National Institute on Drug Abuse

RESEARCH

MONOGRAPH SERIES

Advances in Data Analysis for Prevention Intervention Research

142

Advances in Data Analysis for Prevention Intervention Research

Editors:

Linda M. Collins, Ph.D.
Department of Human Development and Family Studies
The Pennsylvania State University

Larry A. Seitz, Ph.D.
Division of Epidemiology and Prevention Research
National Institute on Drug Abuse

NIDA Research Monograph 142
1994

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

National Institute on Drug Abuse
5600 Fishers Lane
Rockville, MD 20857

ACKNOWLEDGMENT

This monograph is based on the papers from a technical review on "Advances in Data Analysis for Prevention Intervention Research" held on September 9-10, 1992. The review meeting was sponsored by the National Institute on Drug Abuse.

COPYRIGHT STATUS

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. 94-3599
Printed 1994

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*.

Contents

Page

New Statistical Methods for Substance Use Prevention Research

Linda M. Collins and Larry A. Seitz

Some research on drugs and drug use takes place in the laboratory under well-controlled conditions using simple experimental designs. The data from these studies are analyzed easily using standard statistical procedures; sometimes inferential statistics are not even necessary. In contrast, substance use prevention research, particularly intervention research, generally takes place in the field. Field settings offer the tremendous advantage of ecological validity, but they are associated with some disadvantages as well: Field research designs are, of necessity, more complicated, and the researcher can maintain only so much control. Often, the standard, commonly available statistical procedures fall short when applied in complex field research situations. For example, because these procedures are not well suited for the type of data that have been collected, they do not answer directly the research question of interest, Type I error rates are inflated, or statistical power is low. For these reasons and others, it is extremely important for the field of prevention that researchers keep abreast of the very latest developments in statistical methods.

The ultimate purpose of statistics is to provide a means for drawing conclusions from data. At its best, statistics enjoys a symbiotic relationship with substantive research: The need to answer substantive questions in a particular area inspires the development of new statistical methods, and then the new statistical methods in turn prompt substantive researchers—both inside and outside the area in which the method was originally developed—to see their data in new ways and pose new substantive questions. Yet, statistical methods do not always fulfill their potential for playing an important role in substantive research. In the field of prevention, this often is because statistical methods are not made accessible to substantive researchers. Statistical research is unique among scientific disciplines in that new developments must be shared not only with the statistics community but also with the substantive disciplines, such as prevention, most likely to make use of them. The problem is that, while the former goal of sharing with the statistics community is accomplished by means of publications in statistics journals, there is no well-established mechanism for achieving the latter goal.

It has been the experience of the editors of this monograph that prevention researchers display an openness to, and even eagerness for, new statistical methods that would help them obtain the most from their data. Unfortunately, they have nowhere to turn to learn about the very latest methods. Most prevention researchers, like their colleagues in other areas (including statistics), are not trained to read highly technical presentations outside their own area of research, so they typically do not read statistics journals. Even those prevention researchers who do have the background to read technical presentations of statistical material understandably are willing to invest the considerable time that this requires *only if* there is a high probability that the technique presented will be useful to them. However, the likelihood that a technique will be useful cannot be determined without reading the article, resulting in a frustrating "catch-22."

The editors believe that monographs like this one represent a way to disseminate state-of-the-art statistical procedures to the substance use prevention research community while avoiding the frustration described above. This monograph results from a technical review held by the National Institute on Drug Abuse in Bethesda, MD, on September 9 and 10, 1992. Each of the chapters presents a statistical technique or methodological issue chosen because of its immediate relevance to prevention research. The authors of these chapters all have demonstrated an ability to present technical material in an interesting and accessible manner; many of them are prevention researchers and are familiar with the special concerns of this field.

Readers are likely to find the presentation of the material in this monograph to be somewhat different from other presentations of statistical material. Each chapter in this monograph is accompanied by an abstract that summarizes how the technique presented is useful in prevention research. The chapters are written as nontechnically as is possible without sacrificing rigor, with more technical material set off in italics from the rest of the text so that it can be skipped in a first reading. In this way, the editors hope to encourage prevention researchers to think creatively about the kinds of research questions that can be addressed using these procedures. A chapter's purpose is not to make the reader an expert in a statistical procedure, nor even, in most cases, to equip the reader to carry out an analysis. Rather, each chapter provides sufficient conceptual details to enable the researcher to make an informed decision about whether to pursue further study of the procedure. Most of the chapters point readers toward additional literature to read to help them become familiar enough with a particular procedure to apply it to prevention data.

As the reader will see, the chapters in this monograph constitute a broad and varied assortment of introductions to newly developed techniques, introductions to procedures well established in other disciplines but new to substance use prevention research, and new perspectives on well-established techniques.

INTRODUCTIONS TO NEWLY DEVELOPED TECHNIQUES

Multilevel Analysis

The unit of analysis issue has been a contentious one for years in prevention research. Most substance use prevention research is school based and, thus, the subjects are part of a naturally occurring hierarchy: students are clustered in classes, classes are clustered in schools, and schools are clustered in neighborhoods and/or school districts. The costs of ignoring this hierarchy potentially are great. Individuals clustered together in some way tend to give responses that are related to each other's responses; thus, they are not independently sampled data. However, presence of independently sampled data is an assumption of most statistical procedures. If this assumption is violated, Type I error rates go up, sometimes dramatically. One solution that has been offered to this problem is to perform analyses at the aggregate level, using, for example, classroom or school means as the dependent variable. This method does eliminate the problems caused by a lack of independence among individuals but, for many analyses, this is the only benefit associated with this approach. In most cases in prevention research, the questions are posed at the individual level, such as, "Is there an overall decrease in the amount of alcohol used by individual students? For what kinds of students is the program most effective? What are the characteristics of students who seem to be unaffected by the prevention program?" These kinds of questions cannot be answered by aggregate-level analyses because conclusions based on analyses at, say, the classroom level cannot be generalized to either the individual level or the school level.

Kreft's chapter on multilevel analysis offers an elegant solution to this problem. Kreft shows us that, by using multilevel analysis, we can model all the levels occurring in data simultaneously. This approach even makes it possible to examine the effects of interactions among various levels, for example, interactions between characteristics of the classroom environment and characteristics of the individual. Furthermore, the Type I error rate is controlled by this approach to data analysis. Multilevel analyses require special software, but the user is likely to find the time invested in learning the software very worthwhile.

Missing Data Analysis

Another problem that has dogged prevention research is that of missing data. There are numerous sources of missing data. Probably the most pathological source is subject attrition. Most longitudinal substance use prevention studies experience subject dropout over the course of the study. If subject dropout were completely random, the most serious problem would be a loss of statistical power due to a decreasing N.

However, subject attrition in prevention studies is almost never random. Dropouts tend to be those at higher risk for increased substance use or those who already are using at a higher rate. Thus, the problem becomes one not only of statistical power but also of internal and external validity.

There are widely used procedures for dealing with missing data, primarily listwise deletion, pairwise deletion, and mean replacement. The chapter on missing data analysis by Graham and colleagues discusses each of these procedures and introduces some recently developed alternatives. In some ways, the often-used term "missing data analysis" is a misnomer. Missing data are, well, missing, and so they cannot themselves be analyzed. The techniques reviewed by Graham and colleagues do not create data out of thin air, and they are not a substitute for careful experimental design and assiduous efforts to prevent subject attrition. Rather, they help the researcher make the most out of the data that are present in order to obtain more accurate statistical results.

Meta-Analysis

In substance use prevention, as well as in other fields, it is important to integrate the results of years of research in order to draw policy-relevant conclusions. However, this is more easily said than done. Rarely does a series of research studies speak with one voice; usually there are some conflicting findings. For example, some studies might find that a particular prevention program works well overall, while others find that the program works only moderately well or only for a subset of people.

