measures, multiple treatment and control groups, and multiple delayed time points for assessment. Not all of these effect estimates may be relevant for a particular topic, and some relevant estimates may be more important than others. Meta-analysts must then decide which effect estimates should be included in the meta-analysis. Bias may occur when selected effect estimates are just as substantively relevant as those not selected, but differ in average effect size (Matt 1989). A third way that meta-analysts may capitalize on chance is by conducting a large number of statistical tests without adequately controlling for type 1 error.

Bias in Transforming Effect Sizes

Meta-analyses require that findings from primary studies be transformed into a common metric such as a correlation coefficient, a standardized mean difference, or standard normal deviate. Because studies differ in the type of quantitative information they provide about intervention effects, transformation rules were developed to derive common effect size estimates from many different metrics. Bias results if some types of transformation lead to systematically different estimates of average effect size or standard error when compared to others. For instance, this is likely to be the case when primary studies fail to report exact probability levels and truncated levels (e.g., $p < 0.05$) have to be used to estimate an effect size.

Lack of Statistical Independence Among Effect Sizes

Hedges (1990) states that there are at least four reasons why effect size estimates entering into a meta-analysis may lack statistical independence: (a) Different effect size estimates may be calculated on the same respondents using different measures; (b) effect sizes may be calculated by comparing different interventions to a single control group, or different control groups to a single intervention group; (c) different samples may be used in the same study to calculate an effect estimate for each sample; and (d) a series of studies may be conducted by the same research team, resulting in nonindependent results. A predominant bias may occur if stochastic dependencies among effect sizes influence average effect estimates and their precision (Hedges and Olkin 1985).

The simplest approaches for dealing with dependencies involve analyzing only one of the possible correlated effects or an average effect for each study. However, these approaches fail to take into account information concerning the differences between nonindependent effect sizes, and multivariate analyses or hierarchical linear models may be called for (Bryk and Raudenbush 1992; Raudenbush et al. 1988; Rosenthal and Rubin 1986).

Failure To Weight Study Level Effect Sizes Proportional to Their Precision

Even if one obtains unbiased effect estimates within a study, simply averaging them may yield biased average effect estimates and incorrect sampling errors if the effect sizes from different studies vary in precision (i.e., have different standard errors) (Shadish 1992). Similarly, t tests, analyses of variance (ANOVAs), and regression analyses may provide incorrect results unless weighted estimation procedures are used (e.g., weighted least squares).

Underjustified Use of Fixed- or Random-Effects Models

For the statistical analysis of effect sizes, Hedges and Olkin (1985) distinguish between postulating a model with fixed or random effects. In its simplest form, the fixed-effects model assumes that all studies (e.g., social influence programs) have a common but unknown effect size and that estimates of this population value differ only as a result of sampling variability. In the fixed-effect model, analysts are interested in estimating the unknown population effect size and its standard error. In the random-effects model, each treatment is assumed to have its own unique underlying effect and to be sampled from a universe of related but distinct treatments. Under the random-effects model, the effects of a sample of treatments are best represented as a distribution of true effects rather than as a point estimate.

There is no simple indicator for which model is correct. However, two factors should be considered in the decision whether to assume a fixed- or a random-effects model. The first concerns assumptions about the processes generating an effect. For instance, in the context of drug prevention programs, are all the prevention programs labeled "social influence" identical and are they standardized and administered consistently in all studies? Are the processes by which social influence programs affect drug use the same across all studies? If the answer to these questions is "no" or "probably no," a random-effects model is indicated. The second factor to consider is the heterogeneity of the observed effect sizes. A homogeneity test can be conducted to determine whether the observed variance exceeds what is expected based on sampling error alone. If the homogeneity hypothesis is rejected, the analyst may want to consider the possibility of a random-effects model. Alternatively, if one has reason to insist on a fixed-effects model, the search would begin for the variables responsible for the increased variability.

Lack of Statistical Power

When compared to statistical analyses in primary studies, statistical power will typically be much higher in meta-analyses, particularly when meta-analysts are only interested in estimating the average effect of a broad class of interventions. However, as the meta-analyses on drug prevention programs show (Bangert-Drowns 1988; Tobler 1986, 1992), research synthesists are frequently interested in examining effect sizes for subclasses of treatments and outcomes, different types of settings, and different subpopulations. These subanalyses often rely on a much smaller number of studies than the overall analyses and result in a large number of statistical tests. The meta-analyst then has to decide which tradeoff to make between type 1 and type 2 error, or, in other words, between the number of statistical tests and the statistical power of these tests.

THREATS TO INFERENCES ABOUT CAUSATION: ARE THERE ANY NONCAUSAL REASONS FOR THE ASSOCIATION?

Whenever a reliable association between independent and dependent variables is presumed to be causal, some additional threats need to be considered. Note again that inferences about the possible causal nature of a treatment-outcome relationship are not necessarily jeopardized by deficiencies in primary studies. A plausible threat arises only if the deficiencies within each study combine across studies to create a predominant direction of bias. In the following, two aspects are considered: bivariate causal relationship and causal moderator relationship. Table 2 gives a brief summary of the threats. See Matt and Cook (1993) for a discussion of threats to causal mediating relationships.

TABLE 2. *Threats to inferences about causation.*

(1) Failure to assign at random
(2) Deficiencies in the implementation of treatment contrasts
(3) Confounding levels of the moderator with substantively irrelevant study characteristics

Failure To Assign at Random

If experimental units (e.g., students, classrooms, schools) are not assigned to treatment conditions at random, a variety of third-variable explanations can jeopardize causal inference in primary studies. The failure to assign at random jeopardizes meta-analytic conclusions if it results in a predominant bias across primary studies.

For research studies of school-based substance abuse prevention programs, Hansen (1992) argues that selection biases are potential threats in quasi-experimental designs comparing groups that inherently differ in expected drug use. In some studies, higher levels of initial risk for substance abuse may be a precondition for entry into a prevention program. Moreover, Hansen's (1992) research suggests that selection biases may be more likely in some program groups (e.g., alternatives) than in others (affective education). However, despite the potential for selection biases, Tobler's meta-analysis (1986) found little evidence for a predominant bias when comparing randomized trials and quasi-experimental studies.

Deficiencies in the Implementation of Treatment Contrasts

Outside of controlled laboratories, random assignment is often difficult to implement; and even if successfully implemented, it does not ensure that comparability between groups is maintained beyond the initial assignment. Even the most carefully designed randomized experiments and quasi-experiments are not immune to implementation problems such as differential attrition and diffusion of treatments. If the reviewed studies share deficiencies of implementation, a predominant bias may result when studies are combined. However, in trying to examine the implementation of prevention programs more closely, Tobler (1986) found that primary reports often failed to report relevant information.

Hansen (1992) points out another type of implementation issue: studies of school-based prevention programs often involve small numbers of experimental units (i.e., schools), thus jeopardizing the equivalence of control and treatment groups even if experimental units are randomly assigned. While this may threaten the internal validity of a primary study, one would not expect that such nonequivalence necessarily yields a predominant bias when studies are combined in a meta-analysis.

Confounding Levels of a Moderator Variable With Substantively Irrelevant Study Characteristics

Moderator variables condition causal relationships by specifying how an outcome is related to different variants of an intervention, to different classes of outcomes, and to different types of settings and populations. All moderator variables imply a statistical interaction and identify those factors that lead to differently sized cause-effect relationships. Moderators can change the magnitude or the sign of a causal effect, as when Tobler (1986) concluded that peer programs are more effective in reducing drug use than other adolescent drug prevention programs. Threats to valid inference about the causal moderating role of a variable may arise if substantively irrelevant factors are differentially associated with each level or category of the moderator variable under analysis. If the moderator variable (e.g., information/knowledge versus social influence programs) is confounded with characteristics of the design, setting, or population (e.g., urban versus rural schools), differences in the size or direction of a treatment effect brought about by the moderator cannot be distinguished from differential effects brought about by the potentially confounding variable.

Meta-analysts attempt to deal with confounding issues through statistical modeling (e.g., Tobler 1986, 1992) and through the use of within-study comparisons (e.g., Shapiro and Shapiro 1982). Within-study comparisons are particularly useful because they do not require making assumptions regarding the nature of the confounding. For instance, if the moderating role of prevention programs type A and B is at stake, a meta-analysis could be conducted of all the studies with internal comparisons of prevention programs A and B.

THREATS TO GENERALIZED INFERENCES

Research syntheses promise to generate findings that are more generalizable than those of single studies. Following Cronbach (1982) and Campbell and Stanley (1963), generalizations may involve universes of persons, treatments, outcomes, settings, and times. With respect to research syntheses, Cook (1990) distinguishes three separate though interrelated types of generalized inferences. The first concerns general-ized inferences about classes of persons, treatments, outcomes, settings, and times from which the reviewed studies were sampled. These are the generalizations that meta-analysts like to make; for instance, the effects of goal-setting programs (the treatment class) on drug use (the outcome class) among 8- to 12-year-olds (the target population) in public schools (the target setting class) during the 1980s (the target time).

The second type of generalized inferences concern generalizations across universes. Here, the issue is probing the robustness of a relationship across different populations of persons, different classes of interventions, different categories of settings, different outcome classes, and different time periods. When a relationship is not robust, the analyst seeks to specify the contingencies on which its appearance depends. At issue here are moderator variables, and of particular importance are moderator variables that specify the conditions under which a program has no effect or negative effects.

The third type of generalized inferences concern the generalizability of findings beyond the universes of persons, treatments, outcomes, and settings for which data are available. For example, can the effects of comprehensive prevention programs on the onset of drug use observed in school settings be generalized to church, YMCA, and prison settings? Are the effects of social influence programs observed during the 1970s and

1980s generalizable to programs to be implemented during the 1990s? In each of these examples, the issue is how one can justify inferences to novel universes of persons, treatments, outcomes, settings, and times on the basis of findings in other universes.

Generalizing on the basis of samples is most warranted when formal statistical sampling procedures have been used to draw the particular instances studied. That is, a sampling frame has been designed and instances have been selected with known probability. However, in meta-analyses the instances of person, samples, treatments, outcomes, settings, and times rarely if ever constitute probability samples from whatever universes were specified in the guiding research question. Nevertheless, Cook (1990) argues that generalized inferences about persons, treatments, outcomes, and settings can be tentatively justified even in the absence of random sampling. Cook discusses several principles for justifying generalized inferences in meta-analyses; two of these are further elaborated below. The first requires making a case for the proximal similarity of the sample and population (Campbell 1986). This requires identifying the prototypical, identity-inferring elements (Rosch 1978) of the target classes of persons, settings, causes, and effects and then examining whether they are adequately represented in the sample of studies entering a meta-analysis. In addition to the prototypical elements making a study relevant to a target universe, each individual study's setting, population, measure, and treatments are likely to have unique components that are not part of the target classes. It is crucial that these irrelevancies are made heterogeneous in the sample of studies entering a meta-analysis to avoid confounding prototypical and irrelevant characteristics (Campbell and Fiske 1957).

The second principle for generalizing when random selection cannot be assumed is empirical interpolation and extrapolation. Simply put, the more regularly intervention effects occur across different levels of an independent variable (e.g., length of intervention, type of counselor, type of school), the more tenable is the assumption that a causal effect can be extrapolated to not yet studied but related levels (e.g., shorter or longer interventions, different types of schools and counselors). The more dissimilar the yet unstudied levels are from the levels for which intervention effects have been examined, the more difficult interpolations and extrapolations are to justify. The wider and more diverse the conditions under which the intervention effects follow a predictable pattern, the more justified are generalizations to yet unstudied levels. Table 3 lists threats related to the different types of generalized inference desired in meta-analyses.

TABLE 3. *Threats to generalized inferences.*

(1) Unknown sampling probabilities associated with the set of persons, settings, treatments, outcomes, and times entering a meta-analysis
(2) Underrepresentation of prototypical attributes
(3) Failure to test for heterogeneity in effect sizes
(4) Lack of statistical power for studying disaggregated groups
(5) Restricted heterogeneity of substantively irrelevant aspects
(6) Confounding of subclasses with substantively irrelevant study characteristics
(7) Restricted heterogeneity of classes of populations, treatments, outcomes, settings, and times

Unknown Sampling Probabilities Associated With the Set of Persons, Settings, Treatments, Outcomes, and Times Entering a Meta-Analysis

One can rarely assume that the instances of persons, treatments, outcomes, settings, and times represented in a meta-analysis were randomly selected from the population of persons, settings, treatments, and outcomes to which generalization is desired. Even if there are random samples at the individual study level, it is rare that the studies entering into a meta-analysis constitute a formally representative sample of all such possible study-specific populations. The samples entering primary studies are chosen for proximal similarity and convenience rather than for reasons of formal sampling theory, and the studies containing these samples have an unknown relationship to all the studies that have been completed and that might be done on a particular topic. To tentatively justify generalized inferences in the absence of random sampling, the meta-analyst may follow the principles suggested by Cook (1990).

Underrepresentation of Prototypical Attributes

To demonstrate proximal similarity between a sample and its referent universe requires matching theoretically derived prototypical elements of the universe with the elements of the studies at hand. For substance abuse prevention programs, the question is whether the samples of students, prevention programs, settings, outcomes, and times examined in the reviewed studies represent the core attributes of the populations to which one is interested in generalizing. For instance, Hansen (1992) identified a group of school-based programs and labeled them "social influence programs." Hansen explicates that their "… primary purpose is to teach students about peer pressure and other social pressures and develop skills to resist these pressures" (p. 415). Thus, a meta-analysis of all the interventions that teach students about peer pressures but fail to include the development of skills to resist peer pressures might not constitute a social influence program. Consequently, such a meta-analysis would not allow generalized inferences to the target population of social influence programs. In a similar vein, program success could be explicated in terms of long-term abstinence from using illegal substances. A meta-analysis in which the majority of studies examine short-term effects, alcohol and tobacco use, the onset of drug use, and attitudes towards drugs would make questionable generalized inferences to the target population of outcomes (i.e., long-term abstinence from illegal substances).

Failure To Test for Heterogeneity in Effect Sizes

A statistical test for homogeneity has been developed (Hedges 1982; Rosenthal and Rubin 1982) that assesses whether the variability in effect estimates exceeds that expected from sampling error alone. Homogeneity tests play an important role in examining the robustness of a relationship and in initiating the search for factors that might moderate the relationship. If the homogeneity hypothesis is rejected, the implication is that subclasses of studies exist that differ in effect size. The failure to test for heterogeneity may result in lumping manifestly different subclasses of persons, treatments, outcomes, settings, or times into one category (i.e., apples-and-oranges problem). The heterogeneity test indicates when studies yield such different results that average effect sizes need to be disaggregated through blocking study characteristics that might explain the mean differences in effect size. Homogeneity tests also protect against searching for moderator variables when effects are robust.

Lack of Statistical Power for Studying Disaggregated Groups

If there is evidence that effect sizes are moderated by substantive variables of interest, then aggregated classes of treatments, outcomes, persons, or settings can be disaggregated to examine the conditions under which an effect changes in sign or magnitude. Such subgroup analyses rely on a smaller number of studies than main effect analyses and may involve additional statistical tests, thus lowering the statistical power for the subanalyses in question. Large samples mitigate against this problem, as do statistical tests adjusted to take into account the number of tests made. Even more useful are analyses based on aggregating within-study estimates of consequences of particular moderator variables.

Restricted Heterogeneity of Substantively Irrelevant Characteristics

Even if prototypical attributes of a universe are represented in the reviewed studies, a threat arises if a meta-analysis cannot demonstrate that the generalized inference holds across substantively irrelevant characteristics. For instance, if the reviewed studies on social influence programs were conducted by just one research team, relied on voluntary participation by students, depended on teachers and principals being highly motivated, or were all conducted in metropolitan areas of California, the threat would then arise that all conclusions about the general effectiveness of homework are confounded with substantively irrelevant aspects of the research context. To give an even more concrete example, if school-based programs were explicated to involve programs administered and implemented in school during grades 4 to 12, it is irrelevant whether the schools are in urban or rural settings, parochial or nonparochial schools, military schools, or elite academic schools. To generalize to school-based programs in the abstract requires being able to show that relationships are not limited to one or a few of these contexts—say, urban or Catholic schools.

The wider the range and the larger the number of substantively irrelevant aspects across which a finding is robust and the better moderating influences are understood, the stronger the belief that the finding will also hold under the influence of not yet examined contextual irrelevancies. Limited heterogeneity in substantively irrelevant variables will also impede the transfer of findings to new universes because it hinders the ability to demonstrate the robustness of a causal relationship across substantive irrelevancies of design, implementation, or measurement method. Tobler (1986) addresses the issue in examining whether program effects are robust regardless of substantively irrelevant characteristics of research design.

Confounding of Subclasses With Substantively Irrelevant Study Characteristics

Even if substantively irrelevant aspects are heterogeneous across studies, the possibility arises that subclasses of treatments, outcomes, settings, persons, or times are confounded with substantively irrelevant characteristics of studies. This situation arose in a meta-analysis of psychotherapy outcomes; differences in treatment effects were observed across different types of psychotherapy, but psychotherapy types were confounded with such substantively irrelevant research design features as the way psychotherapy outcomes were assessed (Wittmann and Matt 1986). This confounding impedes the ability to identify treatment type as a characteristic that moderates intervention effects.

Restricted Heterogeneity in Classes of Populations, Treatments, Outcomes, Settings, and Times

Generalizations across universes and generalizations to novel universes are facilitated if intervention effects can be studied for a large number and a wide range of persons, treatment, outcomes, settings, and times. This is the single most important potential strength of research syntheses over individual studies. For instance, a generalization to a novel universe of time is required if the question is whether school-based drug prevention programs developed and studied during the 1970s and 1980s can be expected to have similar effects in the 1990s. The confidence in such a generalization would be increased if one could demonstrate that the intervention effects were robust throughout the 1970s and 1980s, across different school settings, across different drugs, across different outcome measures, for students from different backgrounds, and so forth. The more robust the findings and the more heterogeneous the populations, settings, treatments, outcomes, and times in which they were observed, the greater the belief that similar findings will be observed beyond the populations studied.

SUMMARY AND CONCLUSIONS

Meta-analyses of drug prevention programs address questions regarding the causal relationship between prevention efforts and substance abuse. Different from primary studies of substance abuse prevention programs, meta-analyses involve generalized causal inferences. At issue are causal effects involving classes or universes of students, prevention programs, outcomes, settings, and times. This chapter presented threats to drawing such generalized inferences regarding bivariate causal and causal moderator relationships. The first group of threats concerns issues that could lead to erroneous conclusions regarding the existence of a relationship between a class of interventions and a class of outcomes. The second group concerns issues that may lead to erroneous conclusions regarding the causal nature of the relationship. Note that in all these instances, deficiencies in primary studies do not necessarily jeopardize the generalized inferences of a meta-analysis; in theory, such deficiencies may cancel each other out. A plausible threat only arises if deficiencies combine across studies to create a predominant bias. The third group of threats concerns issues that may lead to erroneous conclusions about the universes of persons, treatments, settings, outcomes, and times.

All validity threats are empirical products; they are the result of theories of method and the practice of research. Consequently, no list of validity threats is definite. Threats are expected to change as theories of method are improved and more is learned about the practice of research synthesis. All threats are potential; the existence of a threat by itself does not make it a plausible alternative explanation to a causal claim. Research synthesists have to use the empirical evidence, logic, common sense, and any background information available to determine whether a potential threat indeed provides a plausible alternative explanation.

REFERENCES

Bangert-Drowns, R.L. The effects of school-based substance abuse education—a meta-analysis. *J Drug Educ* 18:243-264, 1988.

Bryk, A.S., and Raudenbush, S.W. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newbury Park, CA: Sage, 1992.

Campbell, D.T. Relabeling internal and external validity for applied social scientists. In: Trochim, W.M.K., ed. *Advances in Quasi-Experimental Design and Analysis*. San Francisco: Jossey-Bass, 1986.

Campbell, D.T., and Fiske, D.W. Covergent and discriminant validation by the multitrait-multimethod matrix. *Psychol Bull* 59:81-105, 1957.

Campbell, D.T., and Stanley, J.C. Experimental and quasi-experimental designs for research on teaching. In: Gage, N.L., ed. *Handbook of Research on Teaching*. Chicago: Rand McNally, 1963.

Cook, T.D. The generalization of causal connections: Multiple theories in search of clear practice. In: Sechrest, L.; Perrin, E; and Bunker, J., eds. *Research Methodology: Strengthening Causal Interpretations of Nonexperimental Data*. DHHS Publication No. (PHS) 90-3454. Washington, DC: U.S. Department of Health and Human Services, 1990.

Cook, T.D., and Campbell, D.T. *Quasi-experimentation. Design and Analysis Issues for Field Settings*. Boston: Houghton Mifflin Company, 1979.

Cooper, H.M. *Integrating Research: A Guide for Literature Reviews*. Newbury Park, CA: Sage, 1989.

Cronbach, L.J. *Designing Evaluations of Educational and Social Programs*. San Francisco: Jossey-Bass, 1982.

Greenwald, A.G. Consequences of prejudice against the null hypothesis. *Psychol Bull* 82:1-20, 1975.

Hansen, W.B. School-based substance abuse prevention: A review of the state of the art in curriculum, 1980-1990. *Health Educ Res Theory Pract* 7:403-430, 1992.

Hedges, L.V. Estimation of effect sizes from a series of independent experiments. *Psychol Bull* 92:490-499, 1982.

Hedges, L.V. Directions for future methodology. In: Wachter, K.W., and Straf, M.L., eds. *The Future of Meta-Analysis.* New York: Russell Sage Foundation, 1990.

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

Hunter, J.E., and Schmidt, F.L. *Methods of Meta-Analysis. Correcting Error and Bias in Research Findings*. Newbury Park, CA: Sage, 1990.

Kulik, J.A., and Kulik, C.-L.C. "Operative and Interpretable Effect Sizes in Meta-Analysis." Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 16-20, 1986.

Little, R.J.A., and Rubin, D.B. *Statistical Analysis with Missing Data*. New York: Wiley, 1987.

Matt, G.E. Decision rules for selecting effect sizes in meta-analysis: A review and reanalysis of psychotherapy outcome studies. *Psychol Bull* 105:106-115, 1989.