Meta-analysis is a method of integrating research findings statistically. The chapter by Tobler presents an annotated example of a meta-analysis performed on prevention data. This chapter demonstrates how meta-analysis can be used to make sense out of inconsistencies in findings across studies by examining what characteristics of studies, such as type of sample or whether or not the study is well controlled, can account for the discrepancies. The task of amassing an exhaustive collection of available studies, coding all relevant variables, computing effect sizes, and performing the required analyses is, as Tobler puts it, "not for the faint of heart." However, meta-analysis is the state of the art in research integration, and those who have the courage to undertake a demanding meta-analysis project will find that it is the clearest way to synthesize findings and arrive at valid policy-relevant conclusions.

Dynamic Modeling

In their chapter, Kibel and Holder demonstrate how to break out of the controlled laboratory or field environment and examine the interplay between various kinds of prevention programs and society at large. Using the dynamic modeling technique advanced by Kibel and Holder, the user can build models of the reciprocal effects of societal factors and substance use. One of the important contributions of this approach is as a heuristic. It forces the user to make explicit every assumption about how societal forces work. It also allows the user to try out different models fairly easily. This is another approach that has likely policy relevance as society considers options like restricting the density of liquor stores in neighborhoods or legalization of certain drugs.

INTRODUCTIONS TO PROCEDURES WELL ESTABLISHED IN OTHER DISCIPLINES

Time Series Analysis

Time series analysis, presented in this monograph by Velicer, grew out of econometrics and has been applied successfully in the social sciences for years. The data needed for time series analysis consist of a long string of repeated observations on an individual taken at regular intervals. Thus, time series designs usually are focused on intensive observation of an individual, in contrast to the typical school-based prevention intervention design, which collects data on a large number of individuals at widely spaced intervals.

For example, Velicer collected data on the cigarette smoking behavior of six individuals twice daily for 62 days. Time series analysis is ideal for modeling the routine habits of substance users. It also is possible to evaluate the effectiveness of interventions designed to interfere with these habits by comparing characteristics of a time series before and after an intervention. This is called interrupted time series analysis. Time series analysis has great potential for use in substance use prevention studies, particularly where subject sample size is limited but intensive measurement of subjects is feasible.

Survival Analysis

Singer and Willett present survival analysis, a statistical technique that is familiar in epidemiology but is beginning just now to be adopted by behavioral researchers. Survival analysis rephrases some of the fundamental questions asked by prevention researchers. For example, in a survival analysis, we identify an event of interest—say, onset of substance use—and ask the question, "Is the amount of time until onset for the program children longer than the amount of time until onset for the control children?" Survival analysis produces some useful quantities, such as the survival function. An example of a survival function in prevention research is the proportion of a sample who have not yet begun the onset process expressed as a function of time. Another useful quantity is the hazard function. This function expresses incidence as a function of time; for example, a hazard function would express the probability of onset at a particular time, given that onset has not occurred already. This function expresses risk (hazard) of substance use onset. The hazard function potentially is tremendously useful in substance use prevention intervention research. For example, a thorough knowledge of the hazard function for people in their preadolescent and adolescent years would be a highly useful tool in the timing of prevention intervention activities and booster sessions.

Latent Class and Latent Transition Analyses

In building models of substance use and its prevention, it often makes sense to identify qualitatively distinct groups. For example, there may be certain patterns of use characterized by frequency, duration, and substance or combination of substances. There may be bingers, light steady users, or specializers in a particular substance. Identifying these kinds of subgroups within data could help prevention efforts by pointing toward directions to go and areas to cover in planning interventions. The chapter by Uebersax illustrates how to use latent class analysis (LCA), a procedure that originally was developed in sociology and psychology, to identify these subgroups or *latent classes.* Uebersax also shows that, once the subgroups are identified, further analyses can be performed to look at quantitative differences among the groups. For example, perhaps bingers are more rebellious or have a poorer relationship with their parents than do light steady users.

Another approach to questions involving latent classes is to ask whether membership in latent classes changes over time. Often these latent classes can be thought of as stages in a process that unfolds over time. Collins and colleagues present latent transition analysis, which is a generalization of LCA to longitudinal data. This approach provides a method of testing stage-sequential models of substance use and related processes.

NEW PERSPECTIVES ON WELL-ESTABLISHED TECHNIQUES

Incorporating Trend Data Along With Individual-Level Cross-Sectional Relationships

Figure 1 in the chapter by Bachman presents an interesting graph showing the increase over time in individuals' perceived risk and disapproval of marijuana use, as well as their corresponding decline in marijuana use during the same period (while availability remained constant). The issue raised is one of causality. Three hypotheses are possible: (1) increases in perceived risk and disapproval led to the decline in marijuana use; (2) changes in use led to changes in attitudes; or (3) changes in some other factor or factors caused changes in both use and attitudes. Bachman provides a series of analyses designed to resolve this issue by incorporating trend data along with individual-level, cross-sectional relationships. These analyses are relatively simple, straightforward, and easy to follow.

The resulting conclusion is in favor of the first hypothesis; individual attitudes about specific drugs seem to affect individual use of those drugs.

Repeated Measures Analysis of Variance

Although the designs of substance use prevention intervention studies often are complex, the bottom-line questions about program effects often boil down to a repeated measures analysis of variance (ANOVA) or analysis of covariance. Many researchers learned the basics of this time-honored approach in graduate school, but they may be a little rusty with these procedures, may not appreciate their subtleties, or may not be aware of the most recently raised issues. The chapter by Barcikowski and Robey, who are noted experts on repeated measures, begins with the basics of repeated measures designs and continues through more complicated designs. Included in this chapter is a wealth of information sure to be helpful to prevention researchers, such as how to detect and adjust for violations of the sphericity assumption.

Statistical Power

Statistical power is an issue that many substance use prevention researchers feel they understand well—just obtain the largest N possible, and power will be maximized. The chapter by Hansen and Collins reminds us that there are other factors that go into power besides the number of subjects at the outset of a study. For example, when subjects are lost to attrition over the course of a study, a loss of statistical power can occur. Hansen and Collins also point out that certain aspects of design under the researcher's control have a direct impact on effect size, which is one of the factors determining power. Hansen and Collins discuss two general strategies for increasing effect size: (1) increasing the size of the difference between the treatment group means and any control group means, and (2) decreasing variance. These authors share many useful practical suggestions for increasing statistical power in the context of prevention research.

Some Important Procedures Not Included in This Monograph

Of course, no monograph of this type can be comprehensive. Of the many exciting statistical procedures that potentially can be of much use in prevention research, only a few could be included in this monograph. The prevention researcher interested in methodology may wish to look into some of the procedures mentioned below.

Structural equation modeling is an exciting procedure that has gone from being virtually unknown 20 years ago to being in almost routine use today. This approach has been used extensively to test models of substance use onset and prevention. There are numerous issues in structural equation modeling that are of interest to prevention researchers, such as assessing goodness of fit (Bentler 1990; McDonald and Marsh 1990), and models for multitrait, multimethod applications (Graham and Collins 1992; Marsh and Bailey 1991; Wothke and Browne 1990).

A related topic is growth curve models. This is a general term for methodology that allows the user to develop and test models of individual growth. Such models can be tested in the context of hierarchical linear models (Bryk and Raudenbush 1992) and structural equation models (McArdle and Hamagami 1991; Willett and Sayer, in press).

A notable omission from this monograph is an extensive discussion about measurement of substance use and related variables. Measurement of substance use is a complex and rich topic and easily could fill a monograph alone. Most researchers have been trained in classical test theory and feel most comfortable using factor analysis and evaluating scales using Cronbach's alpha. In recent years, there have been other approaches developed that researchers potentially would be interested in. For example, item response theory is a different perspective on measurement that has been used successfully in many areas outside of standard achievement testing situations (e.g., Wilson 1992). Under certain conditions, item response theory allows the estimation of item parameters that are independent of the exact sample upon which they are based.

Often researchers measure substance use and related variables with categorical variables, which means in most cases that data analysis is going to involve contingency tables. Use of log-linear models (Agresti 1990) is a methodology for analyzing complicated multiway contingency tables using a framework similar to the familiar ANOVA framework. Latent class models, which are discussed in this monograph, are related to log-linear models but involve latent variables.

CONCLUSION

In this monograph, the editors have attempted to assemble a collection of chapters presenting innovative statistical methods to the substance use prevention research community. The chapters are intended to be accessible conceptual and technical introductions to each method rather than complete tutorials. The editors hope that prevention researchers find the monograph useful. The editors also hope that, in the short run, this monograph helps increase the use of innovative statistical procedures in prevention research, and that, in the long run, two-way communication between the fields of statistics and substance use prevention research is established to the benefit of both.

REFERENCES

Agresti, A. *Categorical Data Analysis*. New York: Wiley, 1990.

Bentler, P.M. Comparative fit indexes in structural models. *Psychol Bull* 107:238-246, 1990.

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

Graham, J.W., and Collins, N.L. Controlling correlational bias via confirmatory factor analysis of MTMM data. *Multivariate Behav Res* 26:607-629, 1992.

Marsh, H.W., and Bailey, M. Confirmatory factor analyses of multitrait-multimethod data: A comparison of alternative models. *Appl Psychol Meas* 15:47-70, 1991.

McArdle, J.J., and Hamagami, F. Modeling incomplete longitudinal and cross-sectional data using latent growth structural models. In: Collins, L.M., and Horn, J.L., eds. *Best Methods for the Analysis of Change: Recent Advances, Unanswered Questions, Future Directions.* Washington, DC: American Psychological Association, 1991.

McDonald, R.P., and Marsh, H.W. Choosing a multivariate model: Noncentrality and goodness of fit. *Psychol Bull* 107:247-255, 1990.

Willett, J.B., and Sayer, A.G. Using covariance structure analysis to detect correlates and predictors of individual change over time. *Psychol Bull,* in press.

Wilson, M. *Objective Measurement: Theory Into Practice.* Norwood, NJ: Ablex Publishing, 1992.

Wothke, W., and Browne, M.W. The direct product model for the MTMM matrix parameterized as a second order factor analysis model. *Psychometrika* 55:255-262, 1990.

AUTHORS

Linda M. Collins, Ph.D.
Professor
Department of Human Development
and Family Studies
The Pennsylvania State University
University Park, PA 16802-6504

Larry A. Seitz, Ph.D.
Mathematical Statistician
Prevention Research Branch
Division of Epidemiology
and Prevention Research
National Institute on Drug Abuse
Parklawn Building, Room 9A54
5600 Fishers Lane
Rockville, MD 20857

Analysis With Missing Data in Drug Prevention Research

John W. Graham, Scott M. Hofer, and Andrea M. Piccinin

ABSTRACT

Missing data problems have been a thorn in the side of prevention researchers for years. Although some solutions for these problems have been available in the statistical literature, these solutions have not found their way into mainstream prevention research. This chapter is meant to serve as an introduction to the systematic application of the missing data analysis solutions presented recently by Little and Rubin (1987) and others. The chapter does not describe a complete strategy, but it is relevant for (1) missing data analysis with continuous (but not categorical) data, (2) data that are reasonably normally distributed, and (3) solutions for missing data problems for analyses related to the general linear model, in particular, analyses that use (or can use) a covariance matrix as input. The examples in the chapter come from drug prevention research. The chapter discusses (1) the problem of wanting to ask respondents more questions than most individuals can answer; (2) the problem of attrition and some solutions; and (3) the problem of special measurement procedures that are too expensive or time consuming to obtain for all subjects.

The authors end with several conclusions:

- Whenever possible, researchers should use the Expectation-Maximization (EM) algorithm (or other maximum likelihood procedure, including the multiple-group structural equation-modeling procedure, or, where appropriate, multiple imputation, for analyses involving missing data [the chapter provides concrete examples]);

- If researchers must use other analyses, they should keep in mind that these others produce biased results and should not be relied upon for final analyses;

- When data are missing, the appropriate missing data analysis procedures do *not* generate something out of nothing but *do* make the most out of the data available;

- When data are missing, researchers should work hard (especially when planning a study) to find the cause of missingness and include the cause in the analysis models; and

- Researchers should sample the cases originally missing (whenever possible) and adjust EM algorithm parameter estimates accordingly.

INTRODUCTION

Missing data problems have been a thorn in the side of prevention researchers for years. Although some solutions for these problems have been available in the statistical literature for some time now, consumers of statistical procedures, in general, and prevention researchers, in particular, generally have not reaped the benefits of these solutions. In large part, drug prevention analyses have dealt with missing data problems in a piecemeal fashion. A systematic solution to missing data problems, which prevention work to date has lacked, has been viewed as something that was at the very top of the second page of the priority list.

This chapter is meant to serve as an introduction to the systematic application of the missing data analysis solutions presented recently by Little and Rubin (1987) and others. The chapter does not describe a complete strategy, but it is relevant for:

- Missing data analysis with continuous (but not categorical) data;

- Data that are normally distributed, or at least close enough to normally distributed that most critics would not complain too much about it; and

- Solutions for missing data problems for analyses related to the general linear model, in particular, analyses that use (or can use) a covariance matrix as input.

The chapter will deal with three missing data situations. The first is *omissions*. The second is the problem of participant *attrition*. The third is *planned missing data*, that is, data that are missing as a result of the measurement strategy. In general, the discussion of these issues will be conceptual and practical, rather than mathematical (see Little and Rubin [1987] for mathematical treatments of these issues). Finally, the examples in this chapter come from drug prevention research, and most of the points are made in this context. However, most of the points have relevance in other research domains as well.

Before discussing the various forms of missing data, consider the philosophy of missing data analysis. Analysis of data with missing values is thought of more appropriately as a set of procedures for analyzing the data one has, rather than for generating the data one does not have. The missing data analysis procedures recommended here are reminiscent of pairwise deletion (or pairwise inclusion) in the sense that they allow full use of the available data, thereby allowing the most statistically powerful analysis possible. The procedures recommended, however, provide additional benefits that far exceed those of pairwise deletion.

OMISSIONS

Omissions are defined as missing data that occur within an otherwise complete survey. In discussing omissions, a distinction is drawn between those that occur somewhere in the middle of the survey and those that occur at the end. Various causes of missingness in both cases also are discussed.

Internal Omissions

Internal omissions occur for various reasons. A subject simply may not see a question. He or she may want to think about a question before answering and simply forget to go back to the skipped question. A subject may have trouble understanding the meaning of a question and may skip it. Finally, a subject may not answer a particular question because he or she is afraid of possible negative consequences of answering it or because the question evokes negative feelings he or she does not want to experience.

Failure To Complete the Survey

This type of omission simply means that the subject began the survey, completed it up to a point, and then stopped responding. Assuming that many subjects do finish the survey, the two main reasons for the failure to complete it are lack of ability and lack of motivation. A subject may lack the ability to finish because he or she is a slow reader or because the survey is in English and the subject is not a native English speaker. A subject may lack the motivation due to general rebelliousness or because he or she feels it is appropriate to make a minimal effort.

ATTRITION

Attrition occurs when a subject is present for the intervention and for at least one wave of measurement but is absent entirely for one or more other waves of measurement. Various patterns of attrition are possible, and each may possess unique problems and solutions. Consider the example shown in table 1: An intervention is completed for seventh graders, a pretest measure is taken at seventh grade, and posttest measures are taken on the same subjects at the eighth and ninth grades.

TABLE 1. *Patterns of attrition*

Is subject present for ___?

Attrition pattern	0_7	X_7	0_8	0_9
1	YES	YES	YES	YES
2	YES	YES	YES	no
3	YES	YES	no	YES
4	YES	YES	no	no
5	no	YES	YES	YES

Some of the patterns shown in table 1 may be more of a problem than others. For example, patterns 2 and 4 have in common the fact that the subject leaves the measurement part of the research and is never heard from again. This could be a problem in that the subject may have dropped out of the study for reasons having to do with the main dependent variable (i.e., drug use). With attrition patterns 3 and 5, this is less of a concern in that later drug use may be used as a reasonable proxy for earlier drug use.

Causes of Attrition

Researchers would like to think that the kind of attrition shown in table 1 is caused by a random process. As discussed in a later section, and as many researchers believe at an intuitive level, data that are missing completely at random (i.e., the cause of missingness is a random process) are a minor nuisance compared to data that are missing for nonrandom reasons. Unfortunately, the cause of attrition probably is never a purely random process.