Matt, G.E., and Cook, T.D. Threats to the validity of research syntheses. In: Cooper, H., and Hedges, L.V., eds. *The Handbook of Research Synthesis.* New York: Russell Sage Foundation, 1993.

Premack, S.L., and Hunter, J.E. Individual unionization decisions. *Psychol Bull* 103:223-234, 1988.

Raudenbush, S.W.; Becker, B.J.; and Kalaian, H. Modeling multivariate effect sizes. *Psychol Bull* 103:111-120, 1988.

Rosch, E. Principles in categorization. In: Rosch, E., and Lloyd, B.B., eds. *Cognition and Categorization*. Hillsdale, NJ: Erlbaum, 1978.

Rosenthal, R. The ‘file drawer problem’ and tolerance for null results. *Psychol Bull* 86:638-641, 1979.

Rosenthal, R. *Meta-Analytic Procedures for Social Research*. Beverly Hills, CA: Sage, 1984.

Rosenthal, R., and Rubin, D.B. Comparing effect sizes of independent studies. *Psychol Bull* 22:500-504, 1982.

Rosenthal, R., and Rubin, D.B. Meta-analytic procedures for combining studies with multiple effect sizes. *Psychol Bull* 99:400-406, 1986.

Rubin, D.B. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley, 1987.

Shadish, W.R., Jr. Do family and marital therapies change what people do? A meta-analysis of behavioral outcomes. In: Cook, T.D.; Cooper, H.M.; Cordray, D.S.; Hartman, 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.

Shapiro, D.A., and Shapiro, D. Meta-analysis of comparative therapy outcome research: A replication and refinement. *Psychol Bull* 92:581-604, 1982.

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

Tobler, N.S. Drug prevention programs can work: Research findings. *J Addict Dis* 11:1-27, 1992.

Wittmann, W.W., and Matt, G.E. Meta-Analyse als Integration von Forschungsergebnissen am Beispiel deutschsprachiger Arbeiten zur Effektivität von Psychotherapie [Integration of German-language psychotherapy outcome studies through meta-analysis.]. *Psychologische Rundschau* 37:20-40, 1986.

AUTHOR

Georg E. Matt, Ph.D.
Department of Psychology
San Diego State University
San Diego, CA 92182-4611

Issues in Classification in Meta-Analysis in Substance Abuse Prevention Research

William B. Hansen and Lynn A. Rose

Meta-analysis as a method holds great promise for allowing fields of research to accomplish synthesis and integration of findings. This goal must be compared to experimental research, which is inherently reductionistic in its approach. Because of this divergence in methods and implicit goals, meta-analysts are often faced with a need to reconceptualize original research in order to fit it into a method that allows comparison. The authors have identified two such issues (classification of variables and classification of correlational results) that will pose continued dilemmas for meta-analysts.

This chapter has two goals. The first is to discuss strategies to create schema for classifying independent variables. The second is to discuss issues of classifying types of correlational relationships between independent and dependent variables. Both have practical relevance for incorporating theory into meta-analytic practice.

CLASSIFYING INDEPENDENT VARIABLES

Creating a classification schema for independent variables is a major dilemma for meta-analysts. Yet, inevitably, completing such work is among the first steps that one must take in beginning a meta-analysis. The resulting schema will ultimately determine much of the meaning that emerges from subsequent analyses.

An extensive literature about analytic methods has emerged. Topics include attention to effect size estimates (Glass et al. 1981; Hall et al. 1994) and controls for methodological problems (Cook et al. 1992; Wortman 1994). What is being analyzed, which lies at the root of wanting to complete analyses in the first place, has unfortunately received less attention (Cooper 1990; Orwin 1994; Stock 1994). It is the authors' observation that creating links with theory in completing meta-analyses gives meaning and value to the methods.

Categorization of independent variables is challenging because no single theory captures all available variables. Theories that guide reductionist research focus only on relevant variables, ignoring variables that are perceived to be irrelevant. Theoretical traditions also reflect diverse scientific disciplines. Meta-analysis needs to account for this diversity and the diversity limits a meta-analyst's ability to use theory a priori. Of necessity, linking theory and categorization in meta-analysis must be done post hoc.

Creation of an Inclusive Categorization Schema

The authors faced the problem of categorization in an analysis of 242 drug abuse correlational studies. The analysis reflects a truly evolutionary process. In the initial phases of meta-analytic work, started in 1986, the authors began abstracting research articles that included correlates of substance use. At that time, there was an a priori interest in topics relevant to current substance use prevention curriculum development. Seven categories of independent variables were created: (1) peer use, (2) parent use, (3) sibling use, (4) rebelliousness, (5) attitudes about substances, (6) normative beliefs, and (7) miscellaneous other variables. The miscellaneous other variables category was divided into subcategories. If a reported independent variable failed to fit within an existing subcategory, a new variable category was created. By the time approximately 100 studies had been entered, nearly 50 miscellaneous subcategories had been created.

The dataset was initially created to answer a limited number of questions about variables relevant to social influence-based substance abuse prevention program development (Hansen 1988). The miscellaneous other variables category was initially ignored for analysis purposes. However, as the number of studies grew, it became clear that a refinement of miscellaneous subcategories was needed. All variables, including those originally grouped in the initial six categories, were recategorized. The first goal of this recategorization was to organize the independent variables with greater precision. The second goal of the reclassification was to create a broad categorical matrix into which measures from newly identified articles could be readily classified. It was apparent that major groupings of variables might be possible. These groups would not be expected to follow the organization of a specific theory. In some respects, these groups were expected to create meta-theoretical constructs that would apply to an entire field.

Ultimately, a two-tier classification system was developed. In the initial tier, 12 categories were established. These included: (1) previous substance use, (2) intentions to use substances, (3) cognitive factors, (4) competency factors, (5) personality factors, (6) use by others, (7) social pressures, (8) institutional affiliations, (9) peer structure, (10) home and family structure, (11) demographic factors, and (12) miscellaneous factors. For each of these 12 major categories, a second tier of subcategories was created (these are described in detail below). The development of subcategories proceeded by examining the descriptions of measures within each group in the first tier and making logical subdivisions where appropriate. Research has made little or no attempt to insure that all potential variables are used equally. The uneven pace of normal research guarantees that some variables will be used frequently and will be highly similar in structure and content. Such variables can be rapidly reduced to the most elemental concepts. Other variables are rarely used, differ markedly in format and meaning, and do not group easily. Both prevalence of an item and similarity of concept were used to create subcategories. When sufficiently large numbers of similar items were identified, they were joined into initial subcategories. If, in comparison to the size of other subcategories, extremely large numbers of cases were present, further logical divisions were attempted until additional subclassification would result in too few cases for analysis. This left numerous items that remained together as loose constructs because similarity and frequency were not sufficient for more precise categorization.

Substance use measures were identified as dependent variables. A number of studies reported the correlation among various substances. To enable analyses, one of the two measures was identified as a dependent variable and one as an independent variable. The variable identified as the dependent variable was either the variable that was measured first (in the case of longitudinal studies) or the most prevalent substance. As an example of the second case, drinking alcohol is generally more prevalent than steroid use. Alcohol would have been identified as the dependent variable in a case where correlations between the two substances were observed. Previous substance use included six subcategories of substances as independent variables: tobacco, alcohol, marijuana, other single drugs, other combined drugs, and being drunk.

Intentions measures included the expected probability of future consumption as well as measures of commitment toward limiting future use and abuse of substances. Intention measures, almost without exception, focused on intentions to use a specific substance. Five subcategories were defined by the substance about which the intention was assessed: tobacco, alcohol, marijuana, other single drugs, or combined drugs.

Measures of cognitive factors addressed beliefs (including knowledge), attitudes, and values. Seven subcategories were developed. Belief and knowledge items were sufficiently prevalent to create three distinct subgroups: beliefs about health consequences, beliefs about social consequences, and beliefs about psychological consequences. Items relevant to values were also sufficiently well represented to create three distinct subgroups: general values, religious values, and values related to achievement. Attitudes about drugs formed the seventh subcategory.

Measures of competency were subcategorized into five groups. Intelligence test scores, primarily from standardized tests, formed a category that was distinguishable from other competency measures. School performance, including grades as well as standardized achievement tests, formed the second group. Self-efficacy, the perceived ability to deal with a variety of social situations including (but not limited to) peer pressure, was a rather heterogeneous subcategory. Decisionmaking skills and stress management skills had sufficient definition in the measures that two clearly defined categories could be created.

Personality factors were grouped into seven subcategories. Personality variables were broadly defined as those that reflected a personal trait or characteristic other than competence. The attribution on the part of researchers was to ascribe relatively stable psychological characteristics to individuals. Several subcategories were separated from the general concept of personality because of the prevalence of highly similar measures. For example, self-esteem, affect (characteristic mood), and locus of control were constructs that were identified frequently enough to create a sizable number of indicators that were specific to each. In the case of self-esteem and locus of control, numerous studies reported similar measures for defining each construct.

Independence and deviance are often thought to be highly related. Sufficient numbers of measures were included that it was possible to separate each construct. Measures of independence included the expressed need for or value of independence, sensation seeking, and risk taking. Deviance measures included definable observances of violence, antisocial behavior, and delinquency.

The remaining personality measures fit into two categories. The first group contained those items that reflected other psychological characteristics of individuals not grouped into the subcategories listed above. These intrapsychic characteristics were distinguished from personality characteristics that described an individual's social personality. Examples of the latter include gregariousness and likability.

The sixth overall category of variables included a variety of institutional affiliations. Three distinguishable subcategories, church attendance, religious affiliation, and moral codes, were created. Church attendance refers to religious practice. Religious affiliation was often noted as the type of religion to which the respondent belonged. Moral codes referred to a belief in or attitude toward a general or specific religious or other moral code. Two additional subcategories, school bonding and academic expectations, were also defined. School bonding reflected a feeling of acceptance by the school as an institution. Academic expectations reflected the hopes and desires of others regarding an individual's academic performance. Finally, two related but distinguishable subcategories, structured and nonstructured activities, were created. Structured activities included self-reports about the extent of participation in extracurricular sports, music, hobbies, and other supervised activities. Nonstructured activities included self-reports about hanging out, spending time in the neighborhood, and other activities that implied or specified a lack of adult supervision.

Use by others was the seventh broad categorical group of variables. Five subcategories were identified. Three of the five were relatively easy to define: drug use by peers (same age, older, and younger friends and acquaintances), drug use by parents, and drug use by siblings. A fourth subcategory included drug use by extended relatives (aunts, uncles, grandparents, and cousins). The fifth subcategory included perceptions about drug availability. These measures typically included ratings of the frequency with which drug use was observed in the community as well as the ease or difficulty of obtaining substances. It is noteworthy that these measures all included perceptions of prevalence that were broad and general as well as those that were specific; it is likely that broad and general perceptions are more likely to be biased by perceptual processes, reducing the degree to which actual use among others is accurately measured. The distinction between perception and documented occurrence was not pursued in classification.

The social pressures category included seven definable subcategories. The first subcategory included reports of receiving offers to use substances from peers as well as parents and nonspecified or miscellaneous others. The original intent was to use the source of the offer to define a specific subcategory. However, there were too few examples to make separate subcategories, and a general offers category emerged instead. The second subcategory included reports about an individual's motivation to comply with social pressures to use substances. Peers' attitudes about drug use, parents' attitudes about drug use, and others' attitudes about drug use each constituted three separate subcategories. Others' attitudes (including parents and peers as well as miscellaneous or nonspecified others) about topics other than substance use (e.g., violence) were also identified. The final subcategory included variables that attempted to measure exposure to or influence from mass media sources related to substance use.

The ninth category of independent variables included peer structure. The first subcategory was labeled "peer group characteristics." This set of measures included various descriptive indices that characterized peer groups as rebellious, risk taking, religious, academically oriented, and so forth. The second subcategory assessed the balance between peer and parent influence, often through self-reports of the respondent. These measures addressed which source (peer or parent) of social influence predominated as well as indices assessing the extent of conflict between parents and the peer group. The third subcategory included the level of intimacy that existed between the respondent and other teens, primarily of the opposite sex. The primary measure included in this factor was self-reports of sexual intercourse. The final peer factor subcategory included assessments of social bonding and attachment to the peer group. In some instances, this included a simultaneous assessment of the positive or negative nature of the peer group. However, these measures primarily addressed the degree to which the adolescent perceived himself or herself to be accepted by or belong to a group of friends.

The 10th major category of variables assessed a variety of home and family structures. This was developed to be as similar as possible to the peer structure category described above. Unfortunately, there were few parallel comparisons across studies. Six subcategories were therefore created. These included parents’ psychological traits, which roughly correspond to peer group characteristics but included measures of clinical personality characteristics as well. Parental relations roughly paralleled measures of peer bonding that assessed feelings of attachment and caring from and for parents. Additional family measures were also identified. The third home factors subcategory included measures that assessed the viability of parents’ marriage. The fourth subcategory included measures of parents’ educational achievement. The fifth subcategory included descriptive measures of the composition of the family, including descriptions of who lived at home. The final home factors subscale assessed participants’ socioeconomic status, including income as well as surrogate measures (e.g., Hollingshead measures).

Demographic information formed an 11th major category. Gender, age, ethnicity of the sample, and geographic identifiers (such as urban-suburban-rural distinctions as well as geopolitical location) were included as subcategories.

Finally, a miscellaneous category was created to include variables that did not fit within any of the other major categories. Included in this were two substantive subcategories (political involvement/social activism and exposure to school information or formal programs). In addition, a truly miscellaneous subcategory that included all other measures was created.

Mapping Classification to Existing Classification Schema

The database created for meta-analytic purposes was broad and comprehensive. It was assumed at the outset that the database could be used to answer a variety of questions. Not only can the database be used to generate summary findings, it is also possible that the database could be used to compare previous work with work in progress. However, the authors learned that a second order of manipulation was needed to complete such tasks. As is usual in the case of research, individual projects address only a limited number of project-specific variables, the construction of which is typically dictated by project-specific theoretical issues. As a result, referencing the meta-theoretical database presented unique problems when the authors began using it to examine convergence with findings from an empirical study.

A review of school-based curriculums (Hansen 1992) identified 12 curriculum approaches common to intervention. Each approach implicitly addressed a mediating variable that has been postulated to account for substance use. As a result of the review, a project was funded to examine the potential of each of the 12 postulated mediating variables. In this study, scales were developed to measure each of the following postulated mediating processes: (1) beliefs about susceptibility to the consequences of using substances, (2) decisionmaking skills, (3) stress management skills, (4) social skills, (5) goal-setting skills, (6) beliefs about alternatives to using substances, (7) self-esteem, (8) skills for resisting peer pressure (self-efficacy), (9) skills for getting and providing assistance for solving problems, (10) normative beliefs about the prevalence and acceptability of substance use, (11) perceptions that substance use would interfere with personal values and lifestyle, and (12) a strong personal commitment to not use substances. Data have been collected on three occasions, each 12 months apart, using these measures and measures of substance use.

The review and the followup study that examined postulated mediating variables were developed independently of the creation of either the classification schema or the meta-analytic database. Connecting the two was not planned. Nonetheless, the presence of both datasets provided an opportunity to examine the potential of the meta-analytic database to be used as a source of cross-validation of initial findings from the empirical project.

The initial step for completing a comparison between a study and the database findings was to find variables in each that provided some degree of correspondence. Table 1 presents the measures for which correspondence appeared appropriate.

Corresponding concepts were identified in the meta-analytic database for all but three of the variables in the ongoing study. In the ongoing study, social skills specifically discussed skills for communicating and resolving interpersonal differences. The nearest corresponding variable in the meta-analytic database was social personality traits. However, this subcategory included more personality measures than skill measures, and many were not relevant. Goal-setting skills also failed to find a match. Achievement values consisted predominantly of motivation and aspiration measures, few of which attempted to assess skills per se. It was felt that achievement values, general values, and religious values corresponded more closely with the ongoing study's variable that addressed incongruence between values and lifestyle and substance use. A match between goal-setting skills and a meta-analysis category was not available. Finally, measures of skills for getting and providing assistance were not observed in the creation of the meta-analysis database. Thus, while incomplete, it was felt that the correspondence between two datasets would prove useful for comparison purposes.

TABLE 1. *Corresponding measures from the ongoing study and the meta-analytic database.*

Ongoing study	Meta-analytic database
Beliefs about suspectibility to the consequences of using substances	Beliefs about social consequences Beliefs about health consequences Beliefs about psychological effects
Decisionmaking skills	Decisionmaking skills
Stress management skills	Stress management skills

Social skills Goal-setting skills	
Beliefs about alternatives to using substances Self-esteem	Participating in structured activities Participating in nonstructured activities Self-esteem
Skills for resisting peer pressure (self-efficacy) Skills for getting and providing assistance for solving problems	Skills for resisting peer pressure (self-efficacy)
Normative beliefs about the prevalence and acceptability of substance use Perceptions that substance use would interfere with personal values and lifestyle	Peer drug use Peer drug attitudes Achievement values General values Religious values
Commitment to not use substances	Intentions/commitment

CLASSIFYING CORRELATIONAL RELATIONSHIPS

The second issue of classification emerged as comparisons were attempted. The meta-analytic database included four types of measures: (1) correlation coefficients (e.g., Pearson r, phi); (2) odds and risk ratios; (3) multivariate coefficients (e.g., standardized regression weights); and (4) group mean comparison statistics (analysis of variance (ANOVA), multivariate ANOVA (MANOVA)). Of these, correlation coefficients were most prevalent, provided greatest standardization, and were most similar to analyses already available from the empirical study (Hansen, under review).

Correlation coefficients pose an additional problem that is related to classification in meta-analysis. Specifically, the issue of calculating and reporting positive and negative signs for correlational values is problematic. The sign of the correlation coefficient is dependent upon four independent factors: (1) the scaling of the independent variable, (2) the scaling of the dependent variable, (3) the empirical relationship, and (4) the actions of the investigator in reporting the findings. Table 2 presents the expected sign values that correspond to different combinations of factors (1) and (2), scaling of the independent and dependent variables.

TABLE 2. *Criteria used for classifying correlation coefficients.*

		Independent variable scaling	
		High = • More of a theoretically undesirable trait or situation • Less of a theoretically desirable trait or situation	High = • More of a theoretically desirable trait or situation • Less of a theoretically undesirable trait or situation
Dependent variable scaling	High = High drug use	Expected correlation Positive Type 1	Expected correlation Negative Type 2
	High = Low drug use	Expected correlation Negative Type 3	Expected correlation Positive Type 4

When the independent variable is scaled so that high values are theoretically less desirable and dependent variable high values are theoretically undesirable (high drug use), the correlation is expected to be negative (type 1). For example, it might be hypothesized that high academic achievement (a socially and theoretically desirable trait) would be inversely related to high drug use. If both are measured so that high values represent fulfilling each condition, a negative correlation coefficient is expected. However, it must be remembered that scaling is relatively arbitrary in social science. Either or both scales can be inverted by either reversing the response categories or by multiplying the final values by negative one (-1). There are no rules that all researchers and research teams follow. If the academic achievement variable, still scored so that high was better, and the drug use (dependent) variable were reversed (e.g., in the case of measuring the degree of abstinence rather than use), the sign would be expected to reverse (type 3). More often, it appears that investigators vary the ordering of the independent variable.

The third factor that may influence the sign of the correlation is the empirical situation. For example, figure 1 presents a theoretical distribution of correlation coefficients based on fictitious data. (For the sake of argument, assume that this is a type 1 coefficient as defined in table 1.) As with all measurement phenomena, there is a distribution of scores and, in this case, some scores are negative. Even though a positive correlational value is expected (the mean of correlations is positive), some values will be negative. In meta-analytic terms, this may be due to differences in populations, differences in methods (specific measures selected), or chance findings.

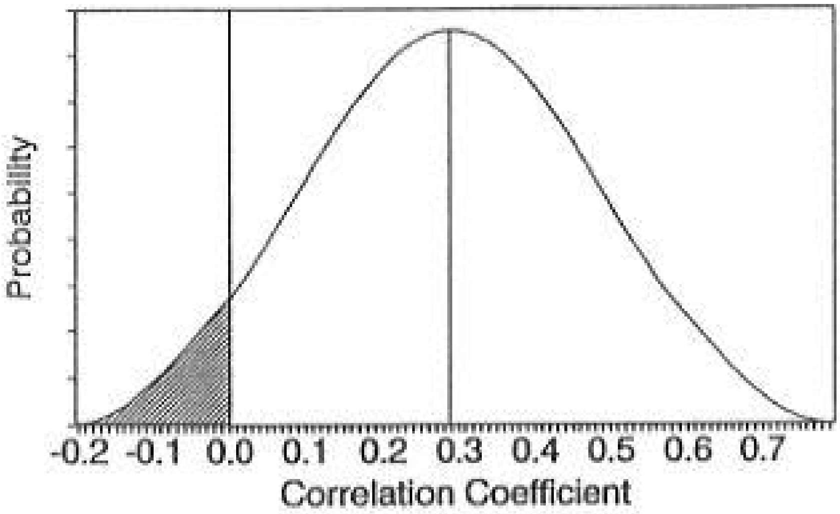


FIGURE 1. *Theoretical sampling distribution of correlation coefficients.*

The final factor that may influence the sign of the correlation coefficient is manipulation on the part of the investigator. When findings are presented, there are occasionally reasons to alter the sign. In most (but not all) cases, this occurs as a transformation to the positive sign. The purpose appears to be ease of presentation on the part of the researcher. The details that underlie the justification for selection of directionality for any given scale may be complicated. It is possible, for instance, that no theoretical model exists for ordering the direction of a scale. The use of multiple scales with mixed directionality may be simplified by creating a uniform direction for presentation purposes. Whatever the intent or reason, it is clear that such practices occur relatively often. Unfortunately for the meta-analyst, such transformations are often undocumented.

What has resulted in the field of substance abuse research is the presentation of correlational data that are relatively noncomparable. Not only is variable scaling often not described sufficiently to inform the reader, unexpected findings are often not highlighted and investigator-induced transformations not documented. This left the authors with a serious dilemma and two options. First, there was the possibility of examining the literature by individual result to determine which of the four types of correlations, adjusted for apparent transformation by the researcher, existed. The authors are actually pursuing this strategy, but it is time consuming and may not result in perfect classification. Second, the authors could arbitrarily transform all the data. Given the time constraints under which this chapter was developed, the authors adopted the latter strategy. All correlation coefficients were transformed to positive values.

Implications of Transforming All Correlations To Be Positive

Before presenting the findings, the implications of this transformation should be clearly documented and understood. Given a theoretical spread of correlation coefficients that corresponds roughly to those presented in figure 1, the transformation of values had a relatively predictable effect. Figure 2 presents the same data with the negative values folded over the positive values. It is apparent that the distribution of values became skewed and the mean of the distribution was inflated. However, in the case presented, the increase in the mean is only slight. Had the distribution of the available values been smaller (i.e., all above zero), no inflation would have been seen at all.

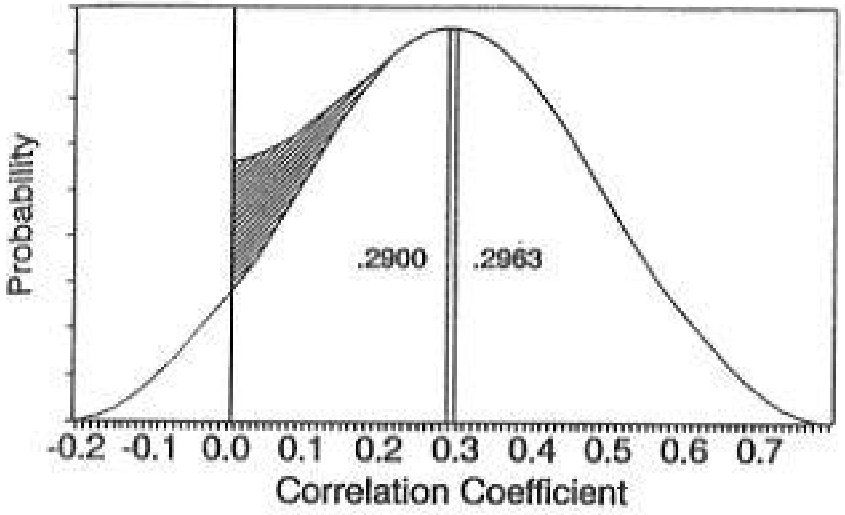


FIGURE 2. *Two theoretical sampling distributions of correlation coefficients with a relatively high true mean correlation (r = 0.29), one allowing and one not allowing negative correlation coefficients.*

Figure 3 presents the same fictitious distribution but with the mean of correlations lowered from 0.2900 to 0.0963. It is readily apparent that in this case, using absolute values of correlations greatly increases skewness and vastly inflates the mean.

These examples illustrate the difficulty of the approach. This procedure violates fundamental statistical assumptions, but with relatively well-known effects. The essential problem that emerges is that values close to zero are expected to be grossly inflated. At the same time, high mean correlations are expected to be relatively accurately portrayed. The point at which confidence is restored is related to skewness and variance. Practical experience from the meta-analysis suggests that correlations of 0.30 and higher experience little inflation and are expected to be accurate.

The utility of this approach is that it allows the identification of correlations that are likely to be valuable for the development of prevention programming. Small correlations are presumed to indicate weak causal linkages. Large correlations are presumed to indicate strong causal

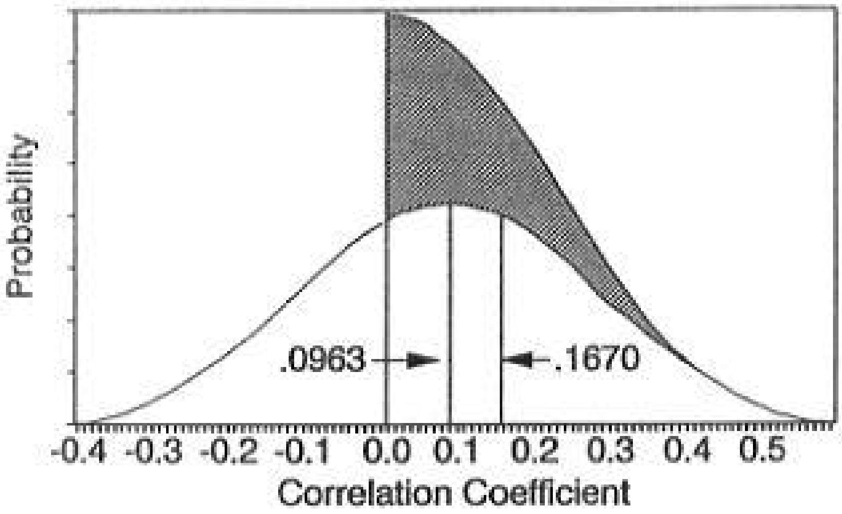


FIGURE 3. *Two theoretical sampling distributions of correlation coefficients with a relatively low true mean correlation (r = 0.0963), one allowing and one not allowing negative correlation coefficients.*

linkages. The later are of most interest and, at least in this case, are most likely to be accurate indicators of the true underlying mean.

CORRESPONDENCE BETWEEN ONGOING AND DATABASE FINDINGS

The ability to compare classification strategies and data makes meta-analytic findings useful. To demonstrate the utility of these strategies, two sets of findings were compared. From the meta-analytic database, variables that corresponded to those in an ongoing empirical study were selected and compared. Because absolute values of correlation coefficients were included in the meta-analytic database, all values are presented as positive numbers. In the case of the ongoing study, this involved an absolute value transformation of otherwise negative (type 2) values.

Table 3 presents cross-sectional data from both the ongoing study of 12 mediating variables and data amassed from the meta-analytic database. Results indicate that there is relatively high concordance among findings. In both cases, commitment and intentions were relatively strong predictors. Similarly, peer drug use and peer drug attitudes, elements of which were captured in the study’s measure of normative beliefs, demonstrated a relatively high correlation with substance use. The ongoing study yielded a relatively high correlation between beliefs about consequences and substance use. Among analyses included in the meta-analysis, only beliefs about social consequences and psychological effects yielded a comparable relationship. Beliefs about health consequences had a lower relationship in the meta-analytic database.

TABLE 3. *Correspondence between meta-analytic and ongoing findings; cross-sectional correlations with substance use; average of alcohol, tobacco, and other substances.*

Meta-analytic findings				Ongoing findings	
	N	Mean	SD		Mean
Commitment	25	0.36	0.11	Commitment	0.42
Health beliefs	87	0.15	0.10	Beliefs about	
Social beliefs	59	0.32	0.19	cons	0.33
Psychological beliefs	104	0.30	0.18	equences	
General values	75	0.21	0.16	Incongruence	
Religious values	30	0.20	0.12	between values and	
Achievement values	42	0.21	0.09	substance use	0.38
Decisionmaking	15	0.14	0.17	Decisionmaking	0.16
Stress management	45	0.25	0.19	Stress management	
					0.14
Self-efficacy	26	0.32	0.20	Self-efficacy	0.27
Self-esteem	72	0.16	0.11	Self-esteem	0.19
Structured activities	69	0.17	0.10	Alternatives	0.28
Nonstructured activities	58	0.27	0.12		
Peer drug use	386	0.36	0.18	Normative beliefs	0.44
Peer drug attitudes	82	0.36	0.25		

KEY: SD = standard deviation.

Other variables that appeared to have similar magnitudes of correlations were decisionmaking skills, self-efficacy for resisting peer pressure, and self-esteem. The alternatives measure appeared to be closest in magnitude to reports of participating in nonstructured activities and was markedly higher than the meta-analytic finding for participating in structured activities.

Several variables had low correspondence. The measure of perceived incongruence between values, lifestyle, and substance use was relatively strong, whereas each of the three categories of values-oriented measures from the meta-analytic database were less predictive. In part this may reflect a different way of measuring this construct with an emphasis on the incongruence rather than the presence or absence of any given value.

The authors also observed a smaller correlation coefficient for stress management than was observed generally. This suggests that the role of stress management skills might be generally more important as a predictor than reflected in the authors' ongoing research.

Longitudinal analyses were also compared (see table 4). For values from the meta-analytic database, all data that involved comparing earlier measures of the predictor with later measures of substance use qualified the data as longitudinal. This resulted in significant variability in the time lag between measures, which is ignored in these analyses. For the ongoing study, measurement of mediators and substance use is delayed 12 months.

TABLE 4. *Correspondence between meta-analytic and ongoing findings; longitudinal correlations with substance use; average of alcohol, tobacco, and other substances.*

Meta-analytic findings				Ongoing findings	
	N	Mean	SD		Mean
Commitment	35	0.17	0.20	Commitment	0.38
Health beliefs	43	0.15	0.11	Beliefs about	cons equences
Social beliefs	23	0.16	0.12		0.30
Psychological beliefs	24	0.15	0.10		
General values	13	0.12	0.07	Incongruence	between values and substance use
Religious values	10	0.20	0.08		
Achievement values	19	0.24	0.09		0.37
Decisionmaking	0	---	---	Decisionmaking	0.11
Stress management	42	0.17	0.11	Stress management	
					0.09
Self-efficacy	15	0.30	0.21	Self-efficacy	0.23
Self-esteem	40	0.15	0.11	Self-esteem	0.16
Structured activities	44	0.22	0.19	Alternatives	0.11
Nonstructured activities	5	0.18	0.13		
Peer drug use	109	0.28	0.17	Normative beliefs	0.40
Peer drug attitudes	55	0.26	0.16		

A generally consistent pattern of relationships was observed in these analyses. Peer drug use, peer attitudes, and normative beliefs measures were similar in magnitude. This convergence suggests that these variables are strong longitudinal predictors of substance use. The magnitude of self-efficacy to resist peer pressure as a predictor in both datasets also remained relatively strong. In contrast to the ongoing study's findings about commitment, beliefs, and values, the meta-analytic longitudinal correlation coefficients were markedly smaller. The authors' measure of alternatives was a relatively weak longitudinal predictor of substance use. Both measures from the meta-analytic database were somewhat stronger, albeit in a moderate range for longitudinal findings. Self-esteem remained a weak predictor of substance use in both datasets. There were no longitudinal studies in the meta-analytic database that examined skill at decisionmaking as a predictor of substance use.

CONCLUSION

The purpose of this chapter was to examine classification issues in meta-analysis. Classification is an inherent underlying activity that receives little attention. Nonetheless, without a well-conceived classification schema at the base of meta-analysis, the theoretical implications of specific analyses lose their meaning. Two specific issues in classification were addressed: classifying variables for analysis and understanding correlation coefficients needed in analysis.

The measurement typology classification schema that resulted in the creation of the database is meta-theoretic in nature. That is, no single theory includes all variable classes. The classification schema appears to be useful in that a diversity of studies and variables can be incorporated within it. The authors nonetheless recognize that the classification model is at least partly dependent upon the topic being studied (substance use), the existing theories that have driven prior research and influenced the development of measures, and the amount of detail that exists in the available studies. The overall pattern of classification involved identifying successive hierarchies of variables, with each level of nesting emerging as sufficient numbers of cases were observed. A two-tier hierarchy was presented. It might have been as easily considered a three-tier hierarchy with some elements complete and some incomplete. With sufficient data from the field, it may be possible to create a full three-tier or four-tier classification schema, the progression being dependent upon refinement of measures and theoretical constructs and the availability of sufficiently large numbers of cases.

A distinct but equally perplexing problem exists for classifying correlational relationships. The field has not progressed sufficiently for a clear typology of relationships to have become standard for presenting data. Four independent elements (the scaling of the independent and dependent variables, the empirically observed relationship, and the needs of the investigator for presentation of findings) were identified as barriers to the consistent application of comparable methods for presenting correlational findings. Of the two available solutions, transformation of all values to positive numbers is the easiest to complete. This method provides inflated estimates of the average correlation coefficient, particularly when correlational values are near zero. Given the difficulties in completing topologies, this method produces results that have utility. Refinements in reporting will significantly improve the ability of meta-analysts to resolve this dilemma.

REFERENCES

Cook, T.D.; Cooper, H.; 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.

Cooper, H. On the social psychology of using research reviews. In: Wachter, K.W.,and Straf, M.L., eds. *The Future of Meta-analysis*. New York: Russell Sage Foundation, 1990. pp. 75-88.

Glass, G.V.; McGaw, B.; and Smith, M.L. *Meta-analysis in Social Research*. Beverly Hills, CA: Sage, 1981.

Hall, J.A.; Rosenthal, R.; Tickle-Degnan, L.; and Mosteller, F. Hypotheses and problems in research synthesis. In: Cooper, H., and Hedges, L.V., eds. *The Handbook of Research Synthesis*. New York: Russell Sage Foundation, 1994. pp. 17-28.

Hansen, W.B. Theory and implementation of the Social Influence Model of primary prevention. *Prev Res Findings* 3:93-107, 1988.

Hansen, W.B. School-based substance abuse prevention: A review of the state of the art curriculum, 1980-1990. *Health Educ Res* 7(3):403-430, 1992.

Hansen, W.B. Comparison of postulated mediators of substance use prevention: A longitudinal examination. *Prev Med*, under review.

Orwin, R.G. Evaluating coding decisions. In: Cooper, H., and Hedges, L.V., eds. *The Handbook of Research Synthesis*. New York: Russell Sage Foundation, 1994. pp. 140-162.

Stock, W.A. Systematic coding for research synthesis. In: Cooper, H., and Hedges, L.V., eds. *The Handbook of Research Synthesis*. New York: Russell Sage Foundation, 1994. pp. 125-138.

Wortman, P.M. Judging research quality. In: Cooper, H., and Hedges, L.V., eds. *The Handbook of Research Synthesis*. New York: Russell Sage Foundation, 1994. pp. 97-110.

AUTHORS

William B. Hansen, Ph.D.

Associate Professor

Lynn A. Rose, M.P.H.
Associate Professor
Bowman Gray School of Medicine
Medical Center Boulevard
Winston-Salem, NC 27157-1068

Improving Meta-Analysis for Policy Purposes

Larry V. Hedges

Many empirical sciences have developed formal methods of combining information across independent research studies, an enterprise with a long history that was named "meta-analysis" (Hedges 1992; National Research Council 1992). When the question to be addressed is a narrow scientific one, the standard methods of meta-analysis provide adequate tools for combining the evidence. These are discussed in Cooper (1984), Hedges and Olkin (1985), Hunter and Schmidt (1990), Light and Pillemer (1984), Rosenthal (1984), or the new "Handbook of Research Synthesis" (Cooper and Hedges 1994) which includes contributions by all the authors previously mentioned.

Society is entering an era in which systematic research syntheses reasonably can be expected to contribute to the formation of public policy. In the area of health care research, this is already happening. In 1989, an act of Congress created the Office of the Forum for Quality and Effectiveness in Health Care within the Agency for Health Care Policy and Research. The forum was created to develop guidelines for clinical practice. A novel aspect of this effort to develop medical practice guidelines is that forum guidelines are required to be based on research evidence whenever possible (National Institute of Medicine 1990; Woolf 1991). Expert opinion or clinical judgment is substituted only when research evidence is not available to support some aspect of a guideline. The forum has already issued guidelines in a number of areas including the management of pain, depression, urinary incontinence, pressure ulcers, and cataracts, and other panels are currently developing guidelines on other issues. It is important to recognize that these clinical practice guidelines are practice policies and thus their development is an act of policymaking (Eddy 1990*a*, 1990*b*, 1990*c*; Woolf 1991, 1992.)

It is a matter of some concern, then, whether systematic syntheses of research can provide reliable evidence from which to gauge the likely effects of policies that might be adopted. The record of meta-analysis in providing valid syntheses of scientific research for purely scientific purposes is unassailable both from an analytic (deductive) standpoint and from an empirical standpoint. In medicine, meta-analytic conclusions have been repeatedly validated by larger clinical trials (see Chalmers et al. 1987) and in the physical sciences by subsequent experiments of higher accuracy (Rosenfeld 1975). However, the record of meta-analysis is not nearly so compelling in the arena of providing reliable evidence for policy purposes. Two examples illustrate the point.

In the health care field, the General Accounting Office’s cross-design synthesis project (Silberman et al. 1992) explored a notable lack of correspondence between estimates of the effectiveness of experimental treatments derived from clinical trials and data derived from population surveys after those experimental treatments became the standard of practice. The clinical trials found that the experimental treatments could drastically reduce death rates among those treated for a particular disease. Consequently, one would expect to see the death rates from the disease drop as the new treatments became standard. However, the population survey data failed to validate the clinical trials estimates of the likely treatment effect when implemented as a practice policy.

A second example comes from educational research, particularly from syntheses of research on classroom learning. A series of such syntheses produced singularly unconvincing recommendations for policy, even though the research foundation is rather sound (Wang et al. 1993). Celebrated examples from this tradition include mastery learning methods; their efficacy and practicality are supported by an enviable body of research, but the practical applications have been disappointing.

The purpose of this chapter is not to provide a comprehensive review of previous work in meta-analysis, but to question its applicability for the purposes of drawing inferences for policy. It is argued that conventional approaches to meta-analysis are ill suited to inform many policy questions—not because they are technically flawed, but because they answer the wrong questions. Thus the failure is one of articulating the problem precisely and insuring that the methods are well suited to address the problem. Because these tasks (particularly stating the problem in a way that is useful for ensuring relevant statistical analysis) are very difficult, some researchers may have fallen into a trap that Tukey (1994) identified as a perennial problem in applied statistics: having a good answer to the wrong question. To avoid this trap, Tukey suggests researchers think carefully about the question and try to get an answer (even a poor answer) to the question that they really care about.

The genesis of the problem is that scientific research literatures consist of studies (experiments) whose designs are selected according to practical and scientific criteria. The criteria used in selecting the study context and variations of the treatment to be studied may change as research on a treatment progresses. Early in a research program, intensive variants of the treatment are likely to be studied in contexts or with subjects believed to be susceptible to the treatment; such studies may continue throughout the research program,. After the treatment efficacy is established under such highly favorable conditions, scientific interest may shift to the efficacy of less intense variants of the treatment under less favorable conditions. These less favorable conditions almost always correspond to the conditions under which treatments will be applied in practice, and hence are more relevant to policy questions. This is well known in evaluation research; for an interesting example in another context, see Feinstein (1985).

This chapter proposes a model or framework for thinking about the problems of drawing inferences from research literatures for policy purposes and suggests how this model may be used in research syntheses. It is argued that use of the model will reveal the nature of the research evidence available, identify knowledge gaps when evidence is unavailable, and better summarize the available evidence for policy decisions. By estimating components of variability, this model will also help quantify the likely generalizability of research findings.

Related Conceptualizations

The model proposed is in the same spirit as other models for inference from collections of studies. Cordray's policy space (Cordray and Fischer 1993) incorporates the idea of classifying studies according to treatment type (intensity) and context (subject type). Rubin's response surface model (Rubin 1990, 1992) incorporates the idea of classifying studies by study design and treatment type. Cronbach's model of construct generalization (Cronbach 1982) incorporates the idea that the relevant population (universe) about which generalizations are desired is multifaceted, including facets for context. Becker's (in press) model of generalizability of study results extends Cronbach's formulation to research synthesis.

The approach of this chapter is informal and generally nontechnical (although some of the content is by its nature technical), but the statements are precise. For example, the term "uncertainty" is used to refer to the variance of certain quantities without fully qualifying the random variables involved or whether it is to be taken as a subjective distribution or a classical sampling distribution (in most cases it could be made precise as either).

INFERENCE MODELS IN META-ANALYSIS

It is convenient to summarize inference models in meta-analysis within three categories: fixed-, random-, and mixed-effects models. These distinctions have been made in other contexts and have been applied before to meta-analysis (Cooper and Hedges 1994; Hedges 1992*;* Hedges and Olkin 1985). To understand these models, assume that researchers are interested in summarizing a collection of independent research studies, each of which can be described by a numerical index (such as a proportion, a correlation coefficient, a mean difference, or a rate ratio). In research synthesis, such indexes are generically known as indexes of effect size because they provide a quantification of the degree of relationship between variables. In any particular meta-analysis, it is usually desirable to work with the same type of index of effect size from all studies.

All three inference models distinguish the concepts of a population effect size or effect size parameter from that of a sample effect size or effect size estimate. When necessary, the effect size estimates from k independent studies are denoted by Roman letters subscripted by the study identification number and the corresponding effect size parameters by Greek letters. Thus T_1...T_k might be the effect size estimates from k studies, ${}_1$,..., ${}_k$ are the corresponding effect size parameters, and T_i differs from ${}_i$ by an amount G_i = T_1 - ${}_i$, which is usually referred to as a "sampling error." Except for biases that arise in some estimation conditions, sampling errors are due to variations across the samples of individuals that might be used to compute effect size estimates. Sampling errors arise because researchers estimate effect size in any individual study from a sample of finite (often quite small) size. If a study had a sample of infinite size available, there would be no sampling error.

The uncertainty of T_1 as an estimate of ${}_i$ is usually quantified by the standard error (the square root of the sampling error variance), which is denoted &$_i$. Indexes of effect size used in meta-analysis have a property that permits the sampling error variance to be analytically derived as a function of the effect size itself and the sample sizes; consequently, sampling error variances can be treated as "known" quantities and not as quantities that have to be estimated from replications in the data.

Fixed-Effects Models

Fixed-effects models are both the simplest and the most widely used statistical models in meta-analysis. They treat the effect size parameters as if they were fixed quantities. The parameters may differ across studies, but such differences are not thought of as a consequence of chance processes. The simplest fixed-effect model, and the model most often used in meta-analysis, treats all studies as having the same effect size parameter = ${}_1$ = . . . = ${}_k$. More complex fixed-effects models posit that the effect size parameters ${}_1$, . . ., ${}_k$ are a simple (usually linear) function of study characteristics. For example, the effect size might be taken as a function of duration or intensity of treatment, and fixed-effects models might be used to test whether studies with short duration or low intensity have smaller effect sizes than studies of long duration or high intensity.

Note that fixed-effects models make rather strong assumptions about the data. One is that between-study variations in effect size parameters are not the consequence of random processes, and thus do not add to uncertainty of summaries such as the average effect. However, various tests of model specification have been developed to determine if sample effect sizes are consistent with fixed-effects models (Hedges and Olkin 1985), and there is a considerable body of evidence that these models are often reasonably consistent with meta-analytic data.