There are numerous nonrandom causes of attrition that are completely unrelated to the measurement: The subject is ill for the measurement session; the subject cuts several classes, and this happens to be one of them; the subject drops out of school to earn money for the family or to take care of a family member; the subject is suspended from school (e.g., for fighting); the subject's parents move away to take a new job in another city; or the subject's parents move around a lot for other reasons.

There also are several nonrandom causes of attrition that are directly related to the measurement: The student refuses to participate because of general rebelliousness; the student refuses to participate due to difficulty with the survey (e.g., he or she is a poor reader); the parents actively withhold permission to participate due to concerns about invasion of privacy; the parents passively fail to give permission due to procrastination; or the parents passively fail to give permission because they do not care about what their child does.

Finally, there could be a nonrandom cause of attrition that is directly related to scores on the dependent variable itself. For example, students who use drugs may be more likely to drop out of the study than are students who do not use drugs. Fortunately, drug use may be a rather distal cause of attrition, and some other variable (e.g., dismissal from school) may be the more proximal cause. If this is the case, it may be possible to find and measure the more proximal cause even though the drug use measure is not available because of attrition.

Differential Attrition

Differential attrition has been thought to be one of the most serious threats to the validity of intervention programs. Two definitions of differential attrition are:

- People who drop out of the study have greater drug use at the posttest than do those who stay, AND more people attrit from the program group than from the control group; and

- Program by attrition status interaction with posttest drug use as the dependent variable.

Note that, in both definitions, it is posttest drug use that is relevant. Unfortunately, when researchers have missing data for the posttest measure of drug use, they never can be certain whether there is differential attrition or not. Procedures have been suggested for testing for differential attrition that involve using the *pretest* measure of drug use as a proxy for posttest drug use (e.g., Biglan et al. 1987; Hansen et al. 1985). However, even when the correlation between pretest and posttest drug use is substantial (e.g., r = .60), pretest use may be a poor proxy for posttest use. Although the jury is still out on these procedures, recent work has suggested that the Biglan and colleagues (1987) and Hansen and colleagues (1985) procedures may be useful in most cases *if they show no differential attrition* using pretest drug use as a proxy (Graham and Donaldson 1993). However, the procedures often may be misleading when they suggest that there *is* differential attrition.

A study is described below showing that differential attrition is a serious problem only when the cause of missingness is the posttest drug use variable itself. When differential attrition is caused by some variable other than the dependent variable, and when that variable is included properly in the model, there is no bias due to attrition. This can be true even when traditional complete cases analyses are performed.

PLANNED MISSING DATA

One of the most important features of planned missing data is that researchers know what caused the missingness—they caused it. If researchers assign subjects randomly to the various measurement conditions, then they know that the cause of missingness is a random process. The advantage of doing this will be discussed in a later section.

The Three-Form Design

One of the biggest problems facing drug prevention researchers is that there simply are too many questions to ask and not enough time to ask them. Models of prevention and prevention effectiveness necessarily are complex (e.g., Flay and Petraitis 1991) and require the measurement of many behavioral and psychosocial constructs. However, in many populations (especially adolescent populations), there simply is not enough time to ask all of the relevant questions. Thus, researchers devise various measurement plans to maximize the total number of questions asked while maintaining a manageable number of questions for any individual.

One such measurement plan is the three-form design, which is depicted in table 2 (Graham et al., submitted). Suppose a research team wants to collect questionnaire data on adolescents in their area. They would like to ask 130 questions, but the children will complete only about 100. With the three-form design, each child receives only 100 items, but 130 questions still are asked overall.

TABLE 2. *Three-form design*

	Answered question set?			
	X	A	B	C
Form 1	YES	YES	YES	no
Form 2	YES	YES	no	YES
Form 3	YES	no	YES	YES

There are two main advantages of the three-form design. First, one can ask approximately 33 percent more questions overall while keeping reasonable the number answered by any individual. Second, although no subject has complete data for item sets X, A, B, and C (as shown in table 2), at least one-third of the subjects respond to each pair of items. Thus, good estimates of covariances can be obtained for all item pairs.

Special Measurement Procedures

Another type of planned missingness has to do with special measurement procedures. For example, in the Adolescent Alcohol Prevention Trial (AAPT), Graham and colleagues (1989), Hansen and Graham (1991), Hansen and colleagues (1988, 1991), and Rohrbach and colleagues (1987) sought to measure the variables hypothesized to mediate prevention program effectiveness (see figure 1). One of these key mediating variables was the resistance skill of subjects receiving various prevention curricula (including a resistance skills-training curriculum).

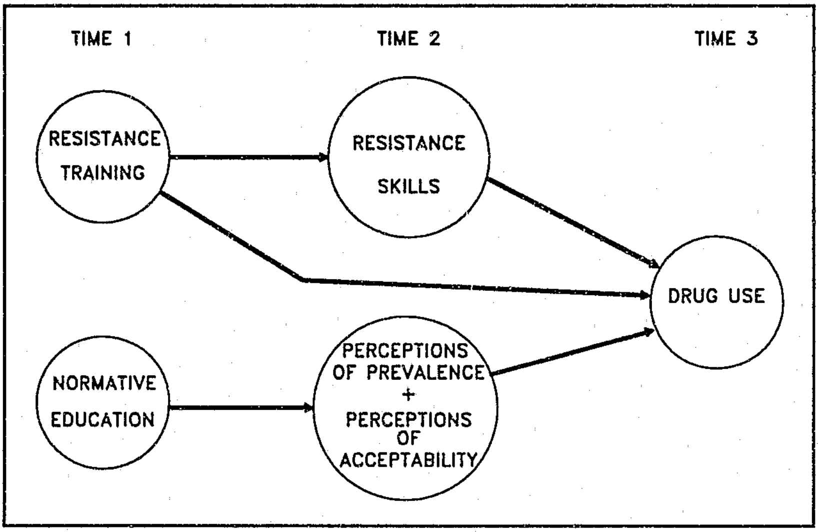


FIGURE 1. *Process model*

However, because the measurement procedure was rather extensive and involved pulling subjects out of class individually, only a random one-third sample of the subjects could receive the skills assessment. Drug use and other related measures were collected for the full sample.

Test of the Interaction: ProgramxGrade of Intervention

One of the key questions for prevention researchers is: "What is the best grade for an intervention?" As a means of answering this question, the AAPT project was implemented fully at the fifth and seventh grades as shown in table 3. Hypotheses regarding grade of intervention could be tested easily with a posttest-only analysis (e.g., by treating posttest drug use, say eighth-grade drug use, as the dependent variable and program, grade, and the programxgrade interaction as the independent variables.

TABLE 3. *Analysis by grade*

Panel	Grade of program	Grade 5 data?	Grade 8 data?
1	5	yes	yes
2	7	yes	yes
3	5	yes	yes
4	7	NO	yes

However, a stronger test of hypotheses involving grade of intervention would include pretest drug use as a covariate. In the AAPT study, because some subjects received the program as fifth graders, it would seem that fifth-grade drug use would be the most appropriate covariate. However, most subjects receiving the program in the seventh grade were not pretested until the seventh grade and had no data for fifth-grade drug use. Thus, if the analyses of covariance (ANCOVAs) were conducted based only on students with complete data, no seventh graders from panel 4 would be involved, and the test of the key interaction would not be possible without significant loss of power.

CAUSES OF MISSINGNESS REVISITED

Many causes of missingness have been suggested in the examples given above. When the cause of missingness is a random process, the problems arising from the missing data are relatively minor and are mainly a matter of statistical power. However, when the cause of missingness is not a random process, the problems are more complex. Two general kinds of nonrandom missing data mechanisms are discussed below: accessible and inaccessible. Ways in which most causes of missing data can be made accessible also will be discussed.