Random-Effects Models

Random-effects models differ from fixed-effects models in that they treat the effect size parameters ${}_1$, . . ., ${}_k$ as if they were sampled from a universe (hyperpopulation) of possible effect size parameters. The conceptual model usually considers the observed studies as a (random) sample from a universe of studies that might have been observed. Since the studies are selected at random, their effect size parameters are a sample from a universe of effect size parameters. The object of the analysis is to estimate the (hyper-) parameters that describe this (hyper-) population of effect size parameters, usually the mean and the variance (which is often called the between-studies variance component).

Although random-effects analysis is superficially similar to fixed-effects analysis, yielding for example an estimate of the mean effect size and its uncertainty (in the form of a standard error), the meaning of these quantities is subtly different. The estimated mean in the random-effects model is the mean of a population of effect size parameters, and it is possible for the average effect size parameter to be positive but for some (perhaps a large proportion) of the effect size parameters to be negative. The characteristics of the distribution of the random effects (in particular, its variance) help determine how likely this is to occur.

Mixed-Effects Models

Mixed-effects models incorporate some of the characteristics of both fixed- and random-effects models. In mixed-effects models, the effect size parameters are partly determined by knowable characteristics of the studies (fixed effects) and partly the result of random processes. The models are typically employed by using a specified set of study characteristics as fixed-effects predictors of the effect size parameters, and defining any remaining between-study variation as random. One can think of this as defining a universe of studies that have precisely the same set of characteristics (the same values of the fixed effects) and treating the observed studies with that set of characteristics as a sample from that universe.

PROBLEMS WITH APPLICATIONS OF META-ANALYTIC MODELS FOR POLICY

Any of the meta-analytic models are quite capable of providing valid answers to the questions they are designed to answer. Unfortunately for policy purposes, they are usually used to address the wrong question. The models are frequently used to summarize studies that have been done; a simple summary of studies is rarely the answer to a question of real interest to policymakers.

PROBLEMS WITH THE RANDOM-EFFECTS CONCEPTUALIZATION FOR DRAWING POLICY RELEVANT INFERENCES

Random-effects models (as conventionally used) make the assumption that the sample of studies is a simple random (or at least, representative) sample of the universe of studies to which generalization is desired. This is an astonishingly naive assumption. Even if it happened to be true in any one case, the fact that multiple perspectives on the same issue would often prescribe different universes makes it impossible for the sample of studies actually conducted to represent all the policy relevant universes to which generalization is desired.

An Illustrative Example. Consider a simplified example of estimating the rate of drug use (or the effect of a drug use prevention program). Suppose that the rate (or effect) depends on the age and ethnicity of the target population, and assume for simplicity that there are two age groups (young and old) and two ethnic groups (African American and European American). Now consider a collection of equally valid research studies, each of which provides data on one of the age or ethnicity categories. How should one combine the information across studies to make an inference about the rate of drug use (or the effect of the prevention program)? It depends on the precise question one wants to answer.

In determining the rate for the entire population, all studies are relevant. But if one wants to know about the rate among young people, only some studies are directly relevant. Moreover, the average rate (or the average effect of the treatment) is highly unlikely to be the relevant summary. For example, if one is interested in young people in general and there are equal numbers of studies of African Americans and European Americans, then (since there are more European Americans than African Americans in the general population) the simple average will overweight the results of studies of African Americans.

Not only is the simple average unlikely to be the relevant summary, but its uncertainty is unlikely to be the relevant estimate of uncertainty for two reasons. First, the variance of the combined estimate depends on the variance of the estimates that go into it. If the uncertainty of the estimated effects within ethnic groups is not the same, misweighting the groups in the combined estimate also misweights data for the purposes of computing uncertainty. Second, even if the within-group uncertainty is the same for each race, misweighting the ethnic groups will lead to misweighting the between-group component of the uncertainty of the combined estimate.

A CONCEPTUAL FRAMEWORK FOR POLICY RELEVANT INFERENCE

The first step in a policy relevant synthesis is to classify the relevant variables that have a systematic effect on study results. It is reasonable to assume that effect parameters depend on three general categories of study characteristics: treatment type, study design, and study context. Treatment characteristics include all of the ways that the nominal treatment may vary systematically across studies including the duration, intensity, and mode of treatment administration. When the treatment itself is diffusely defined, the particular variety of treatment is a relevant variable here. Note that unplanned variations in treatment implementation would not be included as variables here because they are not controlled and add to unsystematic variation.

Study design characteristics include all of the systematic aspects of the research design and procedure except those that are part of study context. These characteristics include procedures used to ensure internal validity (the conventional meaning of experimental design) as well as characteris-tics of the outcome measures used.

Study context characteristics include aspects of the target populations and the settings in which the research study was conducted: all of the usual demographic characteristics of the subject population, and the character-istics of treatment settings as well (e.g., a school-based, community-based, or individualized program). Obviously there will be some ambiguity among these categories of variables, and some treatment types can occur only in some contexts, but in any given policy question, decisions (albeit arbitrary ones) can be made to classify a variable in that way for the purposes of the analysis.

Treatment Type and Context Define the Estimand

The technical development of valid statistical inference depends on unambiguous statement of the quantity to be estimated. Researchers often forget this point when working in areas where the statistical procedures and underlying conceptual models are so well understood as to be conventional. However, problems are often complicated by conceptual ambiguity about the quantity to be estimated. The purpose of this lengthy theoretical development is to provide a framework for achieving clarity on what should be estimated to help meta-analytic summaries better inform policymaking.

In order to define the inference problem precisely, it is necessary to define the treatment type and the study context variables. To be clear about what treatment effect to estimate, researchers must know which treatment variations to count as implemented, with which subject population, in which settings.

The treatment type variables serve to define the treatment itself. Researchers may wish to draw inferences about the likely effect of any particular subtypes of treatment or about a mix of them. If the treatment mix is of interest, it is important to recognize that changing the propor-tions of each subtype in the mix may change both the estimate of the overall effect and its uncertainty.

A more technical way to put the above argument is that, to specify the population to which researchers wish to generalize, the distribution of contexts and the treatment types must be specified. The most convenient way to do that is to define a context stratification system and specify the population weights given to each cross-classified context stratum.

Study design characteristics reflect the standard of evidence of internal validity to be applied in drawing inferences. In principle, these characteristics could be considered technical parameters. The policy-maker is unlikely to be interested in the relations between these characteristics and effect size, in and of themselves, although scientists studying quasi-experimental design would find them substantively interesting. The policymaker wants to know what the effect is, not what design features lead to biases (unless this information helps interpret evidence). An optimist might consider the study design characteristics as a way to categorize the departures of existing studies from hypothetically perfect studies (i.e., studies that provide an unbiased estimate of a conceptual treatment effect). In fact, Rubin (1990, 1992) has suggested that researchers estimate a "response surface," precisely characterizing the relation between study design characteristics and effect size in order to estimate the effect size of such an ideal study.

Estimation and Inference

After an estimand has been precisely defined by specifying the relevant distribution of contexts and treatment types, the problem becomes one of estimating the mean and uncertainty (variance) of the treatment effects as a well-defined statistical problem. It is most natural to carry out the estimation in the context of a random- or mixed-effects model, although the analysis requires some modification of existing methods to accommodate the weighting used to define the relevant distribution of contexts. This would involve a reasonably straightforward adaptation of methods that are already well developed in the analysis of stratified samples from survey designs (Cochran 1977). In principle, these methods would involve stratifying the sample of studies and then carrying out a meta-analysis within each of the strata using standard methods. These summary statistics from the meta-analyses would then be combined using a weighted combination procedure similar to those used in the analysis of stratified sample surveys. In practice, a few difficulties will arise.

Methods for Weighted Combination of Meta-Analyses Do Not Exist

Even with a modest set of context and treatment type strata, some (and perhaps many) of the strata will have no studies. That is, the stratified sample will have missing data. Note that this is a limitation of the data, not a limitation of the synthesis method. This limitation is a strength of the method—it forces researchers to confront the fact that the available data are not adequate to provide an empirically based estimate of the relevant treatment effects.

When faced with missing data, there are three choices: get more data, substitute assumptions for data, or change the question. The first option is typically the best, but least immediately feasible. However, it is important to note that the identification of missing data in a synthesis is equivalent to identifying studies that need to be done and whose results would reduce uncertainty in policy relevant inferences.

One practice in dealing with missing data is imputation of missing data (or more sophisticated model-based inference under models that include missing data in their specification). In this case, the assumptions that substitute for the data are embodied in the imputation (or missing data) model (Little and Rubin 1987; Rubin 1987). Here the assumptions concern the relation between the observed data and the missing data, so that empirical evidence plays some role in values substituted for the missing data.

A different way of adjusting for missing data is to go entirely outside the data set and use expert opinion. Estimates derived via expert opinion could be used in place of empirical research results in strata where data are missing. There are many methods of gathering such information, including a considerable literature on how to elicit prior information for Bayesian statistical analyses (Kadane et al. 1980; Winkler 1967). One particular advantage of the sampling frame for contexts is that it narrows the domain about which expert opinion is elicited. Expertise, by definition, is a consequence of substantial experience and is necessarily context bound. The use of relatively narrow contexts helps make it possible to ensure that the experts provide information within their domains of expertise. Within those domains, it is quite likely that expert opinion can be satisfactorily substituted for empirical research results. Indeed, the adequacy of expert opinion could be monitored by eliciting expert opinion in domains where satisfactory empirical evidence already exists.

Knowledge of Population Weights

In order to specify the population of interest, it is necessary to specify the population weights for each stratum. It is probably possible to specify strata for which these weights would be unknown. However, it seems obvious that these weights are of critical interest; knowing the composition of the target population and settings to affect is critical to formulating wise policy. Perhaps one should be wary of any tools that purport to yield targeted evidence on policy that do not also require knowledge of who is to be affected and in what contexts.

It may not be critical that the weights be known exactly. If effects do not vary profoundly across adjacent strata, then modest variations in the weights will produce only small variations in the overall effect. (If effects do vary profoundly across strata, one should be cautious about averages because they may obscure real variation.) Examining alternate possible values of the weights will permit bracketing of effects and sensitivity analysis. In fact, uncertainty in the weights could (and probably should) be incorporated into the overall estimates and their uncertainty.

Ambiguity in Classification of Studies Into Strata

It is clear that some studies will be difficult to classify into strata. Some may overlap stratum boundaries. In other cases parts of a study may fall into different strata. Such problems are common in meta-analyses and there is little reason to believe that they would be insurmountable in this context.

CONCLUSION

The model of synthesis proposed here defines questions more sharply in a fashion more relevant to policy concerns—what might happen if a policy were implemented in a relevant range of contexts. It is a more difficult approach, but one that is not impossible to carry out. The model will reveal gaps in evidence and make explicit precisely how assumptions have been substituted for empirical evidence to make inferences when some of the necessary empirical evidence was unavailable. This new model could produce estimates of treatment effects that are similar to those produced by more traditional meta-analytic methods. For example, if all studies gave the same estimate of treatment effect regardless of context or treatment type, the overall estimates from a simple meta-analysis and the more complex variety described here would coincide. Most likely they would not. In that case, the model proposed here provides more valid answers to questions of interest to policymakers.

REFERENCES

Becker, B.J. The generalizability of empirical research results. In: Benbow, D., and Lubinski, D., eds. *From Psychometrics to Giftedness: Papers in Honor of Julian C. Stanley*. Baltimore: Johns Hopkins University Press, in press.

Chalmers, T.C.; Levin, H.; Sacks, H.S.; Reitman, D.; Berrier, J.; and Nagalingam, R. Meta-analysis of clinical trials as a scientific discipline, I: Control of bias and comparison with large co-operative trials. *Stat Med* 6:315-315, 1987.

Cochran, W.G. *Sampling Techniques*. 3d ed. New York: Wiley, 1977.

Cooper, H.M. *The Integrative Research Review*. Beverly Hills, CA: Sage Publications, 1984.

Cooper, H.M., and Hedges, L.V., eds. *The Handbook of Research Synthesis*. New York: The Russell Sage Foundation, 1994.

Cordray, D.S., and Fischer, R.L. Practical aspects of evaluation synthesis and its variations. In: Wholey, J.S.; Harty, H.P.; and Newcomer, K.E., eds. *Handbook of Practical Program Evaluation*. San Francisco: Jossey-Bass, 1993.

Cronbach, L.J. *Designing Evaluations of Educational and Social Programs*. San Francisco: Jossey-Bass, 1982.

Eddy, D.M. Practice policies—What are they? *JAMA* 263:877-880, 1990*a*.

Eddy, D.M. Practice policies: Where do they come from? *JAMA* 263:1265-1275, 1990*b*.

Eddy, D.M. Practice policies—Guidelines for methods. *JAMA* 263:1839-1841, 1990*c*.

Feinstein, A. *Clinical Epidemiology*. Philadelphia: W.B. Saunders, 1985.

Hedges, L.V. Meta-analysis. *J Educ Stat* 17:279-296, 1992.

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

Hunter, S.E., and Schmidt, F.L. *Methods of Meta-analysis: Correcting Error and Bias in Research Findings*. Beverly Hills, CA: Sage Publications, 1990.

Kadane, J.B.; Dickey, J.M.; Winkler, R.L.; Smith, W.S.; and Peters, S.C. Interactive elicitation of opinion for a normal linear model. *J Am Stat Assoc* 75:845-854, 1980.

Light, R.J., and Pillemer, D.B. *Summing Up: The Science of Reviewing Research*. Cambridge, MA: Harvard University Press, 1984.

Little, R.J.A., and Rubin, D.B. *Statistical Analysis of Missing Data*. New York: Wiley, 1987.

National Institute of Medicine. *Clinical Practice Guidelines.* Washington, DC: National Academy Press, 1990.

National Research Council. *Combining Information: Statistical Issues and Research Opportunities*. Washington, DC: National Academy Press, 1992. [Reprinted as Draper, D.; Gaver, D; Goel, P.; Greenhouse, J.; Hedges, L.; Morris, C.; and Waternoux, C. *Combining Information: Statistical Issues and Research Opportunities*. Washington, DC: American Statistical Association, 1993.]

Rosenfeld, A. The particle data group: Its nature and operation. *Ann Rev Nuclear Sci* 555-599, 1975.

Rosenthal, R. *Meta-analytic Procedures for Social Research*. Beverly Hills, CA: Sage Publications, 1984.

Rubin, D.B. *Multiple Imputation for Missing Data in Sample Surveys.* New York: Wiley, 1987.

Rubin, D.B. A new perspective on meta-analysis. In: Wachter, K.W., and Straf, M.L., eds. *The Future of Meta-Analysis.* New York: The Russell Sage Foundation, 1990.

Rubin, D.B. Meta-analysis: Literature synthesis or effect-size surface estimation? *J Educ Stat* 17:363-374, 1992.

Silberman, G.; Droitcour, J.A.; and Scullin, E.W. *Cross Design Synthesis: A New Strategy for Medical Effectiveness Research.* Report No. GAO/PEMD-92-18. Washington, DC: U.S. General Accounting Office, 1992.

Tukey, J.W. Methodology and the statistician’s responsibility for BOTH accuracy AND relevance. *J Am Stat Assoc* 74:786-793, 1994.

Wang, M.C.; Haertel, G.D.; and Walberg, H.J. Toward a knowledge base for school learning. *Rev Educ Res* 63:249-299, 1993.

Winkler, R.L. The assessment of prior distributions in Bayesian analysis. *J Am Stat Assoc* 62:1105-1120, 1967.

Woolf, S.H. *AHCPR Interim Manual for Clinical Practice Development*. AHCPR Pub. No. 91-0018. Washington, DC: Agency for Health Care Policy, 1991.

Woolf, S.H. Practice guidelines, a new reality in medicine II: Methods of developing guidelines. *Arch Intern Med* 152:946-952, 1992.

AUTHOR

Larry V. Hedges, Ph.D.
Professor
The University of Chicago
5835 South Kimbark Avenue
Chicago, IL 60637

Using Linked Meta-Analysis To Build Policy Models

Mark W. Lipsey

There are many readily identifiable applications of meta-analysis to the area of drug abuse prevention and related topics. Meta-analysis of preventive intervention research (e.g., Tobler 1986), for instance, can identify more and less effective approaches, as can an analogous meta-analysis of rehabilitative treatment research. Meta-analysis of the correlates of intervention-induced change can illuminate the psychological processes involved in the response to intervention. Meta-analysis of the predictive relationships of risk variables with subsequent abuse can indicate which types of variables are most strongly related to the target behavior and chart the developmental course of drug abuse problems (cf. Loeber and Dishion (1987) on antisocial behavior). Meta-analyses of the relationships among risk factors might better identify their structure and the independent clusters they represent. Meta-analysis of the consequences and correlates of drug abuse can trace the patterns of dysfunction in which abuse is embedded. While each of these individual applications may have considerable merit, the intent here is to look ahead to the prospects of linking a number of such meta-analyses into an integrated whole that covers multiple aspects of problem behavior in a coordinated manner.

It is the purpose of this chapter to sketch a meta-analytic approach to building policy models for certain difficult social problem areas such as drug abuse. The term "policy model" means an interconnected set of statements of relationships that embrace the key variables in the problem (especially those manipulable by social programs or policy), that are descriptively accurate regarding the nature and extent of the problem, that incorporate both predictive/diagnostic risk factors for the problem and the effects of intervention in the problem, and that reflect change over time. Most important, such a model must permit "what if" simulations that yield valid insights into the results of changed risk circumstances, different interventions, and the like.

Meta-analysis offers the potential to integrate the full range of empirical information about a problem into a policy model that may then provide an efficient information base from which to address a number of practical questions in a coordinated manner. To effectively prevent drug abuse, for instance, one must know what risk factors are predictive of subsequent abuse and what interventions may alter those risks. To treat abuse, one must know what range of problems associated with the abuse must be targeted, what treatments are most effective, and how long lasting the effects are. To scale the prevention and treatment effort to the nature of the problem, one must know how widespread abuse and risk for abuse are, how they are distributed in the population, and what trends can be expected in the levels of problem behavior.

There are various identifiable examples of such policy models. In social welfare policy, for instance, rather sophisticated computer simulations have been used to apply different stipulations of government regulations to demographic databases and projections in order to investigate the costs and scope of the different policies (Citro and Hanushek 1991). On other fronts, economists routinely use various forms of economic theory to develop models to explore policy options on a wide range of topics (e.g., markets, labor, housing).

A particularly difficult area for such modeling, however, is presented by social problems that involve a substantial behavioral component and are heavily influenced by personal choices, experiences, and characteristics. Such problems include substance abuse, chronic criminality, domestic violence, school dropout, persistent unemployment, homelessness, and the like. Policy models for these kinds of situations are difficult to develop because the problems are not functions of simple demographics, nor do they lend themselves to analysis in terms of broad economic tenets, incentives, response to law, or other principles of rational behavior. In these areas there are no comprehensive policy models but, rather, various piecemeal models based on the empirical findings of one study or another. Most of these efforts are too limited in scope and have too narrow an empirical base to provide much utility for policy. It is in these difficult problem areas especially that meta-analysis can be used as a tool for integrating empirical findings and contribute to the development of useful policy models. This approach can be illustrated by work underway on antisocial (criminal and delinquent) behavior that is generally applicable to the problem of drug abuse as well.

A DEVELOPMENTAL FRAMEWORK

The focus of this example is on those problem behaviors that can be effectively represented as developmental progressions. This perspective recognizes that there is often a period prior to display of a problem behavior by an individual that may be characterized by the presence of risk factors predictive of the behavior, as well as a period afterwards when the behavior may either go into remission or be established in a persistent, chronic pattern. The early phases of this developmental progression are the appropriate points for any preventive intervention. The later phases are the appropriate points for direct rehabilitative treatment of the problem or, perhaps, supportive treatment to prevent backsliding after the problem is in remission. This framework is most applicable to chronic problems that have distinct precursors in childhood and adolescence. General antisocial behavior can be represented in these terms, as can drug abuse.

To depict this developmental progression in terms of relationships that may be important to a policy model, one must distinguish a variety of elements that can be associated with each other developmentally or concurrently, as described below.

1. *Behavioral progression*. Few problem behaviors represent sharp discontinuities from prior behavior. Typically there are precursor behaviors that share many of the underlying characteristics of the problem behavior. For instance, hyperactivity in early childhood, aggressive behavior in childhood, and criminal violence in adolescence and adulthood are probabilistically linked in a behavioral progression (Loeber 1988). Similarly, abuse of cocaine or heroin is generally preceded by the use of other drugs (Collins 1991). These behavioral progressions have been described as instances of "heterotypic continuity" (Sampson and Laub 1992) to indicate the underlying psychological continuity in what on the surface are different behaviors.

2. *Ancillary problem behaviors*. Serious problem behaviors are rarely manifest in isolation. The problem behaviors themselves cause other problems, as when a person loses employment because of substance abuse. Also, factors that lead to a given problem behavior produce other problem behaviors, as when a person with poor impulse control has problems with delinquency, substance abuse, and personal relationships.

3. *Risk and protective factors*. There is a wide range of variables other than overt precursor or ancillary problem behaviors that are predictive of subsequent problem behaviors. Personal characteristics (e.g., temperament, intelligence), family circumstances (e.g., broken home), nature of peers, socioeconomic status, and many other such factors measured at time 1 can be predictive of problem behavior at time 2. Those that are associated with the emergence of problem behavior are risk variables; those that are associated with less problem behavior than expected at a given risk level are protective variables (Hawkins et al. 1992).

4. *Intervention*. Programs or policies of intervention into the problem behavior cycle can attack the problem behavior itself, ancillary problem behaviors, risk factors, or the social/environmental factors that produce risk. Moreover, they may be preventive interventions that are targeted at the early phases of the developmental progression, rehabilitative interventions during the period when the problem behavior is overt, or maintenance interventions aimed at stabilizing recovery or preventing relapse.

Figure 1 depicts a generic developmental progression in which arbitrary stages of development of the focal problem behavior (e.g., substance abuse, violence) are identified as B_1, B_2, and so forth. The progression of ancillary behavior problems associated with the focal problem is labeled A_1, A_2, and so forth. The risk factors at each stage are identified as R_1, R_2, R_3, and so forth; the potential interventions at each stage are labeled I_1, I_2, and so forth.