Accessible Missing Data Mechanisms

The missing data mechanism is accessible when the cause of missingness has been measured and is available for use in the analysis (Graham and Donaldson 1993). Although one never can know for sure whether the mechanism is accessible, it is important to know that accessible, nonrandom mechanisms cause no bias *when the cause of missingness is included properly in the analysis.* As discussed more fully below, analyses that properly take the cause of missingness into account include: (1) use of the Expectation-Maximization (EM) algorithm; (2) other maximum likelihood procedures (e.g., the multiple-group structural equation-modeling procedures described by Allison [1987] and Muthen and colleagues [1987]); and (3) ANCOVA with complete data in certain situations.

The term "accessible" is related to the term "ignorable" as used by Little and Rubin (1987), except that the term "accessible" refers to the mechanism per se, whereas the term "ignorable" refers to a combination of the mechanism and the analysis used. For example, even when the cause of missingness has been measured, the mechanism is not ignorable if the cause is not used properly in the analysis. The term "accessible" emphasizes the importance of measuring the causes of missingness.

Inaccessible Missing Data Mechanisms

The missing data mechanism is inaccessible when the cause of missingness has not been measured or otherwise is unavailable for analysis (Graham and Donaldson 1993). This is similar to Little and Rubin's (1987) term "nonignorable." Again, however, the term "inaccessible" refers to the mechanism itself, whereas Little and Rubin's term refers to a combination of the mechanism and the analysis used.

Inaccessible missing data mechanisms arise when the variable containing the missing data itself is the cause of missingness. For example, the mechanism would be inaccessible if the people who drop out of a drug use prevention study do so because they currently are high-level drug users.

Inaccessible mechanisms also can arise if another unmeasured variable is the cause of missingness and that variable is correlated with the one containing the missing data (e.g., posttest drug use). If the cause of missingness is unrelated to the variable with missing data, then the cause essentially is a random process with respect to the variable containing missing data. (Keep in mind the fact that a variable can be correlated with *missingness* on the posttest drug use variable and still can be uncorrelated with posttest drug-use itself.) For example, general transiency may be related to attrition *and* may be correlated with drug use. On the other hand, a parent being transferred to another job will be related to attrition but may not be correlated with drug use.

When the cause of missingness is inaccessible, there may or may not be bias in the estimation of key parameters. For example, a recent study (Graham and Donaldson 1993) showed that estimates of program effects were substantially biased if there was differential attrition on the main dependent variable and if that variable was the cause of missingness. The study also showed that, even in the presence of substantial attrition (caused by the dependent variable itself), the estimates of program effects were unbiased if there was no differential attrition on the main dependent variable.

How Can One Know if the Mechanism Is Accessible?

Given the importance of being able to distinguish between accessible and inaccessible missing data mechanisms, the natural question that arises is: "How can one know if the mechanism is accessible or inaccessible?" The answer, unfortunately, is that one cannot know. At least, one cannot know about the mechanism if one collects no new data. However, there may be several courses of action researchers can take.

One can assume that the mechanism is a random process (i.e., data are missing completely at random). Although it never may be reasonable to assume that data missing due to omissions or attrition are missing completely at random, it often may be reasonable to assume that the cause is a random process with respect to the dependent variable.

One can assume that the mechanism is accessible. Following Heckman (1979), Dent (1988) described a procedure to determine how much of the cause of missingness had been measured (also see Graham and Donaldson 1993; Leigh et al. 1993). The procedure involves creating a missingness dummy variable with the value of 1 if the variable of interest was nonmissing and the value of 0 if the variable was missing. This missingness variable then would be regressed on all other variables in the data set. The linear combination of all other variables could be thought of as a single variable representing the known causes of missingness and could be included in all other analyses. In this way, biases from measured causes of missingness would be controlled.

The main problem with this approach is that one still does not know how much of the measurable cause of missingness has been measured. In general, there are three possible causes of missingness: (1) measured variables correlated with the variable containing the missing data, (2) unmeasured variables correlated with the variable containing the missing data, and (3) variables that essentially are a random process with respect to the variable containing the missing data. Suppose one discovers that the first type of cause (measured variables correlated with variable of interest) accounts for 20 percent of the variance in the missingness dummy variable. Although this is a rather substantial amount, it still is not known whether the remaining 80 percent of the causes are of the second or third type (unmeasured and correlated, or random processes). In this situation, researchers must resort to making assumptions about the causes of missingness.

Collect Additional Data. The best way to get around the problem of not knowing about the mechanism of missingness is to collect additional data from those with initially missing data (Graham and Donaldson 1993; Little and Rubin 1987; Rubin 1987). If one can obtain measures from a random sample of the cases originally missing, one has sampled and measured all causes of missingness. That is, the causes of missingness then are accessible. If used properly in the analyses, this addition of cases controls completely for all missing data biases. Using these data properly in the analysis will be discussed further in the **Analysis Possibilities** section, EM Algorithm subsection.

Most Causes of Missingness Are Measurable

Short of collecting additional data, one never can be certain about the causes of missingness. Nonetheless, the better a researcher is able to account for missingness, the stronger his or her argument that the important causes of missingness have been measured and taken into account is. In most cases, the cause of missingness should be measurable.

Table 4 presents a set of possible measures for some of the major causes of missingness discussed in this chapter. This is not meant to be an exhaustive list, but it does provide a starting place for thinking about measuring these important variables.

TABLE 4. *Possible measures of causes of missingness*

Cause: Subject is a slow reader

Possible Measures:

- *Standardized test scores from school records, especially reading scores*
- *What language do you usually speak at home?*
- *What language do you usually speak with your friends?*
- *Grades*

Cause: Subject lacks motivation to complete survey

Possible Measures:

- *Measures of general motivaton*
- *Measures of motivation to complete the questionnaire*

Cause: Subject is rebellious

Possible Measures:

- *Measures of rebelliousness*

Cause: Parents move away/transiency

Possible Measures:

- *How many schools have you attended since first grade?*
- *How many times have you moved in the past 5 years?*
- *How likely is it that next year, you will be in this school, or in the next higher school in this school system?*

Cause: Parents actively fail to give permission (are political activists, fear invasion of privacy, etc.)

Possible Measures:

- *How bad is invasion of privacy?*
- *How bad do your parents think invasion of privacy is?*
- *Possible to get classroom teacher to ask questions such as these in general classroom context (i.e., even those without permission may respond)*

Cause: Parents passively fail to give permission (are procrastinators, couldn't care less about what their kids do or don't do, etc.)

Possible Measures:

- *How much do your parents care what you do?*
- *Possible to get classroom teacher to ask questions such as these in general classroom context*

Cause: Child refuses to participate because of scores on the dependent variable

Possible Measures:

- *How would your best friends react if you used drugs?*
- *How would your parents react if they found out you used drugs?*
- *If you used drugs, and you said so on this questionnaire, how likely is it that your friends would find out?*
- *If you used drugs, and you said so on this questionnaire, how likely is it that your parents would find out?*

ANALYSIS POSSIBILITIES

In this section, several possibilities for analysis with missing data will be discussed. Some of the procedures employed in the past, as well as procedures that have emerged more recently, will be explored. A strong stand is taken in this chapter on what should and should not be used for analysis with missing data.

Mean Substitution

One of the most common forms of analysis with missing data involves simply substituting the mean for the variable whenever a value is missing. As illustrated in an example below, mean substitution can produce very wrong estimates of variances and covariances. In general, substituting the mean for the missing value has the effect of underestimating the magnitude of both variances and covariances.

In short, mean substitution should *never* be used. Other procedures to be described below are as easy, or easier, to use and are far more defensible.

Complete Cases Analysis

The advantage of analyzing only those cases with complete data is that it is easy to do. For many procedures, analysis of complete cases (i.e., listwise deletion) is the default option. If the cause of missingness is a random process, there are no biases in such analyses. Under some circumstances, there may be no biases even if the cause of missingness is nonrandom, provided the nonrandom cause is accessible. For example, consider a simple program evaluation ANCOVA model with program and pretest drug use predicting posttest drug use. If there are missing data only for the posttest drug use measure and, if the cause of missingness is pretest drug use, then the complete cases analysis is unbiased for estimating the program effect (Graham and Donaldson 1993). However, if the cause of missingness is nonrandom and unmeasured (i.e., inaccessible), serious bias can occur with complete cases analysis.