Figure 1 represents a generic sketch of a policy model for problem behaviors characterized by a developmental progression. If one had information about the nature and magnitude of all the relationships depicted in that figure, one would have a tool with which to support decisionmaking about appropriate social responses to the problem. For instance, this model and data about the distribution of various early risk factors would be a basis for projecting the extent to which the problem behavior will subsequently develop among a population of interest. Moreover, one could estimate how much the problem behavior might change if the risk factors were to change at different stages, whether naturally or as a result of policy initiatives. Especially important, of course, is the information that this model might provide about the effects of intervention at any stage, and in particular how it might affect the

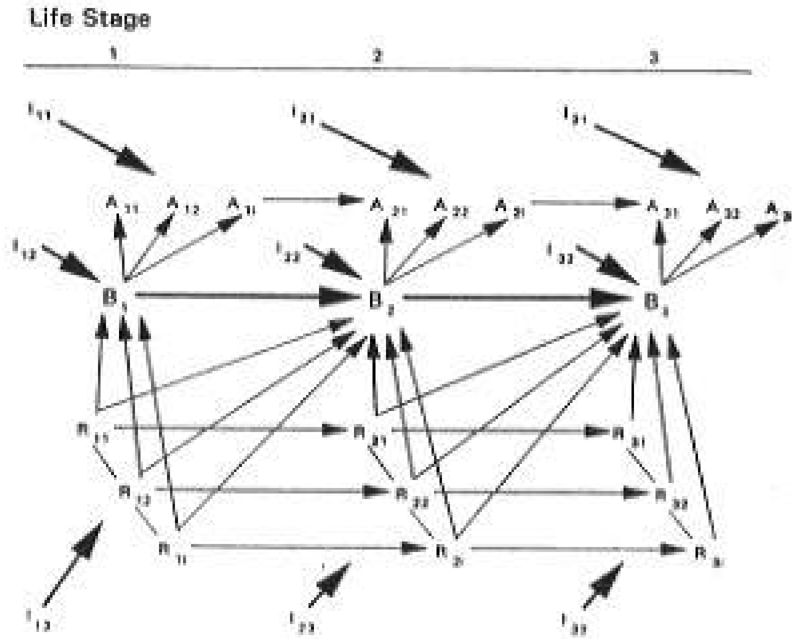


FIGURE 1. *Generic sketch of a policy model for problem behaviors characterized by a developmental progression.*

KEY: I = intervention; B = problem behavior; A = ancillary behavior; R = risk protective factor.

progression of the problem behavior directly and, indirectly, the ancillary problem behaviors.

Is Such a Policy Model Feasible?

It is apparent that, even for a rather simple behavioral problem, a model of the sort shown in figure 1 would be very complex. There are potentially a large number of variables that are relevant, and the information needed is very nearly the relationship of every variable with every other variable at each developmental stage and across all stages. No doubt this complexity is the reason why researchers do not have anything resembling this sort of policy model in the social sciences for many of the troubling social-behavioral problems being studied. Nonetheless, researchers must aspire to as complete an understanding as possible along lines such as these to effectively address the question of how best to ameliorate those problems. One can perhaps draw inspiration from the physical sciences, where it is now not uncommon to model such complex systems as weather patterns.

It is also quite apparent that there are likely to be many more relationships researchers would like to understand than any single study can investigate. Moreover, even if some study were to heroically cover the entire domain of interest, it would have inherent constraints that would make it inappropriate as the sole basis for a policy model. For instance, no matter what samples of persons and sites were involved, there would be some uncertainty about generality to other persons and sites. In addition, the research procedures, construct operationalizations, and even data analysis would represent only limited selections from the set of reasonable approaches. Ideally, one would want to base a policy model on a sufficient sampling of research to ensure some robustness or generality across the methodological and procedural options researchers might exercise and, especially, across the persons, sites, and situations that constitute the domain of the social problem under study.

It follows that the construction of policy models is best approached as a task of research synthesis. Not only does synthesis make use of all the available and relevant research, but it has inherent generality as a function of its integration of multiple studies with all their diversity of methods, samples, and situations (Cook 1993).

Obviously, a synthesis of research bearing on the relationships of interest for a particular policy model is itself limited by the availability of relevant research. There is little likelihood that sufficient research exists in any domain of problem behavior to permit solid meta-analytic estimates of the nature and magnitude of every relationship of interest. For many problem areas, however, there is a corpus of research more than sufficient to permit a start on model development. In addition, one advantage of systematic meta-analysis is that it yields very specific identification of variables and relationships that have not been adequately covered in research and warrant more attention. Development of policy models, therefore, will inevitably be an iterative process in which the quality of the meta-analyses supporting the models will improve as gaps in the primary research are identified and attended to by the research community.

This chapter now takes a closer look at how meta-analysis might be employed to begin the process of constructing useful policy models for social problems reflecting progressions of problem behavior.

Meta-analysis revolves around the effect size, a statistical index of the magnitude of a relationship. The most fully developed procedures are for the product-moment correlation as an effect size index of the degree of association between two variables, and for the standardized difference between means as an effect size index of group differences, whether natural or experimentally induced (Durlak and Lipsey 1991). However, effect sizes in one of these metrics can be algebraically transformed to the other. For present purposes, think of the relationships depicted in the scheme of figure 1 as entirely correlational. This begs the important question of the extent to which certain key relations among them are causal and hence have predictable results when the independent variable is manipulated. This chapter will return to that issue later.

In analyzing the relationships pertinent to the scheme shown in figure 1 for the categorically different types of relationships that must be synthesized in order to give a full accounting of the developmental progression, one finds the following (not all shown in figure 1 to limit clutter).

1. Predictive relationships between a variable measured at time 1 and a variable measured at time 2, representing different stages in the developmental progression:

   BøB relationships
   AøA and AøB relationships
   RøB and RøA relationships
   IøB, IøA, and IøR relationships

2. Cross-sectional relationships between two variables measured at the same time (i.e., during the same stage in the developmental progression):

   AöA relationships
   AöB relationships
   RöR relationships
   RöB and RöA relationships

All of these types of relationships are typically studied and reported in research bearing on the problem behaviors of interest. Longitudinal and panel studies of the problem behavior and, sometimes, of general human development provide information on relevant time 1—time 2 predictive relationships. Cross-sectional surveys and other such studies provide information on the concurrent relationships. Experimental and quasi-experimental investigations provide information on the relationship between an intervention and subsequent outcome variables.

Rarely would all of these types of relationships be investigated in a single study, however. Indeed, experimental studies of intervention, cross-sectional surveys, and longitudinal studies are, for the most part, cate-gorically different research paradigms that study certain subsets of these relationships and almost never examine the other subsets. Meta-analysts typically, and quite reasonably, restrict themselves to synthesizing research in one of these domains (e.g., intervention studies), where comparable issues are investigated with comparable methods across studies.

Constructing a policy model that involves all of the types of relationships shown in figure 1 with meta-analytic techniques, therefore, requires information from multiple meta-analyses—those synthesizing intervention studies, those synthesizing cross-sectional studies, and those synthesizing developmental relationships. Moreover, the natural boundaries of the respective research literatures in these paradigms are likely to be differentiated according to developmental stage. A meta-analyst might, for instance, synthesize intervention research for programs aimed at preventing drug abuse before it begins, but would not necessarily include programs aimed at treating abuse after it is established. Complete coverage of the relationships shown in figure 1, thus, requires something more like a family of meta-analyses than a single one.

Given the natural distinctiveness of the different research paradigms and issues studied within them, and the corresponding distinctiveness of the meta-analyses that would synthesize research within each of those categories, it seems apparent that it will require a set of linked meta-analyses to cover all the relationships relevant to even a simple policy model. But if these research paradigms are distinct, how can the various different research literatures and corresponding meta-analyses be linked into such a model in an integrated manner?

The answer is that such linkage is not possible unless there is substantial overlap among the various research categories in the variables studied. Fortunately, such overlap is relatively common. Intervention studies target as outcome measures much the same problem behavior and risk variables that are of interest to longitudinal and survey researchers. Longitudinal and survey researchers, in turn, often study much the same variables despite their different methods. Moreover, the variables that are of interest at one developmental stage generally overlap those that are of interest at a later stage. By organizing relationships around the key variables of the model, therefore, it should be possible to link information from different literatures and different meta-analytic domains.

The central concept here is the notion of linked meta-analyses—integrating meta-analyses of different but related research literatures via overlapping variables to cover all the relationships needed to synthesize an overall policy model. This policy model will consist of a complex, integrated set of synthesized empirical relationships covering interconnections among the stages of the developmental progression for the problem behaviors, predictive risk factors, and protective factors across those stages, and the effects of intervention at different stages.

Clearly what is envisioned here is a rather complex undertaking, though it builds directly upon existing method and experience in meta-analysis. The remainder of this chapter briefly discusses what seem to be the most important issues that must be resolved in order to proceed along these lines.

CHALLENGING ISSUES

Aside from the sheer complexity of identifying, acquiring, and meta-analyzing all the empirical research relevant to one or another relationship in a policy model of the sort described here, there are some special challenges such an endeavor poses that go beyond current experience and techniques in meta-analysis. Some of the most salient are itemized below.

Punctuating Developmental Stages

Using a developmental framework is central to the version of a policy model proposed here. Organizing information in terms of a developmental progression makes it possible to examine the potential effects of preventive intervention and also gives a basis for projecting likely trends and future problem levels as a function of the frequency and distribution of predictive risk factors. The research base that contributes the most to estimating the nature and magnitude of relationships involved in such a developmental progression consists of studies that investigate the associations between risk or precursor variables at time 1 and problem behavior or subsequent risk at time 2. However, the time 1—time 2 intervals represented in longitudinal research of this sort are likely to vary widely from one study to another. This poses a problem for the meta- analyst of how to organize and aggregate effect sizes representing different intervals covering different portions of the presumed developmental progression.

The most straightforward solution is to divide the developmental progression into different stages indexed to characteristics of the persons moving through those stages. The simplest such characteristic is age for those problems that have childhood precursors and tend to stabilize in chronic form for adults. For aggressive antisocial behavior, for instance, a developmental progression can be charted by dividing the age continuum into segments from birth to adulthood, since there are clear childhood antecedents of aggressive behavior and considerable stability thereafter. For problems like alcoholism that may cycle throughout adulthood, other developmental markers may be needed to segment useful stages (e.g., degree of social impairment).

Once meaningful segments are established, the meta-analyst can categorize any time 1—time 2 effect size according to the stages of the progression represented by times 1 and 2. When times 1 and 2 both fall within a single stage, the relationship can be treated as virtually cross-sectional (perhaps with some statistical adjustment for minor variations in interval length). When times 1 and 2 represent different stages, the corresponding effect size can be aggregated with all like effect sizes that link those same two stages.

Multiplicity of Variables

Nearly all meta-analysis must deal with variability in the operationalization of constructs. This variability requires the meta-analyst to apply some higher order categorization by which certain ranges of operationalizations are judged to represent the same construct while others are judged to represent different constructs. Relationships involving similar constructs under that scheme can be aggregated across studies to produce corresponding effect size estimates. Meta-analysis of relationships for policy models as described here raises this same issue but, because it is likely to involve so many more variables of such diverse sorts, the complexity of the situation is greatly increased.

One approach is to use a hierarchical scheme that first categorizes variables into very broad groups (e.g., personal characteristics, family situation, environmental factors), and then subdivides those groups into smaller, more coherent clusters. Aggregation of effect sizes can then be performed at both broader and narrower levels depending on the amount of detail judged desirable in the policy model. Inevitably there will be variable types in the research literature that are unique or sufficiently infrequent so that no aggregation is possible. Setting standards for the minimum number of effect size estimates necessary for aggregation on any one relationship will likely exclude a large number of peripheral variables and somewhat simplify the meta-analyst's task.

Different Empirical Bases for Different Relationships

Since virtually no research studies are expected to include data on all the relationships pertinent to even a simple policy model of the sort envisioned here, it follows that different relationships in the model will be estimated from different studies. Since those studies are likely to vary in terms of methods, procedures, nature of samples, and the like, a question is raised about whether the different effect size estimates for different relationships will be comparable enough to be included in the same model. Though not well developed in meta-analysis, study comparability is not a new issue; it arises in any synthesis in which effect sizes for more than one categorically different relationship are being estimated (Becker 1992; Premack and Hunter 1988).

With present techniques, there seems to be little that can be done to examine this issue of comparability other than to include as full a range as possible of descriptors of the characteristics of the studies and samples employed in the meta-analysis. Such descriptors allow a side analysis of the extent to which the effect size estimates are functions of study method, procedures, setting, sample characteristics, and other factors (Lipsey 1992). To the extent that such relationships are found, statistical adjustments can be applied to better equate the study findings to be aggregated into effect size estimates for the policy model.

Missing Data, Incomplete Linkages

Even with a focus on only those variables that are most frequently and fully represented in the empirical literature, construction of a full policy model will require synthesis of effect size values for a large number of relationships. Inevitably, the empirical literature eligible for synthesis will not provide even coverage of all those relationships. Some relationships will be widely documented and many studies will contribute to the synthesis; others will not have been examined. In order to move ahead to develop a usable policy model under these circumstances, it will be necessary to fill in the gaps via some imputation or estimation strategy. The critical question is how to go about this.

Several approaches deserve consideration. One possibility would be to estimate the magnitude of underdocumented relationships on the basis of theory, hypothesis, or expert judgment. In this approach, the intuitions of knowledgeable persons, or whatever theory was available or could be developed, would be used to assign an order of magnitude estimate to the missing relationship. Alternatively (or in combination), an empirical technique could be applied (e.g., estimating the magnitude of a relationship between two variables as the mean of the relationship of each of those variables to other "similar" variables). Rubin (1990) has proposed a scheme in which effect sizes might be arrayed along defined dimensions in ways that could permit unmeasured effect sizes to be interpolated. More sophisticated empirical imputation techniques may also be applicable, but only limited work has been done along these lines for missing meta-analysis effect sizes (Pigott 1993).

Whatever the approach applied, it seems clear that any relationships in a policy model that are not derived from directly relevant empirical estimates must be flagged as weak points in the model. Ideally, they would be updated as soon as possible with empirical estimates based on new research designed to fill in the most crucial gaps in the model.

Causality

Many of the questions one would want a policy model to address have to do with cause-and-effect relationships. The most obvious example would be assessment of the likely effects of intervention of a given sort at a given stage on developmental progression. For direct intervention effects, available experimental research can be expected to provide information interpretable in causal terms. Less direct effects (e.g., those on ancillary problem behaviors, long-term effects, and effects on subsequent risk factors not generally studied in intervention research) will not necessarily be described by the available experimental research. However, correlational research may link those variables to outcomes that are represented in the experimental research. The question is how to estimate the indirect causal influences within the constraints of the known correlations.

A similar question is implied when one attempts to use the policy model to estimate the effects of changed risk circumstances. For example, if one wanted to know how much difference a stronger family life would make in adolescent drug abuse (e.g., reduced frequency of single parent families, higher socioeconomic status of families in poverty), one would need to estimate the effects at time 2 of altered risk factors at time 1. Available literature permits synthesis of time 1—time 2 correlations between risk factors and subsequent drug abuse but, for obvious ethical and practical reasons, there is no experimental research to identify the strength of the respective causal relationships.

Therefore, while some direct causal evidence may be gleaned from synthesis of experimental research, especially where intervention issues are involved, many of the causal issues of interest will have to be addressed on the basis of correlational data. This task is much the same as that for which path analysis and structural equation modeling were developed. An important part of constructing a policy model from research synthesis, therefore, will be the estimation and testing of causal influences among variables on the basis of theory and consistency with empirically derived correlations using structural modelling techniques (Becker 1992; Premack and Hunter 1988).

Base Rates and Frequencies

Researchers are often content to learn the nature and magnitude of the relationships among the variables pertinent to an issue. For policy and decisionmaking purposes, however, it is also often necessary to have information on the number of persons involved in a social problem and the number (or proportion) likely to be affected by any ameliorative efforts. Base rate information about the number of persons affected by a problem, or evidencing risk factors for the problem, is generally available from surveys and other such descriptive research. What is needed in addition is some means of interpreting the effect sizes for key relationships that are derived from research synthesis in terms of the number of persons affected when circumstances described by that relationship change. For example, if one knows the effect size for the impact of an intervention on drug abuse and then imagines applying that intervention to all abusers, how much will the number of drug abusers decrease?

The easiest way to represent such situations is with a set of proportions that represent transition probabilities from one state to another. For example, imagine that in a given population 10 percent were drug abusers and 90 percent were not. Say that the mean effect size for an intervention is such that 60 percent of those treated stop using drugs and 40 percent continue and, further, that the effect sizes for risk factors suggest that, of the 90 percent who don't use drugs, 2 percent will begin over the period when treatment is applied to the users. With proportions like these and base rate data for the size of the population at issue, one can estimate the number of persons in each category at any stage of the sequence.

While such proportions are often available in the literature (e.g., in cross-tabulation tables), the correlational and standardized mean difference effect sizes employed in meta-analysis do not capture all the information necessary to reconstruct their values. For purposes of constructing policy models, it would be desirable to synthesize the crucial proportions, where available, in tandem with the customary effect size indices. However, there is a minor technical problem. The meta-analysis literature has not yet adequately addressed the question of synthesizing proportions and other such univariate descriptive statistics (i.e., how to construct weighted means from different estimates, test for homogeneity, determine statistical significance, and the like). Development and explication of such techniques would be useful and should not be difficult.

Costs

Any useful policy model should integrate information on the economic factors associated with the problem situation being modeled. Most important are the costs associated with the problem itself and for the various forms of intervention that ameliorate the problem. Unfortunately for this purpose, the behavioral science research that investigates the developmental progression of behavioral problems, the associated risk factors, and the effects of intervention

does not typically include cost variables. While there may be separate economic analyses available for various aspects of these problem situations, they are not necessarily configured in such a way that they can be readily integrated with the behavioral information. A significant challenge for the construction of
useful policy models of the sort described here, therefore, is the identification and effective integration of cost factors into the model

Environmental/Social Versus Personal Risk Factors

There is a strong skew in the behavioral science literature toward identifying and measuring variables at the individual (person) level. This means that much of the literature available for synthesis for a policy model expresses risk factors as personal characteristics. However, many risk factors important to a policy model are characteristics of the social conditions and environment with which the persons at risk must cope. Omitting such risk factors from the policy model biases it, on at least some factors, in a victim-blaming direction that implies that the source of the problem is located exclusively in personal deficiencies. Given that the empirical literature itself has this skew, it is not apparent how a policy model based on meta-analysis of that literature can altogether avoid the same skew.

Nonetheless, categorization of risk variables for a policy model should at least attempt to differentiate those that reflect social conditions most directly from those that are inherently more personal (e.g., temperament). For example, socioeconomic status and risk variables involving peers, family structure, and the like are amenable to intervention programs that target social conditions rather than behavior change of individuals. A full model must consider such social intervention and provide some estimate of which risk variables would likely change, and with what results, if the social conditions were changed. Giving fair representation to this dimension of the problems and interventions represented in a policy model presents an important challenge that, at present, has no ready solution.

Implementation of the Model

The basic structure of a policy model as described here is a network of relationships among variables configured so that it is possible to estimate the effects on some variables of changing others. The scope of what is proposed, however, ensures that this network will involve numerous variables and be relatively complex. The question is how one can implement this model in a fashion that will make it useful without compromising its validity (e.g., by oversimplification). Because of the primarily correlational data that provide the empirical base to the model and the causal questions that one would want to ask of it, structural equation modeling would seem to be an appropriate approach to representing the statistical relationships that comprise the policy model.

Structural equation models, however, are not especially accessible or useful for exploring options to those who do not have specialized backgrounds. A better approach might be to use structural equation modeling results and processes as the information base in a dynamic computer simulation of an expert system. Such a simulation could present the user with an interface that depicted the crucial variables, options, and outcomes in readily understandable form. "What if" simulations could then be run to explore the expected effects on problem behavior, costs, and other aspects of changing the risk and/or intervention components in the simulation. Such an implementation of a policy model could, in principle, retain the complexity and detail of the meta-analytic results and relationships derived from the empirical literature, as well as the analytical sophistication of structural equation modeling, while still presenting the problem description, policy options, and expected results in a form that would not require specialized skills to explore or understand. Some such implementation will be necessary if the policy model is to prove useful to the policy and decisionmaking community it is intended to serve.

CONCLUSION

Certainly there are many difficulties with the concept of policy models based on linked meta-analyses. Perhaps the greatest problem, however, is that, even in the best case, developing such a model will require a leap beyond established, detailed knowledge in order to fill in the gaps and make the linkages that are required for the model but inadequately investigated in the extant research literature.

Behavioral scientists are characteristically quite conservative about moving beyond the specifics documented in established research. The level of aggregation inherent to meta-analysis and the likely insufficiencies of available research for portions of a policy model make the approach described here seem ambitious and risky. (Curiously, economists are much less inhibited about these matters, which may explain why they are often more influential in policy domains.)

However, policy and program decisions will be made whether behavioral research is deemed sufficient or not. It is the premise of this chapter that, under such circumstances, decisionmakers should be offered the best available information and, moreover, that it should be systematically synthesized and integrated rather than provided piecemeal. The approach described in this chapter is an attempt to look ahead to how meta-analysis, as an advanced technique of research synthesis, can help build a representation of empirical knowledge that is robust, general, and directly applicable to a range of program and policy issues involving recalcitrant behavioral problems.

REFERENCES

Becker, B.J. Models of science achievement: Forces affecting performance in school science. In; Cook, T.D.; Cooper, H.; 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.

Citro, C.F., and Hanushek, E.A. *Improving Information for Social Policy Decisions: The Uses of Microsimulation Modeling.* Washington, DC: National Academy Press, 1991.

Collins, L.M. Measurement in longitudinal research. In: Collins, L.M., and Horn, J.L., eds. *Best Methods for the Analysis of Change: Recent Advances, Unanswered Questions, and Future Directions*. Washington, DC: American Psychological Association, 1991. pp. 137-148.

Cook, T.D. A quasi-sampling theory of the generalization of causal relationships. *New Dir Program Eval* 57:39-82, 1993.

Durlak, J.A., and Lipsey, M.W. A practitioner's guide to meta-analysis. *Am J Community Psychol* 19:291-332, 1991.

Hawkins, J.D.; Catalano, R.F.; and Miller, J.Y. Risk and protective factors for alcohol and other drug problems in adolescence and early adulthood: Implications for substance abuse prevention. *Psychol Bull* 112:64-105, 1992.

Lipsey, M.W. Juvenile delinquency treatment: A meta-analytic inquiry into the variability of effects. In: Cook, T.D.; Cooper, H.; 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.

Loeber, R. The natural histories of conduct problems, delinquency, and associated substance abuse: Evidence for developmental progressions. In: Lahey, B.B., and Kazdin, A.E., eds. *Advances in Clinical Child Psychology*. Vol. 11. New York: Plenum, 1988. pp. 73-124.

Loeber, R., and Dishion, T. Antisocial and delinquent youths: Methods for their identification. In: Burchard, J.D., and Burchard, S.N., eds. *Prevention of Delinquent Behavior*. Newbury Park, CA: Sage, 1987. pp. 75-89.

Pigott, T. Missing data and imputation. In: Cooper, H., and Hedges, L.V., eds. *The Handbook of Research Synthesis*. New York: Russell Sage Foundation, 1993.

Premack, S.L., and Hunter, J.E. Individual unionization decisions. *Psychol Bull* 103:223-234, 1988.

Rubin, D.B. A new perspective. In: Wachter, K.W., and Straf, M.L., eds. *The Future of Meta-Analysis*. New York: Russell Sage Foundation, 1990.

Sampson, R.J., and Laub, J.H. Crime and deviance in the life course. *Ann Rev Sociol* 18:63-84, 1992.

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 Issues* 16:537-567, 1986.

AUTHOR

Mark W. Lipsey, Ph.D.
Professor of Public Policy
Vanderbilt Institute of Public Policy Studies
Box 508, Peabody College
Vanderbilt University
Nashville, TN 37212

Some Limiting Factors in Meta-Analysis

Robert L. Bangert-Drowns

In first explicating the notion of quantitative literature review for the social sciences, Glass (1976) argued that knowledge is not built from any individual study, but from the integration of findings from many studies. Individual studies do not so much yield knowledge as evidence with which knowledge can be built. Knowledge is socially constructed. To overemphasize a single study's findings or integrate research only impressionistically leaves researchers knowing less than the evidence offers, insufficiently exploiting the wealth of data scattered in separate studies.

Quantitative research integration, or meta-analysis, has a history in both the physical and social sciences that precedes Glass' formulation (Bangert-Drowns 1986; Hedges 1987). Most generally, meta-analysis is a perspective rather than a method, a recognition that research findings can be interpreted probabilistically in the context of collections of studies. The meta-analytic perspective is consistent with, and perhaps newly empowers, communal and cumulative activities of science in refining method and transforming data into knowledge (Schmidt 1992).

A number of writers initially responded with skepticism or even overt hostility to this apparently new method of inquiry (e.g., Eysenck 1978). It is hard now to find critics opposed to meta-analysis in principle (Wachter 1988). However, two kinds of concerns are still expressed about meta-analysis. The first suggests that quantitative review communicates an appearance of precision and comprehension which is in fact unreal and thus misleading. The second concern is that meta-analysis is not doing what it claimed it could do: settle important theoretical and practical questions in the midst of contradictory research findings.

These concerns arise from the fact that there is plenty of room for subjectivity and imprecision in meta-analysis (Guzzo et al. 1987; L'Hommedieu et al. 1988; Wanous et al. 1989). In spite of advances in meta-analytic method that are meant to increase the precision of literature review, meta-analysis is still, in many ways, a very human enterprise. Though in principle meta-analysis offers simple means for rendering primary research more useful, meta-analysts disagree about appropriate method (Bangert-Drowns 1986), implementation of method (Carlberg et al. 1984; Slavin 1984), and interpretation of findings (Clark 1985). Implementations of meta-analyses vary in quality and must be read with the same scrutiny afforded primary research. Primary research itself presents vagaries and biases to the reviewer that surely confound precise conclusions about underlying parameters.

Meta-analysis promises to simplify complex literatures, but will be indelibly marked with the many human decisions that shaped the original data and then integrated it in new ways. Consumers of meta-analytic products therefore must carefully review meta-analytic findings. This chapter will alert readers to critical strengths and limitations of meta-analysis for policy, theory, and practice.

COMMON CRITICISMS OF META-ANALYTIC METHOD

Meta-analytic method consists of six phases: formulation of a purpose, retrieval of studies, coding of study characteristics, calculation of effect sizes, analysis of central tendency and variation in effect sizes, and interpretation and publication of findings. Meta-analysts hear many criticisms of this process, but most criticisms target specific phases of meta-analytic implementations rather than meta-analysis in principle.

Apples and Oranges

Some critics argue that meta-analysis, in its effort to be comprehensive, necessarily mixes elements that are too dissimilar to warrant integration. Meta-analysts have been said to use "overly broad categories" which in fact confuse rather than clarify important distinctions in the literature (Gallo 1978; Presby 1978).

This apples-and-oranges problem can affect both dependent and independent variables at the levels of constructs and operationalizations of constructs. Most readers would not be concerned if a meta-analyst mixed different operationalizations of the same construct, for example, finding an average attitude toward personal drug use by aggregating standardized outcome measures (effect sizes) associated with the different attitude toward drug use instruments. However, a meta-analyst could also aggregate across constructs, combining, for example, measures of knowledge, attitude, and behavior to study a more generalized construct, effect of substance abuse education. A meta-analyst can define treatment or outcome constructs and operationalizations narrowly or broadly, and critics can complain about the breadth of such definitions.

Most importantly, however, meta-analysts control the scope of the constructs and operationalizations they wish to review. How meta-analysts formulate their purposes for review, and, secondarily, how they code study characteristics and calculate effect sizes, determine the breadth of categories they employ. Colleagues may complain that a construct is too broad to be interpretable or practical, or too narrow to provide an overview of a literature. But meta-analysts, not meta-analytic method, determine whether apples and oranges are mixed in overly broad categories.

Garbage In, Garbage Out

Another common criticism of meta-analysis (e.g., Eysenck 1978) concerns the quality of the primary research included in reviews. It has been claimed that meta-analysis is too inclusive and too willing to accept data from poorly designed studies in an effort to be comprehensive. Would it not be better to highlight the findings from a handful of well-designed studies than to give equal attention to the results of good and bad studies alike?

In principle, exclusivity has some merit, but reviewers invariably disagree about what constitutes good quality research. Glass (1976; Glass et al. 1981) argued that excluding studies a priori may lose data needlessly if quality of research has no relation with study outcomes. Glass' empirical response was to code threats to validity as independent variables and test their relation to treatment effects. If no relations exist, studies can be combined regardless of quality.

Glass' response is not an entirely satisfactory one. Good and poor studies may not differ in mean effect size, but in distribution. Differential distributions related to study quality could add considerable imprecision to average effect sizes, especially when categorizing studies into smaller groups according to study features. One also needs to consider the meta-analysis' credibility. Some studies are so notoriously or obviously flawed that to include them would cast doubt on overall findings.

No reviewer can escape issues of inclusion. Even the most inclusive meta-analysts exclude some studies from their reviews, perhaps case studies or pre-post designs. In all cases, meta-analysts should report inclusion criteria explicitly so that readers can determine how the sample of studies was formed and if adequate attention was paid to study quality.

The "garbage in, garbage out" complaint reflects concern with the study retrieval phase of meta-analysis. Like the complaint about apples and oranges, it is directed more at implementation than at meta-analysis itself. Meta-analysts may attend insufficiently to study quality, but nothing about meta-analytic method necessitates such attention or inattention.

Oversimplification of Research

It is tempting to see meta-analysis' walk-away message in terms of main effects, and results of meta-analyses are sometimes cited solely for their average findings (Bloom 1984; Niemiec et al. 1986). Critics have complained that meta-analysis collapses complex and subtle scholarship into single numerical representations (Cook and Leviton 1980). Such oversimplification does gross injustice to hard-fought debates in a field.

Historical accident may have fostered the idea that average effects are meta-analyses' most important products. Some early meta-analyses emphasized average results and only secondarily examined effect size variation (Cooper 1979; Rosenthal 1976). Early meta-analyses that studied effect size variation often defined their constructs broadly and thus appeared to oversimplify the reviewed literature (Smith and Glass 1977).

Ironically, meta-analyses also may appear to oversimplify a literature when they suggest a resolution to confusion in findings. For example, excitement about using simple computer applications as instructional tools for improving student achievement has not been justified by meta-analyses (Bangert-Drowns 1993; Hembree and Dessart 1986; Russell 1991). For researchers and practitioners who have committed considerable resources to such issues, or policymakers who publicly advocated some side of a debate, reviews that yield such convincing evidence may seem too simple.

Certainly meta-analysis is a method of data reduction, but it does not oversimplify a literature necessarily. In fact, most current meta-analyses examine variation in study outcomes and thus describe not just overall effect magnitude, but relations among variables. A particular meta-analysis could be criticized for defining its domain or its constructs too broadly, analyzing data in an overly simplistic way, or only emphasizing measures of central tendency in the findings. When valid, these criticisms reflect problematic implementation rather than a fault of meta-analytic method per se.

LIMITATIONS OF META-ANALYSIS

Given that common criticisms of meta-analysis more often describe problematic implementations than the method itself, does this mean that meta-analysis is limited only by the ingenuity of the reviewer? In spite of apparent objectivity and precision in systematic, quantitative review, two fundamental factors independent of statistical issues determine the validity and replicability of meta-analytic findings. First, the conclusions of a meta-analysis reflect the many judgments of a meta-analyst as much as the reviewed literature. Second, meta-analysis depends on characteristics of the reviewed literature.

Empirical Examinations of Human Judgments and Literature Characteristics in Meta-Analysis

Several investigators looked at ways in which human judgment and literature characteristics affect the process and outcomes of meta-analysis. Steiner and colleagues (1991), for example, found 35 meta-analyses in the literature on organizational behavior and human resources management. They coded these reviews on 10 variables: degree to which the review is theory based, method for locating studies, attention to potentially unretrieved studies ("file drawer problem"), elimination of studies, assumption of independent effect sizes, control for artifacts, type of meta-analysis used, method for locating moderators, quality of data presentation, and subtlety of interpretation. Steiner and colleagues then analyzed trends among the coded features of the 35 meta-analyses.

Most of the meta-analyses did not test theoretical propositions but averaged effects for different relations under different conditions. The meta-analysts showed insufficient sensitivity to the limits of their data, making causal claims from correlational findings or claiming generalizations on the basis of small data sets. Steiner and colleagues noted time trends in meta-analytic methods. Meta-analysts combined probabilities less frequently and used methods recommended by Hunter and Schmidt (Hunter et al. 1982; Hunter and Schmidt 1990) more frequently. Meta-analysts also more regularly took one effect

size from each study to maintain the independence of their data points.
Wanous and colleagues (1989) located four pairs of meta-analyses, each pair reviewing identical topics in organizational psychology and behavior. The authors divided meta-analytic method into 11 subtasks: defining the domain, establishing inclusion criteria, searching for studies, selecting studies, extracting data, coding for independent variables, deciding whether to group independent and dependent variables, determining the mean and variance of effect sizes, deciding whether to search for moderators, selecting potential moderators, and determining means and variances for effect sizes of subgroups. According to the authors, all of these tasks except those based on numerical calculation (determining means and variances for effect sizes and deciding whether to search for moderator variables) are acts of human judgment. The authors attempted to isolate the causes of discrepant findings within each pair in terms of the 11 subtasks.

The Wanous study is a conservative test of the effects of human judgment on meta-analytic findings. Pairs were selected for conceptual similarity, so they could not differ on step 1 (defining the domain). All pairs used the same meta-analytic techniques (Hunter et al. 1982) and their overall conclusions, not the analyses of moderators, were the products that primarily were compared. In short, pairs were selected and analyzed on criteria that favored similarity to simplify comparison.

Despite the conservative features of the Wanous study, human judgment did affect meta-analytic findings. In the early phases of these meta-analyses (e.g., determining inclusion criteria, locating studies, selecting studies), reviewers created different collections of effect sizes to analyze, and these differences explained most discrepancies in findings. Some discrepancies in findings resulted from minor judgment differences, the inclusion of a single unpublished study in one case. Fortunately, the explicit nature of meta-analysis allowed Wanous and colleagues to identify the specific sources of discrepancies within pairs.

Abrami and colleagues (1988) compared six meta-analyses of the validity of student ratings of instructional effectiveness to determine causes for their discrepant conclusions. They resolved meta-analysis into five subtasks: specifying inclusion criteria, locating studies, coding study features, calculating individual study outcomes, and data analysis.

The reviews differed greatly on each subtask, even though one author produced three of the six meta-analyses. The reviewers agreed on five inclusion criteria, but irregularly employed another seven. Evaluated against an independent exhaustive search of the literature, reviews differed greatly in comprehensiveness (ranging from 20 percent to 88 percent) and in the number of studies incorrectly included. Only one meta-analysis looked for relations between study features and study outcomes. There was only 47 percent agreement among the six meta-analyses regarding which effects to include and their estimates of magnitude. Finally, the reviewers differed in the ways they analyzed the effect sizes, some using weighting, others not, some using conventional statistical tests, others checking for variance attributable to sampling error.

Matt (1989) examined one facet of one feature checked by Abrami and colleagues (1988), scrutinizing a single decision point: How does one decide which effect sizes to include when several can be obtained from one study? Matt recoded 25 studies used in Smith and Glass' (1977) psychotherapy meta-analysis, applied Smith and Glass' original decision rule (the conceptual redundancy rule), and compared it to three other decision rules (the coder agreement, outcome reliability, and outlier truncation rules). The author and two other coders independently calculated effect sizes for the 25 studies and compared them to Smith and Glass' findings.

In terms of number of effect sizes and their magnitudes, all the raters showed considerable differences; and the differences were even greater when the raters compared their results to those of Smith and Glass. This single decision point made considerable differences among the raters' outcomes. The author concluded: "Point estimates of an intervention effect have particularly captured the attention of consumers of meta-analyses. Unfortunately, such point estimates are particularly affected by variation in the mostly implicit rules regarding the selection of effect sizes within studies, and it will often be desirable to present a range of defensible and appropriate estimates based on a number of different techniques, all imperfect but with different weaknesses" (Matt 1989, p. 113).

Dependence on Human Judgment

At each phase of meta-analysis, reviewers must make significant judgments guided by common sense and informed personal preference. These decisions can affect meta-analytic outcomes and deserve careful consideration.

Formulation of the Problem. The most fundamental decisions in a review—the domain to be reviewed, the nature and breadth of the constructs and operationalizations to be considered, and the specific questions to be addressed—are all products of human judgment. These most fundamental decisions constrain all subsequent phases of a meta-analysis.

Retrieval of Studies. This stage includes three important substeps. First, a reviewer must decide the comprehensiveness of the search. Studies can be located from various sources and with varying completeness. Some reviewers limit the extent of their searches to published research, research cited in previous prominent reviews, or studies conducted after a certain date.

Once potentially useful studies are identified, they must be obtained. Actually obtaining copies of identified documents is not always possible, but this is typically a logistical problem, not an issue of reviewer judgment.

Human judgment enters this phase of review most significantly after documents are obtained. A reviewer must determine which studies to include in the review. Decisions to exclude studies are sometimes quite easy, as in cases when an obviously irrelevant study was obtained erroneously. Other inclusion criteria, such as those based on quality of research, may be more unreliable and personal. By clearly and explicitly describing search strategies and inclusion criteria, meta-analysts at least open these decisions to public scrutiny and evaluation, but such explicitness does not mitigate the effects of meta-analysts' judgments.

Coding of Study Characteristics. At least two kinds of judgment operate in the coding of study features. First, meta-analysts must choose which study characteristics will receive detailed examination. They choose these variables for many reasons. Theory or practice suggests relations between some treatment variables and effect size. Reviewers test methodological variables to see if they are confounded by study outcomes. Other variables describe the range of settings and subjects represented in the studies. Because choice of study features is the result of personal insight and preference, scholars may disagree about the most important features to select.

After features are selected, coding itself reflects many acts of judgment. Reports often lack detail or clarity and require some guesswork to reconstruct the most probable research scenario. Some variables require personal judgment even with the clearest reports. Variables that estimate treatment intensity or qualities of interpersonal relations, for example, are difficult to code but likely to be influential in social science phenomena.

Calculation of Effect Sizes. Meta-analysts translate measures in studies to a common metric of treatment effect or relation between variables. Usually, the common measure is either the correlation coefficient or the standardized difference between two group means (i.e., the difference between group means divided by the pooled standard deviation).

Though meta-analysts agree about how to calculate effect sizes (Glass et al. 1981), meta-analysts must exercise personal judgment in deciding when to calculate them. Imagine, for example, an evaluation of a substance abuse education program that employed three measures of knowledge. One is a more reliable instrument than the others, the second provides more comprehensive coverage of the program’s content, and the third is a locally developed measure and thus most likely to be sensitive to local context. Should the meta-analyst select one dependent measure that somehow provides the "best" representation of treatment effects on knowledge, average the effects measured on all three tests, or include them all in the meta-analysis? Alternatively, the reviewer could calculate effect sizes for all three and meta-analyze the dependent measures separately: a meta-analysis for most reliable measures, a meta-analysis for most comprehensive measures, and a meta-analysis for local tests.

Internal contradictions and apparent reporting errors, research biases, selective presentation of only significant findings, or extremely positive or negative scores indicate potential problems for calculation of effect sizes. The careful meta-analyst must develop consistent and reasonable strategies for treatment of each kind of problem.

Investigation of Central Tendency and Variation in Effect Sizes. If the effect sizes obtained from a group of studies were identical, there would be no need for a literature review. Generally speaking, there are two approaches to analyzing effect size variation. One can consider each effect size as an irreducible data point and treat variation among effect sizes as analogous to variation among independent subjects in primary research (Glass et al. 1981; Kulik and Kulik 1989). Reviewers who take this view tend to use conventional statistical tests for research integration. Alternatively, meta-analysts can examine variation among effect sizes in light of the variation that one might expect from sampling error within each study (Hedges and Olkin 1985; Hunter and Schmidt 1990). Some of these researchers advocate the use of tests of homogeneity. In either case, the meta-analyst seeks to find relations between the coded study features and study effects.

Meta-analysts continue to debate the appropriateness of various analytic strategies. Assumptions of conventional statistical tests are often not met in research integration. However, meta-analytic approaches accounting for sampling error favorably weight studies with larger samples regardless of their quality. Tests of homogeneity overvalue statistical significance; statistically significant heterogeneity may be practically unimportant, and nonsignificance does not disprove heterogeneity. Some authors criticize any univariate analyses in research integration as overly simplistic and advocate multivariate analysis techniques.

At present, it is impossible to identify any one analytic strategy as trouble free. Selection of analytic method is a decision that balances the quality of available data with the various risks of alternative methods. A multimethod approach only postpones the decision. If the results of such a multimethod approach are contradictory, the reviewer then must decide which conclusions are most accurately descriptive of the literature.

Interpretation and Publication of Findings. Publication of findings requires significant decisions on the part of the reviewer. The reviewer must interpret the results of data analysis in light of the initial problem statement. Though quantitative analysis may indicate the statistical significance of relations among variables, the meta-analyst must decide which relations are practically significant for theoretical, practical, or policy implications. When several variables are significantly related to study outcomes, the meta-analyst must attempt to explain how these variables are interrelated.

Given constraints on publication space, meta-analysts cannot report many of their decisions. The meta-analyst must balance thorough and explicit exposition with conciseness and select which aspects of method will be reported. Judgments regarding publication link with another series of judgments that also determine the effectiveness of a review: the judgments of readers. The meta-analyst not only aims for accurate and valid integration, but for presentation that is both convincing and useful for the intended audience, whether researcher, policymaker, or practitioner.

Dependence on Primary Research

Obviously, human subjectivity and judgment interject into the meta-analytic process in many ways and with significant impact. This is not to disparage meta-analysis. In spite of its efforts to be precise, comprehensive, and objective, meta-analysis is not a technical feat, but demands as much subtle expertise as any other act of scholarship.

In addition to its dependence on judgment, meta-analysis is also fundamentally dependent on the primary research it integrates. Though an obvious observation, there are a number of less obvious implications that constrain interpretations that are possible from meta-analysis.

Meta-Analysis as a Particular Form of Literature Review. At least four types of literature review can be distinguished (Cooper 1982; Jackson 1980). Meta-analysis is a quantitative form of integrative review. Integrative reviews summarize findings from numerous studies that obtain apparently contradictory results, although the studies use a consistent research design to ask the same fundamental question. The integration of many such literatures in the social sciences is an important scholarly effort. But the comprehensive, statistical integration of contradictory empirical findings, the chief purpose of meta-analysis, is not the only goal of literature review.

Reviews can have at least three other purposes. Some may highlight pioneering methodological developments or theoretical formulations, examining only preliminary research at the cutting edge. Other reviews integrate concepts that appear in disparate literatures, drawing parallels among constructs previously considered distinct or connecting distinct constructs in larger theoretical formulations. Other reviews examine evidence to confirm or refute particular theories. These types of review might benefit from statistical analysis, but they rely primarily on conceptual analysis and do not strive to resolve contradictory findings through comprehensive integration of consistent studies.

Constraints on Questions That a Meta-Analysis Can Ask. Only their resources and creativity constrain primary researchers in the kinds of theoretical, practical, or policy-related questions they can investigate. Certainly good primary research builds on relevant work that precedes it, but the researcher is relatively unfettered in developing hypotheses, operationalizing constructs, and determining the complexity of research design.

Meta-analysts are far more constrained in their work. Meta-analysts must frame their inquiry in terms that permit the inclusion of a reasonably sized sample of studies. Meta-analysts typically frame their questions in terms of constructs frequently used in the literature of interest, and labels for these constructs and their most common operationalizations become the keywords in the search for useful studies.

Meta-analysis has some independence from primary research. Meta-analysts, for example, can integrate different literatures if some underlying construct unites them (e.g., combining teenage pregnancy prevention, smoking prevention, alcohol education, and drug prevention interventions to answer questions about public health prevention programs). Also, meta-analysts ask questions that can only be answered in a multistudy context. For example, only integrative research can ask, "Have public health prevention programs been more effective under different federal administrations?"