The greatest drawback with the complete cases analysis is loss of statistical power. If the amount of missing data is substantial, one may have to discard much data in order to have cases with complete data. In research designs that call for planned missingness, for example, the three-form design (see table 2), one simply cannot perform analyses involving all sets of variables. In other designs (e.g., the process model in figure 1), one may have to discard a large amount of relevant data to obtain complete cases. Finally, for some designs, analysis with complete data would produce a serious imbalance in the data and would make important analyses impossible (e.g., see the analysis of grade by program interaction shown in table 3).

Pairwise Deletion

The main advantage of analyzing by pairwise deletion (or pairwise inclusion) is that one makes use of all the available data. Also, if the cause of missingness is a random process, then analysis by pairwise deletion produces unbiased estimates of each correlation. For example, Graham and colleagues (submitted) have shown that pairwise deletion provides unbiased estimation for analysis of the three-form design when the only cause of missingness is a random process.[1]

Although pairwise deletion may produce pairwise unbiased estimates of covariances, there is no guarantee that the estimates will be matrixwise unbiased. In other words, there is no guarantee that the resulting matrix will be positive-definite, and if it is not, some analyses will not be possible. Furthermore, if the cause of missingness is nonrandom, pairwise deletion does not provide protection from bias, even if the cause of missingness is included in the model. This point is illustrated in a later section with an example of analyses with attrition.

Regression-Based Single Imputation

Another alternative for dealing with missing data is regression-based single impuıation. In this case, the variable containing missing data is predicted by all other relevant variables to be used in the final model.

The regression equation obtained for cases with data present is used to predict the variable for cases with missing values. The predicted scores are substituted (i.e., imputed), and analyses are conducted as if there were no missing data.[2]

Although there is a good rationale for doing such analyses, there are some drawbacks. First, the regression-based imputation procedure has a statistical basis only with certain patterns of missingness called *monotone* missing data patterns (Little and Rubin 1987). The missing data pattern is monotone when the variables and cases can be organized in a way similar to that depicted in table 5. That is, for every subject, if a variable has a nonmissing value, then all variables to the left also have nonmissing values. Also, for every subject, if a variable has a missing value, then all variables to the right also have missing values.

TABLE 5. *Monotone missing data pattern*

	Variable			
Case	A	B	C	D
1	1	1	1	1
2	1	1	1	1
3	1	1	1	0
4	1	1	1	0
5	1	1	0	0
6	1	1	0	0
7	1	0	0	0
8	1	0	0	0

KEY: 0 = Missing
1 = Nonmissing

For missing data patterns that do not conform to the monotone pattern, one can discard data to achieve the monotone pattern, but this can result in a substantial loss of statistical power.

A second disadvantage of regression-based single imputation is that the resulting variance estimates are negatively biased (i.e., smaller than they should be). Covariance estimates also are negatively biased when variables are missing jointly. The problem of negatively biased variance estimates can be understood as described below.

Suppose a regression equation is used to predict scores for data that are *non*missing. Everyone knows that the regression equation does not predict these known scores perfectly. Rather, each score is predicted with some amount of error. That is, there is a component of variability in the known scores that goes beyond the variability accounted for by the regression equation.

So, why should any regression equation be expected to predict the *missing* scores without error? In fact, this is the most serious problem with single imputation: The missing scores are predicted without error. That is, the component of variability (due to random error) is missing. Thus, the total variability of scores is less than would be expected if they were nonmissing. This point is explored further in the next section.

Multiple Imputation

There are two key parts to multiple imputation as described by Rubin (1987): restoring error to the singly imputed values and performing the error restoration multiple times. One way the error restoration could be done follows. Suppose researchers have a situation with three variables, X_1, X_2, and Y. Further suppose that only Y contains any missing data. For those subjects who have no missing data, the regression equation is:

$$\mathbf{Y = b_0+b_1X_1+b_2X_2+e} \tag{1}$$

The degree to which the regression equation is not perfect in predicting the nonmissing Y scores is described by the distribution of error terms, e. It often is reasonable to assume that the distribution of error terms is about the same for both the cases with and without missing data for Y. Thus, the appropriate variability can be restored to the prediction of Y by adding a randomly selected element of the distribution of e to the singly imputed Y score.

There really is nothing multiple about this correction. The "multiple" in multiple imputation comes from performing the sampling and addition of error multiple times. Although the imputed scores (with the error added) are expected to provide unbiased estimates of variances even if performed just once, more precise estimates of the variances can be obtained by performing the random draws multiple times. Rubin (1987) recommends creating multiple full data sets, each with a different set of random draws. He suggests that even two sets of random draws provide substantial improvement in estimation.

The main disadvantage with multiple imputation is that it is a bulky procedure. In order to do the procedure, one must: (1) perform the basic single imputation, (2) generate a distribution of residuals, (3) perform the random selection of error terms (with replacement), (4) create a new data set, adding errors to the singly imputed scores, and (5) perform the analysis of interest. For *multiple* imputation, one repeats steps 3-5 the desired number of times. One must then (6) average the key parameter estimates over the number of imputation steps.

Another limitation is that, if the multiple imputation procedure is based on regression-based single imputation, a monotone missing data pattern still is required. Alternatively, one could perform the single imputation step (step 1, above) based on the EM algorithm (see next section). With this approach, one would not be limited to monotone missing data patterns. However, once a maximum likelihood estimate of the variance-covariance matrix is obtained based on the EM algorithm, adding the bulk of the multiple imputation procedure seems unnecessary.

There certainly are situations in which the multiple imputation procedure is superior to an EM algorithm designed to produce a covariance matrix. One example is the analysis of difference scores. Still, when the analysis to be done is based on a covariance matrix (or means and covariance matrix), use of the EM algorithm to produce maximum likelihood estimates of the covariance matrix seems preferable.

EM Algorithm

The EM algorithm (Dempster et al. 1977; Little and Rubin 1987) achieves much the same result as multiple imputation in that it adds an error component to the imputed values. With the EM algorithm, however, the error is added to the sums of squares and cross-products rather than directly to an imputed score. In this section, the operation of the EM algorithm for the continuous variable case with covariance matrix as output is described briefly. It is important to note different EM algorithms are required for different kinds of analysis. However, because so many common analyses can be performed with the covariance matrix as input (e.g., anything involving the general linear model), this particular version of the EM algorithm can be extensively useful.

For the Expectation (E) step of the EM algorithm, sums of squares and sums of cross-products are collected. If the score for a particular variable is present, the algorithm collects sums in the usual way. If the score is missing, the algorithm uses the best estimate of the score (i.e., the singly imputed value based on a regression involving all other variables).

Collection of sums of squares and sums of cross-products is straightforward if neither variable is missing or if just one variable is missing. If the score is present for both of the two variables involved, sums of squares and sums of cross-products are collected in the usual way. If one of the two values is missing, sums of squares and sums of cross-products are collected in the usual way except that the missing value is replaced by the singly imputed score using all other variables as predictors.

Finally, if both values are missing, the sums of squares and sums of cross-products are collected in the usual way except that the scores are replaced by the singly imputed scores, *and* a correction term is added. For sums of squares, the correction term is the *residual variance* of the variable after being predicted by all other variables in the model. For sums of cross-products, the correction is the residual covariance between the two variables after being predicted by all other variables in the model. This concludes the E step.

The Maximization (M) step is very straightforward in this case. Based on the estimates of sums of squares and sums of cross-products, one calculates the means and covariance matrix.

The EM algorithm is an iterative procedure: The covariance matrix generated at one iteration is used to generate b-weights, and the E step (collecting sums of squares and sums of cross-products) is repeated using the revised b-weights for prediction of missing values. The iterative process continues until the changes in the covariance matrix from one iteration to the next are deemed trivially small.

One clear advantage of the EM algorithm is that it handles virtually any pattern of missing data (i.e., it is not restricted to monotone patterns of missingness). Second, this version of the EM algorithm produces maximum likelihood estimates of the means and the variance-covariance matrix. Third, the procedure is available in its general form in BMDP AM (Dixon 1988; Frane 1988).[3,4]

In practical terms, advantages of the EM algorithm are that (1) all parameter estimates are unbiased[5] and more efficient than other methods of estimation (e.g., pairwise deletion);[6] (2) the covariance matrix is positive-definite (i.e., usable for all analyses requiring a covariance matrix as input); and (3) it makes full use of all available data.