However, the meta-analyst cannot answer questions from literature that does not provide necessary data, and, because meta-analysis is a statistical analysis, the data must be drawn from a number of studies. Primary researchers must describe treatment and setting characteristics in sufficient detail to permit reviewers to code their salient features. Does substance abuse education affect males and females differently? The meta-analyst would be helpless to answer unless primary researchers distinguish their findings by gender.

Meta-analysts commonly conceive of effect magnitude in terms of relations between two variables, partly for the sake of simplicity of interpretation, but also because, if an effect size measures an interaction within a large group of variables, few studies will measure that same interaction. Meta-analysis then tends to favor simpler research designs that highlight comparisons between two variables at a time.

Biases in the Literature. Whole collections of studies sometimes can reflect biases that may or may not be readily detectable. Even if detectable, correction or interpretation of such biases is not always straightforward.

Meta-analysts often test for publication bias to see if study findings are related to their source. The average effect size from unpublished studies is commonly different from, and often smaller than, the average effect from published studies. It is not clear, however, why published research would yield different findings from unpublished research. Perhaps journal editors prefer statistically significant findings, thereby inadvertently elevating published treatment effects. Such a claim, though plausible, casts doubt on all scholarly publication. Alternatively, doctoral students and researchers with limited methodological experience may produce the bulk of unpublished research, while more experienced researchers publish their work. It is not possible to interpret confidently this common meta-analytic finding.

Publication bias is but one example of biases that can permeate a group of studies. Primary researchers do not research topics at random, but select ones likely to attract funding, employ constructs developed by previous successful researchers, produce statistically significant findings, and finally find publication. Such a researcher preference bias will determine whether or not there are sufficient studies to do an integrative review on a given question. Meta-analysis, and literature review in general, is by definition retrospective and therefore reflects what has been done rather than what could be done. The retrospective bias of meta-analysis may significantly misrepresent phenomena that experience rapid innovation, such as computer-based instruction.

Finally, conventions within a domain may bias findings and leave the meta-analyst helpless to correct it. For example, those who research drunk driving generally agree that rearrest rate is a problematic measure of rehabilitation effectiveness. Localities differ considerably in enforcement intensity and strategy and in the severity with which offenders are prosecuted. Even with rigorous enforcement the likelihood of arrest for driving while intoxicated is quite low. Researchers admit that rearrest rates are too insensitive to accurately measure the effects of rehabilitation programs, but rearrest is still the most commonly used measure of treatment effectiveness, primarily because it is such an attractive bottom-line measure for policymakers. It is impossible for the meta-analyst to substitute a more sensitive measure of rehabilitation effectiveness because the meta-analyst is dependent on the primary research.

Small Samples. Data drawn from the same study are not truly independent. The resources, settings, implementations, and personal impact of the researcher leave an indelible mark on all the subjects in a particular study. Though findings might come from numerous effect sizes and thousand of subjects, the number of studies, which roughly correlates to the number of research settings, is a better indicant of the comprehensiveness of a meta-analysis.

Given this standard of comprehensiveness, most meta-analyses are based on relatively small samples. For example, the median number of studies included in 35 meta-analyses reviewed by Steiner and colleagues (1991) was 43. It is not unusual to examine hundreds of documents, but finally settle on a sample of well less than 100 studies.

Nonexperimental Design. Reviewers do not randomly sample studies or randomly assign them to conditions for comparison; they take study conditions as delivered by the primary researcher. Meta-analysis is nonexperimental correlational research, defining important relations among variables but rarely able to determine causal links. The meta-analyst may only speculate about causal relations and triangulate evidence to bolster causal claims. For example, between-study comparisons might relate peer leadership to higher effect sizes in substance education programs. If within-study treatment comparisons show the same pattern, a reviewer could claim more confidently that the nature of program leadership influences program effectiveness.

SOME CRITERIA FOR JUDGING THE QUALITY OF A META-ANALYSIS

As a quantitative integrative review, meta-analysis possesses limited aims: the integration of studies with similar research goals and methods but contradictory results. The quality of a meta-analysis is defined in part by statistical adequacy, but, perhaps even more by the reviewer's craft knowledge and constraints imposed on that craft by the available literature.

Given the importance of craft, how does one determine the quality of a quantitative review? There are some general features by which readers can evaluate the quality of a quantitative literature review.

Comprehensiveness

How was the sample of studies gathered? Some reviews limit searches to specific sources, and this should be explicitly stated in the review. How- ever, other factors being equal, readers should give greater weight to more exhaustive reviews. Exhaustive reviews are not necessarily those with large numbers of effect sizes or subjects, but reviews that include virtually all the published and unpublished research reasonably available on the defined question. Such reviews take into account conclusions from the largest number of researchers drawn from the largest number of settings.

A comprehensive search strategy should locate studies in relevant databases and institutional clearinghouses as well as previous prominent literature reviews on the topic of interest. Inclusion criteria should be stated explicitly and reflect a balance between attention to the internal validity of the studies as well as the external validity and comprehensive-ness of the review. Exclusion of large bodies of research that may bias the outcomes of the review should be carefully evaluated.

Calculation of Effect Sizes

Effect sizes must be calculated correctly. Many meta-analyses report names, major features, and effect sizes of included studies, and readers can scan these lists for unusual outliers or noticeable errors. Authors should define explicitly how they calculated effect sizes. When the effect size is the standardized mean difference (such as Cohen's ‘d', Glass' ‘ES', or Hedges' ‘g'), effect sizes all must be standardized by raw score variation rather than variation corrected for covariance. Effect sizes calculated from corrected variances are incomparable with effect sizes from uncorrected variance. Corrections reduce raw score variance; effect sizes calculated with reduced variances will appear larger and thus spuriously appear to represent superior treatments.

Studies often offer more than one effect size either from multiple criteria or from various subdivisions of the sample. How does the meta-analyst handle multiple effect sizes? The reader should check first for the apples-and-oranges problem. A meta-analyst may define broad constructs to investigate, but combining some operationalizations, especially dependent variables, may not be defensible. For example, no common construct underlies measures of knowledge, attitude, and behavior, and an average effect across measures is difficult to interpret. Such an average might suggest that substance abuse education is highly successful when in fact it may only be successful with knowledge outcomes but not with attitude and behavior.

Readers also should check the ratio of the number of effect sizes to the number of studies. For analysis of any criterion, it is best to have nearly a one-to-one correspondence between studies and effects. If a study contributes more than one effect to analysis, those effect sizes cannot be considered independent, and studies contributing the most effect sizes are overrepresented in the calculation of averages. Occasional violations of one-to-one correspondence are permissible, but as the ratio of effects to studies increases, it becomes more difficult to interpret the analysis of effect sizes.

In some meta-analyses, effect sizes are weighted by study features such as sample size, sampling error, or quality. Such weighting strategies complicate the interpretation of meta-analytic findings. An advantage of effect size over other statistics such as ‘t' and ‘F' is precisely that it is independent of sample size; weighting by sample size or sampling error (including strategies for testing homogeneity) gives greater importance to studies with large samples regardless of the quality of their design or implementation. Weighting by quality introduces other problems. Scholars differ about how to define quality of research, but even if there were agreement in definition, there certainly would be disagreement about the appropriate weights for different qualities. In general, if weighted effect sizes are analyzed, these results should be compared to analyses of unweighted effect sizes to check if differences are meaningful or artifactual.

Analysis of Effect Size Variation

A good meta-analysis not only calculates effect sizes and their average, but attempts to identify variables that explain variation in study findings. Analysis of effect size variation poses several problems. First among these is sample size. The more variables involved in an effort to explain variation, the larger the number of effect sizes (and thus of studies) needed. Overall there should be a large ratio of effect sizes (studies) to variables examined, and categorical variables should have respectable numbers of effects in each level. Other things being equal, reviews examining the larger number of studies are better suited to investigating effect size variation.

A second problem with analysis of effect size variation is the disagreement among meta-analysts about appropriate methods. Visual methods, conventional statistical tests, tests of homogeneity, consideration of sampling error and variation due to artifacts without significance testing, and multivariate and path analytic techniques have been recommended. Any statistical procedure could potentially be applied to research integration, so the reader must keep informed about alternate methods and judge whether a particular implementation is convincing and competent.

Interpretation of Findings

The meta-analyst must not conclude more than the data suggest. With small samples, nonrandom assignment of studies to conditions, the vagaries of human judgment, and the limitations of the literature, conclusions of meta-analysis are largely speculative, "best guesses" of treatment effects and relations among variables. Meta-analysts need to avoid making causal claims on the basis of correlational data, unless such claims are explicitly tentative or unless there are within-study comparisons that support the between-study findings.

A meta-analysis rarely completes the research in a domain and, in fact, often can raise new questions about methodology or relations among variables. An important part of the interpretative portion of a meta-analysis identifies remaining questions or new questions that require additional research.

A META-ANALYTIC VIEW OF META-ANALYTIC FINDINGS

Given the many ways in which human judgment and limitations of the literature can determine the findings of a meta-analysis, it is best to keep a meta-analytic attitude toward meta-analytic findings. That is, a careful reader should compare the findings of any given meta-analysis to the conclusions of other reviews on the same topic, looking for consistencies and inconsistencies among them. Consistencies among reviews, especially when they were independently developed or used different techniques, contribute to the confidence one can place in the findings. The reader should attempt to locate reasons for inconsistencies in findings and either resolve the inconsistency or leave the debate for further primary research.

REFERENCES

Abrami, P.C.; Cohen, P.A.; and d’Apollonia, S. Implementation problems in meta-analysis. *Rev Educ Res* 58(2):151-179, 1988.

Bangert-Drowns, R.L. Review of developments in meta-analytic method. *Psychol Bull* 99(3):388-399, 1986.

Bangert-Drowns, R.L. The word processor as an instructional tool: A meta-analysis of word processing in writing instruction. *Rev Educ Res* 63(1):69-93, 1993.

Bloom, B.S. The 2 sigma problem: The search for methods of instruction as effective as one-to-one tutoring. *Educ Res* 13(6):4-16, 1984.

Carlberg, C.G.; Johnson, D.W.; Johnson, R.; Maruyama, G.; Kavale, K.; Kulik, C.-L.C.; Kulik, J.A.; Lysokowski, R.S.; Pflaum, S.W.; and Walberg, H.J. Meta-analysis in education: A reply to Slavin. *Educ Res* 13(8):16-23, 1984.

Clark, R.E. Confounding in educational computing research. *J Educ Comput Res* 1(2):137-148, 1985.

Cook, T., and Leviton, L. Reviewing the literature: A comparison of traditional methods with meta-analysis. *J Pers* 48(4): 449-472, 1980.

Cooper, H.M. Statistically combining independent studies: A meta-analysis of sex differences in conformity research. *J Pers Soc Psychol* 37(1):131-146, 1979.

Cooper, H.M. Scientific guidelines for conducting integrative reviews. *Rev Educ Res* 52(2):291-302, 1982.

Eysenck, H.J. An exercise in mega-silliness. *Am Psychol* 33(5):517, 1978.

Gallo, P. Meta-analysis—mixed meta-phor? *Am Psychol* 3(5):515-517, 1978.

Glass, G.V. Primary, secondary, and meta-analysis of research. *Educ Res* 10(5):3-8, 1976.

Glass, G.V.; McGaw, B.; and Smith, M.L. *Meta-Analysis in Social Research*. Beverly Hills, CA: Sage Publications, 1981.

Guzzo, R.A.; Jackson, S.E.; and Katzell, R.A. Meta-analysis. In: Cummings, L.L., and Staw, B.M., eds. *Research in Organizational Behavior*. Greenwich, CT: JAI Press, 1987. pp. 407-422.

Hedges, L.V. How hard is hard science, how soft is soft science? *Am Psychol* 42(5):443-455, 1987.

Hedges, L.V., and Olkin, I. *Statistical Methods for Meta-Analysis*. San Diego, CA: Academic Press, 1985.

Hembree, R., and Dessart, D.J. Effects of hand-held calculators in precollege mathematics education: A meta-analysis. *J Res Math Educ* 17(2):83-99, 1986.

Hunter, J.E., and Schmidt, F.L. *Methods of Meta-Analysis*. Newbury Park, CA: Sage Publications, 1990.

Hunter, J.E.; Schmidt, F.L.; and Jackson, G.B. *Meta-Analysis: Cumulating Research Findings Across Studies*. Beverly Hills, CA: Sage Publications, 1982.

Jackson, G.B. Methods for integrative reviews. *Rev Educ Res* 50(3):438-460, 1980.

Kulik, J.A., and Kulik, C.-L.C. Meta-analysis in education. *Int J Educ Res* 13(3):221-340, 1989.

L'Hommedieu, R.; Menges, R.J.; and Brinko, K.T. Validity issues in meta-analysis: Suggestions for research and policy. *Higher Educ Res Dev* 7(2):119-130, 1988.

Matt, G.E. Decision rules for selecting effect sizes in meta-analysis: A review and reanalysis of psychotherapy outcome studies. *Psychol Bull* 105(1):106-115, 1989.

Niemiec, R.P.; Blackwell, M.C.; and Walberg, H.J. CAI can be doubly effective. *Phi Delta Kappan* 67(10):750-751, 1986.

Presby, S. Overly broad categories obscure important differences between therapies. *Am Psychol* 33(5):514-515, 1978.

Rosenthal, R. Interpersonal expectancy effects: A follow-up. In: Rosenthal, R., ed. *Experimental Effects in Behavioral Research.* New York: Irvington, 1976. pp. 440-471.

Russell, R.G. "A Meta-Analysis of Wordprocessing and Attitudes and the Impact on the Quality of Writing." Paper presented at the annual meeting of the American Educational Research Association, Chicago, 1991.

Schmidt, F. What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *Am Psychol* 47(10):1173-1181, 1992.

Slavin, R.E. Meta-analysis in education: How has it been used? *Educ Res* 13(8):6-15, 1984.

Smith, M.L., and Glass, G.V. Meta-analysis of psychotherapy outcome studies. *Am Psychol* 32(9):752-760, 1977.

Steiner, D.D.; Lane, I.M.; Dobbins, G.H.; Schnur, A.; and McConnell, S. A review of meta-analyses in organizational behavior and human resources management: An empirical assessment. *Educ Psychol Meas* 51(3):609-626, 1991.

Wachter, K.W. Disturbed by meta-analysis? *Science* 241(4872):1407-1408, 1988.

Wanous, J.P.; Sullivan, S.E.; and Malinak, J. The role of judgment calls in meta-analysis. *J Appl Psychol* 74(2):259-264, 1989.

AUTHOR

Robert L. Bangert-Drowns, Ph.D.
Associate Professor
State University of New York at Albany
School of Education
1400 Washington Avenue
Albany, NY 12222\

National
Institute on
Drug
Abuse
Research
MONOGRAPH SERIES

While limited supplies last, single copies of the following monographs may be obtained free of charge from the National Clearinghouse for Alcohol and Drug Information (NCADI). Please also contact NCADI for information about other publications of the National Institute on Drug Abuse relevant to drug abuse research.

Additional copies may be purchased from the U.S. Government Printing Office (GPO) and/or the National Technical Information Service (NTIS) as indicated. NTIS prices are for paper copy; add $3.00 handling charge for each order. Microfiche copies also are available from NTIS. Prices from either source are subject to change.

Addresses are:

NCADI
National Clearinghouse for Alcohol and Drug Information
P.O. Box 2345
Rockville, MD 20852
(301) 468-2600
(800) 729-6686

GPO
Superintendent of Documents
U.S. Government Printing Office
P.O. Box 371954
Pittsburgh, PA 15220-7954
(202) 738-3238
FAX (202) 512-2233

NTIS
National Technical Information Service
U.S. Department of Commerce
Springfield, VA 22161
(703) 487-4650

For information on availability of NIDA Research Monographs from 1975-1996 and those not listed, write to NIDA, Public Information Branch, Room 10A-39, 5600 Fishers Lane, Rockville, MD 20857.

26 THE BEHAVIORAL ASPECTS
OF SMOKING.
Norman A. Krasnegor, Ph.D., ed. (Reprint from 1979 Surgeon General's
Report on Smoking and Health.)
NCADI #M26 NTIS PB #80-118755/AS (A09) $27.00

42 THE ANALYSIS OF CANNABINOIDS IN BIOLOGICAL FLUIDS.
Richard L. Hawks, Ph.D., ed.
NCADI #M42 NTIS PB #83-136044/AS (A07) $27.00

50 COCAINE: PHARMACOLOGY, EFFECTS, AND TREATMENT OF
ABUSE. John Grabowski, Ph.D., ed.
NCADI #M50 NTIS PB #85-150381/AS (A07) $27.00

52 TESTING DRUGS FOR PHYSICAL DEPENDENCE POTENTIAL
AND ABUSE LIABILITY. Joseph V. Brady, Ph.D., and Scott E. Lukas,
Ph.D., eds.
NCADI #M52 NTIS PB #85-150373/AS (A08) $27.00

53 PHARMACOLOGICAL ADJUNCTS IN SMOKING CESSATION.
John Grabowski, Ph.D., and
Sharon M. Hall, Ph.D., eds.
NCADI #M53 NTIS PB #89-123186/AS (A07) $27.00

54 MECHANISMS OF TOLERANCE AND DEPENDENCE.
Charles Wm. Sharp, Ph.D., ed.
NCADI #M54 NTIS PB #89-103279/AS (A19) $52.00

56 ETIOLOGY OF DRUG ABUSE: IMPLICATIONS FOR
PREVENTION. Coryl LaRue Jones, Ph.D., and
Robert J. Battjes, D.S.W., eds.
NCADI #M56 NTIS PB #89-123160/AS (A13) $36.50

61 COCAINE USE IN AMERICA: EPIDEMIOLOGIC AND CLINICAL
PERSPECTIVES. Nicholas J. Kozel, M.S., and
Edgar H. Adams, M.S., eds.
NCADI #M61 NTIS PB #89-131866/AS (A11) $36.50

62 NEUROSCIENCE METHODS IN DRUG ABUSE RESEARCH.
Roger M. Brown, Ph.D., and David P. Friedman, Ph.D., eds.
NCADI #M62 NTIS PB #89-130660/AS (A08) $27.00

63 PREVENTION RESEARCH: DETERRING DRUG ABUSE AMONG
CHILDREN AND ADOLESCENTS. Catherine S. Bell, M.S., and
Robert J. Battjes, D.S.W., eds.
NCADI #M63 NTIS PB #89-103287/AS (A11) $36.50

64 PHENCYCLIDINE: AN UPDATE. Doris H. Clouet, Ph.D., ed.
NCADI #M64 NTIS PB #89-131858/AS (A12) $36.50

65 WOMEN AND DRUGS: A NEW ERA FOR RESEARCH.
Barbara A. Ray, Ph.D., and Monique C. Braude, Ph.D., eds.
NCADI #M65 NTIS PB #89-130637/AS (A06) $27.00

69 OPIOID PEPTIDES: MEDICINAL CHEMISTRY.
Rao S. Rapaka, Ph.D.; Gene Barnett, Ph.D.; and
Richard L. Hawks, Ph.D., eds.
NCADI #M69 NTIS PB #89-158422/AS (A17) $44.50

70 OPIOID PEPTIDES: MOLECULAR PHARMACOLOGY,
BIOSYNTHESIS, AND ANALYSIS. Rao S. Rapaka, Ph.D., and
Richard L. Hawks, Ph.D., eds.
NCADI #M70 NTIS PB #89-158430/AS (A18) $52.00

72 RELAPSE AND RECOVERY IN DRUG ABUSE.
Frank M. Tims, Ph.D., and Carl G. Leukefeld, D.S.W., eds.
NCADI #M72 NTIS PB #89-151963/AS (A09) $36.50

74 NEUROBIOLOGY OF BEHAVIORAL CONTROL IN DRUG ABUSE.
Stephen I. Szara, M.D., D.Sc., ed.
NCADI #M74 NTIS PB #89-151989/AS (A07) $27.00

77 ADOLESCENT DRUG ABUSE: ANALYSES OF TREATMENT
RESEARCH. Elizabeth R. Rahdert, Ph.D., and
John Grabowski, Ph.D., eds.
NCADI #M77 NTIS PB #89-125488/AS (A0) $27.00

78 THE ROLE OF NEUROPLASTICITY IN THE RESPONSE TO
DRUGS. David P. Friedman, Ph.D., and
Doris H. Clouet, Ph.D., eds.
NCADI #M78 NTIS PB #88-245683/AS (A10) $36.50

79 STRUCTURE-ACTIVITY RELATIONSHIPS OF THE
CANNABINOIDS. Rao S. Rapaka, Ph.D., and
Alexandros Makriyannis, Ph.D., eds.
NCADI #M79 NTIS PB #89-109201/AS (A10) $36.50

80 NEEDLE SHARING AMONG INTRAVENOUS DRUG ABUSERS:
NATIONAL AND INTERNATIONAL PERSPECTIVES. Robert J.
Battjes, D.S.W., and
Roy W. Pickens, Ph.D., eds.
NCADI #M80 NTIS PB #88-236138/AS (A09) $36.50

82 OPIOIDS IN THE HIPPOCAMPUS. Jacqueline F. McGinty, Ph.D., and
David P. Friedman, Ph.D., eds.
NCADI #M82 NTIS PB #88-245691/AS (A06) $27.00

83 HEALTH HAZARDS OF NITRITE INHALANTS.
Harry W. Haverkos, M.D., and John A. Dougherty, Ph.D., eds.
NCADI #M83 NTIS PB #89-125496/AS (A06) $27.00

84 LEARNING FACTORS IN SUBSTANCE ABUSE.
Barbara A. Ray, Ph.D., ed.
NCADI #M84 NTIS PB #89-125504/AS (A10) $36.50

85 EPIDEMIOLOGY OF INHALANT ABUSE: AN UPDATE.
Raquel A. Crider, Ph.D., and Beatrice A. Rouse, Ph.D., eds.
NCADI #M85 NTIS PB #89-123178/AS (A10) $36.50

86 COMPULSORY TREATMENT OF DRUG ABUSE: RESEARCH AND
CLINICAL PRACTICE. Carl G. Leukefeld, D.S.W., and Frank M.
Tims, Ph.D., eds.
NCADI #M86 NTIS PB #89-151997/AS (A12) $36.50