Disadvantages of the general implementation found in BMDP AM include: (1) standard errors are not readily available; (2) even if they were, one generally needs the standard errors for the analysis based on the covariance matrix, not for the variances and covariances themselves; and (3) there is no method within BMDP AM to modify the results for inaccessible (nonignorable) missing data mechanisms.

Hybrid Version of EM Algorithm: EMCOV.EXE

It is possible to write a hybrid version of the EM algorithm. One such program is EMCOV.EXE (Graham and Hofer, unpublished manuscript). The advantage of this program is its flexibility. For example, the program can be modified easily to adjust the EM algorithm to account for inaccessible missing data mechanisms (for a brief discussion, see the next section; for additional details, see Graham and Donaldson [1993]).

It also is possible to revise the program for special missing data problems. For example, Graham and Hofer (1992) have revised the program to handle missing data problems involving interactions. If the variables making up the interaction have missing data, most procedures must throw away data unless both variables are nonmissing. For some missing data designs (the three-form design, for example), this could mean a substantial loss of data and statistical power. With the hybrid EM algorithm program, however, Graham and Hofer (1992) were able to make use of all the available data and to obtain estimates of interaction terms with smaller standard errors.

The disadvantage of this and other similar hybrid programs is that they are not readily available. However, such programs are becoming more available. The EMCOV.EXE program (Graham and Hofer, unpublished manuscript) is available from the authors as a beta-test program. The current version provides the correct solution for all situations (i.e., any number of variables missing for each subject). An MS-DOS compatible 486 computer with math coprocessor is recommended. The program is FORTRAN compiled with a DOS extender and can handle any number of cases, variables, and missing data patterns, provided one's computer has sufficient memory. Four MB RAM may be sufficient for smaller problems (in the neighborhood of 20-60 variables with N = 1,000), but

8 MB RAM or more may be requıred for larger problems. Results obtained are the same as those obtained from the BMDP AM program.

Adjusting the EM Algorithm Estimates for Inaccessible Missing Data Mechanisms. The adjustment to the EM algorithm is applicable in a theoretical sense to any missing data problems. However, in practical terms, it is best applied to the case of attrition where relatively few variables have missing data. The example presented here examines the case in which there are three variables—a program variable (Program), pretest covariate (X), and posttest dependent variable (Y), with data missing only for the dependent variable.

The correction begins with the collection of data from a *random sample* of cases with previously missing data on the dependent variable. There are three relevant samples. Sample 1 is the sample of cases having complete data at the outset. Sample 2 is the small random sample of those with initially missing data. Sample 3 is the sample of those for whom posttest data are still missing. The main idea is that the data from sample 2 will be used to extrapolate to those in sample 3. The usual EM algorithm would make use of all nonmissing data to extrapolate to those in sample 3.

Although this correction is conceptually simple, it is computationally complicated. If multiple imputations were being performed, it would be a simple matter to use the sample 2 data to impute values for those in sample 3. However, because the EM algorithm computes the covariance matrix directly for the entire sample, a simpler computational solution must be found.

The computational solution is based on the prediction of scores in sample 2 using the regression equation from sample 1. As described in Graham and Donaldson (1993), the first equation is:

$$\hat{Y}'' = b_0 + b_1 Program + b_2 X \qquad (2)$$

where b_0, b_1, and b_2 are the intercept and b-weights derived in sample 1. These predicted scores ($\hat{Y}''$) then are compared to the actual scores in sample 2. Without going into detail here, the correction to be applied to the EM algorithm comes from the regression of the actual scores in sample 2 on these predicted scores:

$$\hat{Y}*=b_0*+b_1*\hat{Y}''$$

where $\hat{Y}*$ is the estimated score in sample 2, $\hat{Y}''$is the predicted score based on the regression equation from sample 1, and b_0* and b_1* are the intercept and regression weight from that regression analysis.

At the point in the EM algorithm where one must use the best guess of the missing value, one estimates the value in the usual way but adjusts the estimate by multiplying by b_1* and adding b_0*.

A General Solution for Estimating Standard Errors: Bootstrapping

A general solution for the problem of estimating standard errors is *bootstrapping* (Efron 1982). Bootstrapping begins with the assumption that the data sample is a random sample of the population. If this is true, then a random sample of cases from the original sample (with replacement) also is a random sample of the population. Furthermore, the standard deviation for any given parameter estimate across several such new samples is an estimate of the standard error for that parameter estimate.

The bootstrap procedure is outlined as follows:

1. Estimate the variance/covariance matrix using the EM algorithm (e.g., BMDP AM or EMCOV.EXE).

2. Use some statistical package (e.g., LISREL, SAS) to perform the analysis of ultimate interest based on the EM covariance matrix.

3. Do the following 50 times (or 20-1,000 times depending on precision required for hypothesis-testing):
   a. Sample cases with replacement from the original data set to obtain a new data set with the same N as the original[7],
   b. Obtain the EM algorithm estimated covariance matrix (e.g., BMDP AM or EMCOV.EXE),
   c. Analyze covariance matrix (with LISREL, SAS, etc.) to obtain parameter estimates of interest, and
   d. Save parameter estimates.

4. The standard deviation obtained for each parameter estimate over the 50 data sets is an estimate of the standard error for that parameter estimate.

The DOS, BASIC, LISREL, and EMCOV.EXE (and BMDP AM) code necessary to perform a simple bootstrap can be obtained from John Graham.

Multiple-Group Structural Equation-Modeling Procedure

An alternative to the EM algorithm is the use of multiple-group structural equation-modeling analyses. These analyses have been outlined recently by Allison (1987), Jöreskog and Sörbom (1989), Muthen and colleagues (1987), and others. When the data are missing completely at random, or when the cause of missingness has been measured and is included in the model, this procedure gives maximum likelihood estimation for most models that can be estimated in LISREL or comparable programs.

The procedure makes use of the multiple-group capabilities of LISREL (or comparable programs). One divides the data into groups corresponding to each distinct missing data pattern and creates a covariance matrix and vector of means for each group. For groups with missing data, the input covariances and means involving a missing variable are set to 0, and input variances are set to 1.

The basic idea of the procedure is that parameters are estimated based on all data that are available for that parameter. All latent-variable variances, covariances, and regressions are constrained to be equal across groups. If the relevant variable is nonmissing, then factor loadings and residuals are estimated and constrained to be equal across groups. If the relevant variable is missing for a particular group, then all factor loadings and residual variance and covariances corresponding to that variable are not estimated in that group; factor loadings and residual covariances are fixed at 0 and residual variances are fixed at 1. The control statement for running a simple LISREL VI or VII program can be obtained from John Graham.

For models based on manifest variables only, this procedure gives results that are the same as those given by the EM algorithm (EMCOV.EXE or BMDP AM). For latent-variable models, the results from this procedure and the EM algorithm are very similar (both unbiased) but, as might be expected, the estimates based on the multiple-group procedure are very slightly more efficient (i.e., have lower standard errors).

Two clear advantages of using this procedure over use of the EM algorithm are (1) that one can analyze directly the model of ultimate interest, and (2) that, as a byproduct of the analysis, correct standard errors routinely are obtained for the model of ultimate interest.

One disadvantage of the multiple-group procedure is that it can be extremely tedious. One look at the LISREL control statements shows that this is not a procedure for the faint of heart. In fact, the procedure may be useful only for those with considerable LISREL experience.

A second disadvantage is that there may be a practical upper limit to the number of different patterns that can be analyzed. For example, the already bulky procedure becomes unwieldy when the number of different patterns or groups gets larger than four or five (however, such analyses have been conducted with as many as 24 groups, and others have reported using the procedure with even more groups). Also, there is a lower limit to the number of cases present for each pattern: There must be more cases within each group than there are variables. One result of these two problems is that data often must be discarded when using this procedure in order to meet the sample size requirements. Although the amount of data to be discarded generally is small, it could be a deciding factor in choosing this procedure.