87 OPIOID PEPTIDES: AN UPDATE. Rao S. Rapaka, Ph.D., and
Bhola N. Dhawan, M.D., eds.
NCADI #M87 NTIS PB #89-158430/AS (A11) $36.50

88 MECHANISMS OF COCAINE ABUSE AND TOXICITY.
Doris H. Clouet, Ph.D.; Khursheed Asghar, Ph.D.; and
Roger M. Brown, Ph.D., eds.
NCADI #M88 NTIS PB #89-125512/AS (A16) $44.50

89 BIOLOGICAL VULNERABILITY TO DRUG ABUSE.
Roy W. Pickens, Ph.D., and Dace S. Svikis, B.A., eds.
NCADI #M89 NTIS PB #89-125520/AS (A09) $27.00

92 TESTING FOR ABUSE LIABILITY OF DRUGS IN HUMANS.
Marian W. Fischman, Ph.D., and Nancy K. Mello, Ph.D., eds.
NCADI #M92 NTIS PB #90-148933/AS (A17) $44.50

93 AIDS AND INTRAVENOUS DRUG USE: FUTURE DIRECTIONS
FOR COMMUNITY-BASED PREVENTION RESEARCH. Carl G.
Leukefeld, D.S.W.; Robert J. Battjes, D.S.W.; and Zili Amsel, D.S.C.,
eds.
NCADI #M93 NTIS PB #90-148933/AS (A14) $44.50

94 PHARMACOLOGY AND TOXICOLOGY OF AMPHETAMINE AND
RELATED DESIGNER DRUGS. Khursheed Asghar, Ph.D., and Errol
De Souza, Ph.D., eds.
NCADI #M94 NTIS PB #90-148958/AS (A16) $44.50

95 PROBLEMS OF DRUG DEPENDENCE, 1989. PROCEEDINGS OF
THE 51st ANNUAL SCIENTIFIC MEETING. THE COMMITTEE ON
PROBLEMS OF DRUG DEPENDENCE, INC.
Louis S. Harris, Ph.D., ed.
NCADI #M95 NTIS PB #90-237660/AS (A99) $67.00

96 DRUGS OF ABUSE: CHEMISTRY, PHARMACOLOGY,
IMMUNOLOGY, AND AIDS. Phuong Thi Kim Pham, Ph.D., and
Kenner Rice, Ph.D., eds.
NCADI #M96 NTIS PB #90-237678/AS (A11) $36.50

97 NEUROBIOLOGY OF DRUG ABUSE: LEARNING AND MEMORY.
Lynda Erinoff, Ph.D., ed.
NCADI #M97 NTIS PB #90-237686/AS (A11) $36.50

98 THE COLLECTION AND INTERPRETATION OF DATA FROM
HIDDEN POPULATIONS.
Elizabeth Y. Lambert, M.S., ed.
NCADI #M98 NTIS PB #90-237694/AS (A08) $27.00

99 RESEARCH FINDINGS ON SMOKING OF ABUSED SUBSTANCES.
C. Nora Chiang, Ph.D., and
Richard L. Hawks, Ph.D., eds.
NCADI #M99 NTIS PB #91-141119 (A09) $27.00

100 DRUGS IN THE WORKPLACE: RESEARCH AND EVALUATION DATA. VOL II. Steven W. Gust, Ph.D.; J. Michael Walsh, Ph.D.; Linda B. Thomas, B.S.; and
Dennis J. Crouch, M.B.A., eds.
NCADI #M100 GPO Stock #017-024-01458-3 $8.00

101 RESIDUAL EFFECTS OF ABUSED DRUGS ON BEHAVIOR.
John W. Spencer, Ph.D., and John J. Boren, Ph.D., eds.
NCADI #M101 NTIS PB #91-172858/AS (A09) $27.00

102 ANABOLIC STEROID ABUSE. Geraline C. Lin, Ph.D., and
Lynda Erinoff, Ph.D., eds.
NCADI #M102 NTIS PB #91-172866/AS (A11) $36.50

103 DRUGS AND VIOLENCE: CAUSES, CORRELATES, AND CONSEQUENCES. Mario De La Rosa, Ph.D.;
Elizabeth Y. Lambert, M.S.; and Bernard Gropper, Ph.D., eds.
NCADI #M103 NTIS PB #91-172874/AS (A13) $36.50

104 PSYCHOTHERAPY AND COUNSELING IN THE TREATMENT OF DRUG ABUSE. Lisa Simon Onken, Ph.D., and Jack D. Blaine, M.D., eds.
NCADI #M104 NTIS PB #91-172874/AS (A07) $27.00

106 IMPROVING DRUG ABUSE TREATMENT.
Roy W. Pickens, Ph.D.; Carl G. Leukefeld, D.S.W.; and
Charles R. Schuster, Ph.D., eds.
NCADI #M106 NTIS PB #92-105873(A18) $50.00

107 DRUG ABUSE PREVENTION INTERVENTION RESEARCH: METHODOLOGICAL ISSUES. Carl G. Leukefeld, D.S.W., and William J. Bukoski, Ph.D., eds.
NCADI #M107 NTIS PB #92-160985 (A13) $36.50

108 CARDIOVASCULAR TOXICITY OF COCAINE: UNDERLYING MECHANISMS. Pushpa V. Thadani, Ph.D., ed.
NCADI #M108 NTIS PB #92-106608 (A11) $36.50

109 LONGITUDINAL STUDIES OF HIV INFECTION IN INTRAVENOUS DRUG USERS: METHODOLOGICAL ISSUES IN NATURAL HISTORY RESEARCH. Peter Hartsock, Dr.P.H., and Sander G. Genser, M.D., M.P.H., eds.
NCADI #M109 NTIS PB #92-106616 (A08) $27.00

111 MOLECULAR APPROACHES TO DRUG ABUSE RESEARCH: RECEPTOR CLONING, NEUROTRANSMITTER EXPRESSION, AND MOLECULAR GENETICS: VOLUME I. Theresa N.H. Lee, Ph.D., ed.
NCADI #M111 NTIS PB #92-135743 (A10) $36.50

112 EMERGING TECHNOLOGIES AND NEW DIRECTIONS IN DRUG ABUSE RESEARCH. Rao S. Rapaka, Ph.D.;
Alexandros Makriyannis, Ph.D.; and Michael J. Kuhar, Ph.D., eds.
NCADI #M112 NTIS PB #92-155449 (A15) $44.50

113 ECONOMIC COSTS, COST EFFECTIVENESS, FINANCING, AND COMMUNITY-BASED DRUG TREATMENT.
William S. Cartwright, Ph.D., and James M. Kaple, Ph.D., eds.
NCADI #M113 NTIS PB #92-155795 (A10) $36.50

114 METHODOLOGICAL ISSUES IN CONTROLLED STUDIES ON EFFECTS OF PRENATAL EXPOSURE TO DRUG ABUSE.
M. Marlyne Kilbey, Ph.D., and Khursheed Asghar, Ph.D., eds.
NCADI #M114 NTIS PB #92-146216 (A16) $44.50

115 METHAMPHETAMINE ABUSE: EPIDEMIOLOGIC ISSUES AND IMPLICATIONS. Marissa A. Miller, D.V.M., M.P.H., and
Nicholas J. Kozel, M.S., eds.
NCADI #M115 NTIS PB #92-146224/ll (AO7) $27.00

116 DRUG DISCRIMINATION: APPLICATIONS TO DRUG ABUSE RESEARCH. R.A. Glennon, Ph.D.;
Toubjörn U.C. Järbe, Ph.D.; and J. Frankenheim, Ph.D., eds.
NCADI #M116 NTIS PB #94-169471 (A20) $52.00

117 METHODOLOGICAL ISSUES IN EPIDEMIOLOGY, PREVENTION, AND TREATMENT RESEARCH ON DRUG-EXPOSED WOMEN AND THEIR CHILDREN.
M. Marlyve Kilbey, Ph.D., and Kursheed Asghar, Ph.D., eds.
GPO Stock #O17-024-01472-9 $12.00
NCADI #M117 NTIS PB #93-102101/LL (A18) $52.00

118 DRUG ABUSE TREATMENT IN PRISONS AND JAILS.
Carl G. Leukefeld, D.S.W., and Frank M. Tims, Ph.D., eds.
GPO Stock #O17-024-01473-7 $16.00
NCADI #M118 NTIS PB #93-102143/LL (A14) $44.50

120 BIOAVAILABILITY OF DRUGS TO THE BRAIN AND THE BLOOD-BRAIN BARRIER. Jerry Frankenheim, Ph.D., and
Roger M. Brown, Ph.D., eds.
GPO Stock #017-024-01481-8 $10.00
NCADI #M120 NTIS PB #92-214956/LL (A12) $36.50

121 BUPRENORPHINE: AN ALTERNATIVE TREATMENT FOR OPIOID DEPENDENCE. Jack D. Blaine, Ph.D., ed.
GPO Stock #017-024-01482-6 $5.00
NCADI #M121 NTIS PB #93-129781/LL (A08) $27.00

123 ACUTE COCAINE INTOXICATION: CURRENT METHODS OF TREATMENT. Heinz Sorer, Ph.D., ed.
GPO Stock #017-024-01501-6 $6.50
NCADI #M123 NTIS PB #94-115433/LL (A09) $27.00

124 NEUROBIOLOGICAL APPROACHES TO BRAIN-BEHAVIOR INTERACTION. Roger M. Brown, Ph.D., and
Joseph Fracella, Ph.D., eds.
GPO Stock #017-024-01492-3 $9.00
NCADI #M124 NTIS PB #93-203834/LL (A12) $36.50

125 ACTIVATION OF IMMEDIATE EARLY GENES BY DRUGS OF ABUSE. Reinhard Grzanna, Ph.D., and
Roger M. Brown, Ph.D., eds.
GPO Stock #017-024-01503-2 $7.50
NCADI #M125 NTIS PB #94-169489 (A12) $36.50

126 MOLECULAR APPROACHES TO DRUG ABUSE RESEARCH VOLUME II: STRUCTURE, FUNCTION, AND EXPRESSION. Theresa N.H. Lee, Ph.D., ed.
NCADI #M126 NTIS PB #94-169497 (A08) $27.00

127 PROGRESS AND ISSUES IN CASE MANAGEMENT.
Rebecca S. Ashery, D.S.W., ed.
NCADI #M127 NTIS PB #94-169505 (A18) $52.00

128 STATISTICAL ISSUES IN CLINICAL TRIALS FOR TREATMENT OF OPIATE DEPENDENCE.
Ram B. Jain, Ph.D., ed.
NCADI #M128 NTIS PB #93-203826/LL (A09) $27.00

129 INHALANT ABUSE: A VOLATILE RESEARCH AGENDA. Charles W. Sharp, Ph.D.; Fred Beauvais, Ph.D.; and
Richard Spence, Ph.D., eds.
GPO Stock #017-024-01496-6 $12.00
NCADI #M129 NTIS PB #93-183119/LL (A15) $44.50

130 DRUG ABUSE AMONG MINORITY YOUTH: ADVANCES IN RESEARCH AND METHODOLOGY. Mario De La Rosa, Ph.D., and Juan-Luis Recio Adrados, Ph.D., eds.
GPO Stock #017-024-01506-7 $14.00
NCADI #M130 NTIS PB #94-169513 (A15) $44.50

131 IMPACT OF PRESCRIPTION DRUG DIVERSION CONTROL SYSTEMS ON MEDICAL PRACTICE AND PATIENT CARE.
James R. Cooper, Ph.D.; Dorynne J. Czechowicz, M.D.;
Stephen P. Molinari, J.D., R.Ph.; and
Robert C. Peterson, Ph.D., eds.
GPO Stock #017-024-01505-9 $14.00
NCADI #M131 NTIS PB #94-169521 (A15) $44.50

132 PROBLEMS OF DRUG DEPENDENCE, 1992: PROCEEDINGS OF THE 54TH ANNUAL SCIENTIFIC MEETING OF THE COLLEGE ON PROBLEMS OF DRUG DEPENDENCE.
Louis Harris, Ph.D., ed.
GPO Stock #017-024-01502-4 $23.00
NCADI #M132 NTIS PB #94-115508/LL (A99)

133 SIGMA, PCP, AND NMDA RECEPTORS.
Errol B. De Souza, Ph.D.; Doris Clouet, Ph.D., and
Edythe D. London, Ph.D., eds.
NCADI #M133 NTIS PB #94-169539 (A12) $36.50

134 MEDICATIONS DEVELOPMENT: DRUG DISCOVERY, DATABASES, AND COMPUTER-AIDED DRUG DESIGN.
Rao S. Rapaka, Ph.D., and Richard L. Hawks, Ph.D., eds.
GPO Stock #017-024-01511-3 $11.00
NCADI #M134 NTIS PB #94-169547 (A14) $44.50

135 COCAINE TREATMENT: RESEARCH AND CLINICAL PERSPECTIVES. Frank M. Tims, Ph.D., and
Carl G. Leukefeld, D.S.W., eds.
GPO Stock #017-024-01520-2 $11.00
NCADI #M135 NTIS PB #94-169554 (A13) $36.50

136 ASSESSING NEUROTOXICITY OF DRUGS OF ABUSE. Lynda Erinoff, Ph.D., ed.
GPO Stock #017-024-01518-1 $11.00
NCADI #M136 NTIS PB #94-169562 (A13) $36.50

137 BEHAVIORAL TREATMENTS FOR DRUG ABUSE AND DEPENDENCE. Lisa Simon Onken, Ph.D.; Jack D. Blaine, M.D.; and John J. Boren, Ph.D., eds.
GPO Stock #017-024-01519-9 $13.00
NCADI #M137 NTIS PB #94-169570 (A15) $44.50

138 IMAGING TECHNIQUES IN MEDICATIONS DEVELOPMENT: CLINICAL AND PRECLINICAL ASPECTS. Heinz Sorer, Ph.D., and Rao S. Rapaka, Ph.D., eds.
NCADI #M138

139 SCIENTIFIC METHODS FOR PREVENTION INTERVENTION RESEARCH. Arturo Cazares, M.D., M.P.H., and Lula A. Beatty, Ph.D., eds.
NCADI #M139

140 PROBLEMS OF DRUG DEPENDENCE, 1993: PROCEEDINGS OF THE 55TH ANNUAL SCIENTIFIC MEETING, THE COLLEGE ON PROBLEMS OF DRUG DEPENDENCE, INC. VOLUME I: PLENARY SESSION SYMPOSIA AND ANNUAL REPORTS. Louis S. Harris, Ph.D., ed.
NCADI #M140

141 PROBLEMS OF DRUG DEPENDENCE, 1993: PROCEEDINGS OF THE 55TH ANNUAL SCIENTIFIC MEETING, THE COLLEGE ON PROBLEMS OF DRUG DEPENDENCE, INC. VOLUME II: ABSTRACTS. Louis S. Harris, Ph.D., ed.
NCADI #M141

142 ADVANCES IN DATA ANALYSIS FOR PREVENTION INTERVENTION RESEARCH. Linda M. Collins, Ph.D., and Larry A. Seitz, Ph.D., eds.
NCADI #M142

143 THE CONTEXT OF HIV RISK AMONG DRUG USERS AND THEIR SEXUAL PARTNERS. Robert J. Battjes, D.S.W.; Zili Sloboda, Sc.D.; and William C. Grace, Ph.D., eds.
NCADI #M143

144 THERAPEUTIC COMMUNITY: ADVANCES IN RESEARCH AND APPLICATION. Frank M. Tims, Ph.D.; George De Leon, Ph.D.; and Nancy Jainchill, Ph.D., eds.
NCADI #M144

145 NEUROBIOLOGICAL MODELS FOR EVALUATING MECHANISMS UNDERLYING COCAINE ADDICTION. Lynda Erinoff, Ph.D., and Roger M. Brown, Ph.D., eds.
NCADI #M145

146 HALLUCINOGENS: AN UPDATE. Geraline C. Lin, Ph.D., and Richard A. Glennon, Ph.D., eds.
NCADI #M146

147 DISCOVERY OF NOVEL OPIOID MEDICATIONS.
Rao S. Rapaka, Ph.D., and Heinz Sorer, Ph.D., eds.
NCADI #M147

148 EPIDEMIOLOGY OF INHALANT ABUSE: AN INTERNATIONAL PERSPECTIVE.
Nicholas J. Kozel, M.S.; Zili Sloboda, Sc.D.;
and Mario R. De La Rosa, Ph.D., eds.
NCADI #M148

149 MEDICATIONS DEVELOPMENT FOR THE TREATMENT OF PREGNANT ADDICTS AND THEIR INFANTS.
C. Nora Chiang, Ph.D., and Loretta P. Finnegan, M.D., eds.
NCADI #M149

150 INTEGRATING BEHAVIORAL THERAPIES WITH MEDICATIONS IN THE TREATMENT OF DRUG DEPENDENCE. Lisa Simon Onken, Ph.D.; Jack D. Blaine, M.D.; and John J. Boren, Ph.D., eds.
NCADI #M150

151 SOCIAL NETWORKS, DRUG ABUSE, AND HIV TRANSMISSION. Richard H. Needle, Ph.D., M.P.H.;
Susan L. Coyle, Ph.D.; Sander G. Genser, M.D., M.P.H.; and
Robert T. Trotter II, Ph.D., eds.
NCADI #M151

152 PROBLEMS OF DRUG DEPENDENCE 1994: PROCEEDINGS OF THE 56TH ANNUAL SCIENTIFIC MEETING, THE COLLEGE ON PROBLEMS OF DRUG DEPENDENCE, INC. VOLUME I: PLENARY SESSION SYMPOSIA AND ANNUAL REPORTS. Louis S. Harris, Ph.D., ed.
NCADI #M152

153 PROBLEMS OF DRUG DEPENDENCE 1994: PROCEEDINGS OF THE 56TH ANNUAL SCIENTIFIC MEETING, THE COLLEGE ON PROBLEMS OF DRUG DEPENDENCE, INC. VOLUME II: ABSTRACTS. (1995) Louis S. Harris, Ph.D., ed.
NCADI #M153 GPO Stock #017-024-01564-4 $22.00

154 MEMBRANES AND BARRIERS: TARGETED DRUG DELIVERY. (1995) Rao S. Rapaka, Ph.D., ed.
NCADI #M154 GPO Stock #017-024-01583-1 $10.00

155 REVIEWING THE BEHAVIORAL SCIENCE KNOWLEDGE BASE ON TECHNOLOGY TRANSFER. (1995)
Thomas E. Backer, Ph.D.; Susan L. David; and Gerald Soucy, Ph.D., eds.
NCADI #M155 GPO Stock #017-024-01581-4 $12.00

156 ADOLESCENT DRUG ABUSE: CLINICAL ASSESSMENT AND THERAPEUTIC INTERVENTIONS. (1995)
Elizabeth Rahdert, Ph.D.; Zili Sloboda, Ph.D.; and Dorynne Czechowicz, M.D., eds.
NCADI #M156 GPO Stock #017-024-01585-7 $14.00

157 QUALITATIVE METHODS IN DRUG ABUSE AND HIV RESEARCH. (1995)
Elizabeth Y. Lambert, M.S.; Rebecca S. Ashery, D.S.W.; and Richard H. Needle, Ph.D., M.P.H., eds.
NCADI #M157 GPO Stock #017-024-01581-4

158 BIOLOGICAL MECHANISMS AND PERINATAL EXPOSURE TO DRUGS. (1995) Pushpa V. Thadani, Ph.D., ed.
NCADI #M158 GPO Stock #017-024-01584-9

159 INDIVIDUAL DIFFERENCES IN THE BIOBEHAVIORAL ETIOLOGY OF DRUG ABUSE. (1996) Harold W. Gordon, Ph.D., and Meyer D. Glantz, Ph.D., eds.
NCADI #M159

161 MOLECULAR APPROACHES TO DRUG ABUSE RESEARCH. VOLUME III: RECENT ADVANCES AND EMERGING STRATEGIES. (1996) Theresa N.H. Lee, Ph.D., ed.
NCADI #M161

162 PROBLEMS OF DRUG DEPENDENCE, 1995. PROCEEDINGS FROM THE 57TH ANNUAL SCIENTIFIC MEETING OF THE COLLEGE ON DRUG DEPENDENCE, INC. (1996)
Louis Harris, Ph.D., ed.
NCADI #M162

163 NEUROTOXICITY AND NEUROPATHOLOGY ASSOCIATED WITH COCAINE/STIMULANT ABUSE. (1996)
Dorota Majewska, Ph.D., ed.
NCADI #M163

164 BEHAVIORAL STUDIES OF DRUG-EXPOSED OFFSPRING: METHODOLOGICAL ISSUES IN HUMAN AND ANIMAL RESEARCH. (1996) Cora Lee Wetherington, Ph.D.;
Vincent L. Smeriglio, Ph.D.; and Loretta P. Finnegan, Ph.D., eds.
NCADI #M164

165 BEYOND THE THERAPEUTIC ALLIANCE: KEEPING THE DRUG-DEPENDENT INDIVIDUAL IN TREATMENT. (1996) Lisa Simon Onken, Ph.D.; Jack D. Blaine, M.D.; and
John J. Boren, Ph.D., eds.
NCADI #M165

166 TREATMENT FOR DRUG-EXPOSED WOMEN AND CHILDREN: ADVANCES IN RESEARCH METHODOLOGY. (1996) Elizabeth Rahdert, Ph.D., ed.
NCADI #M166

167 THE VALIDITY OF SELF-REPORTED DRUG USE:
IMPROVING THE ACCURACY OF SURVEY ESTIMATES.
(1996) Lana Harrison, Ph.D., and Arthur Hughes, M.D., eds.
NCADI #M167

168 RURAL SUBSTANCE ABUSE: STATE OF KNOWLEDGE AND ISSUES. (1996) Zili Sloboda, Sc.D.;
Eric Rosenquist; and Jan Howard, Ph.D., eds.
NCADI #M168

169 TREATMENT OF PAIN IN ADDICTS AND OTHERS WHO MAY HAVE HISTORIES OF DEPENDENCE. (1996)
Alan I. Trachtenberg, M.D., M.P.H., F.A.A.F.P., D.AA.P.M., ed.
NCADI #M169