There also are some limitations in the kinds of missing data patterns that can be handled by this procedure. For example, for the analysis of the program by grade interaction presented in table 3, the group containing missing data on fifth-grade drug use had no variability for any variables relating to grade of intervention, including the key programxgrade interaction. Because all variables relating to grade of intervention were defined only in the total sample, the multiple-group procedure did not work, whereas the EM algorithm worked well.

Finally, in the multiple-group procedure, there is no way to adjust for inaccessible (nonignorable) missing data mechanisms.

EXAMPLE ANALYSES

Analysis of Three-Form Design

The first example, taken from Graham and colleagues (submitted), is a simulation involving analysis of the three-form design. For this example, there were two simulated variables with no missing data (drug use 1 and drug use 2) and three others simulating data from the three-form design. A master data set with no missing data was generated with these five variables (N = 500). Data then were removed completely at random from the three-form design variables such that exactly one of the three variables had missing data for each subject. This random deletion of data was performed 20 times, producing 20 data sets with missing data.

The covariance matrix for the five variables then was reproduced in the 20 data sets. Five different analysis methods were used: EM algorithm, pairwise deletion, mean substitution, single imputation (based on the EM algorithm), and multiple imputation (also based on the EM algorithm).

For the simulation, all variances for the master data set were around 1.0, and all covariances were positive, ranging from .36 to .70.

The results for the simulation appear in table 6. The values in table 6 are deviations from the actual values obtained in the analysis of the master data set containing no missing data (deviations are averaged over all variances and over all covariances). If the estimation procedure is unbiased, the mean of the estimate of each variance and covariance element over the 20 data sets should be very close to the parameter value estimated in the master data set with no missing data. A positive deviation means that the estimate is too high (i.e., positively biased); a negative deviation means that the estimate is too low (i.e., negatively biased).

EM Algorithm. The analysis by EM algorithm was performed using EMCOV.EXE, the hybrid EM algorithm program; the same results were obtained using BMDP AM. Details regarding the program can be obtained elsewhere (Graham and Donaldson 1993; Graham et al., submitted) or by writing to John Graham. The EM algorithm performed very well, producing the least biased and most efficient estimates.[8]

Pairwise Deletion. In this example, pairwise deletion performed nearly as well as the EM algorithm. The variance and covariance elements were estimated virtually without bias (on average), and the standard errors for the estimation were only slightly higher than those obtained with the EM algorithm. However, the lack of bias in this example is due to the fact that the data were missing completely at random. In addition, despite the fact that there is very little bias with pairwise deletion, there is no guarantee that the matrix itself will be positive-definite.

Mean Substitution. It should be very clear from this simple example that mean substitution is the worst of the analysis options. Both variance and covariance elements were seriously negatively biased.

TABLE 6. *Results for three-form design simulation*

Mean deviations from true parameter values
Cause of missingness: Random process

	Estimation procedure				
		pair-	single	mean	mult
	EM	wise	imp	replc	avg
Variances	.001	.002	-.201	-.310	-.015
Covariances	.002	.002	.002	-.185	-.000
Average Standard Error	.037	.045	.036	.025	.040

KEY: EM = EM algorithm; Pairwise = pairwise deletion (inclusion); Single imp = single imputation (based on EM algorithm); mean replc = mean replacement; mult avg = average of 5 multiple imputations.

Single Imputation. The single imputation procedure was included here to illustrate the problem with variance estimates. These single imputations were produced as a byproduct of the EMCOV.EXE program, not based on simple regression. In fact, because data from the three-form design do not conform to the monotone missing data pattern, performing regression-based single imputation would not be appropriate.

The results for single imputation were identical to the EM algorithm for covariance estimates (in this example, covariances were estimated in the same way for the two approaches). As expected, however, the variance elements were estimated with serious negative bias.

Multiple Imputation. Multiple imputation began with the single imputation described in the previous section. As a byproduct of the EMCOV.EXE program, singly imputed values are output along with a vector of residuals for each variable. For each missing score, one element from the distribution of residuals for that variable was sampled (with replacement), thereby restoring variability to the estimate of the sums of squares (and, hence, the variance). This process was repeated five times. The entries in table 6 are average parameter estimates over the five replications of this process.

The results show that the multiple imputation procedure provided estimates that were approximately equal to those obtained with the EM algorithm. The multiple imputation estimates were about equally unbiased, with only slightly larger standard errors.

Examples of Analyses To Deal With Attrition

The attrition example is taken from Graham and Donaldson (1993), where additional details of the study may be found. In this example, data were simulated from a simple drug prevention analysis as shown in figure 2. There were no missing data on pretest drug use or on the program variable, but some data were missing for the posttest drug use variable.

A master data set was generated with no missing data. The relationships between variables were modeled after actual drug prevention data. The correlation between pretest and posttest drug use was r = .60, and the correlation between the program variable and posttest drug use simulated a modest program effect, r = -.10. The correlation between the program variable and pretest drug use was nearly 0, r = .03.

From the master data set (N = 500), missing data were generated for the simulated posttest drug use variable producing the following four patterns: (1) differential attrition caused by *pre*test drug use (i.e., an accessible missing data mechanism); (2) differential attrition caused by *post*test drug use (i.e., an inaccessible missing data mechanism); (3) no

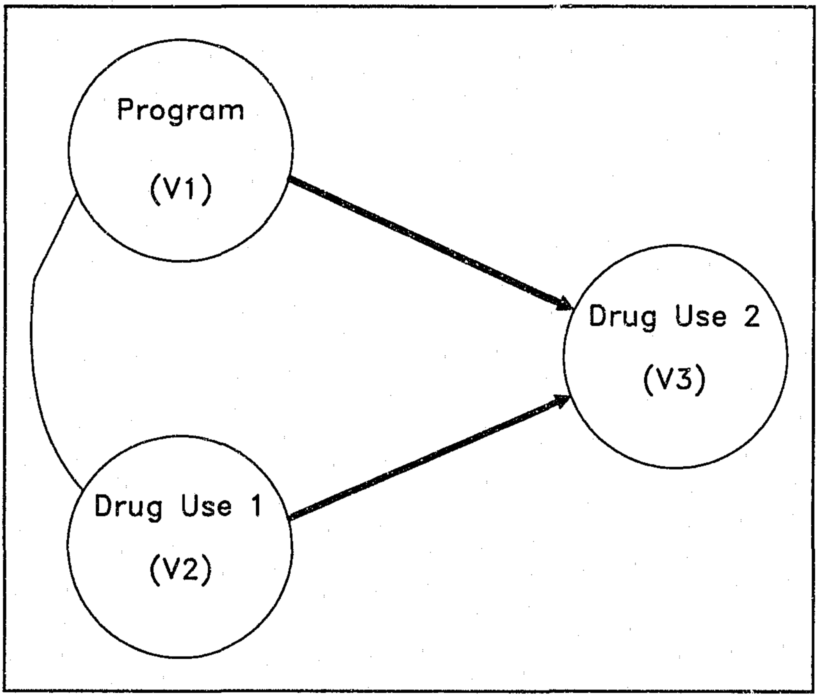


FIGURE 2. *Simple attrition model*

differential attrition, missingness caused by pretest drug use (i.e., accessible); and (4) no differential attrition, missingness caused by posttest drug use (i.e., inaccessible).

Twenty different data sets were generated for each of the four attrition patterns. All data sets were analyzed by standard complete cases analyses and with the EM algorithm (the hybrid EMCOV.EXE program was used). The standard complete cases analyses were zero-order correlation analysis and ANCOVA with posttest drug use as the dependent variable and pretest drug use as the covariate. For the EM algorithm, the same two analyses were repeated but were based on the EM algorithm estimates of the covariance matrix.

For cell (2) of the design (inaccessible missing data mechanism, differential attrition), the data also were analyzed using a correction to the EM algorithm. The details of the correction appear in Graham and Donaldson

(1993). In brief, the cases originally missing were randomly sampled, and the data were restored for this random sample. Then, rather than using the regression equation in the original sample to predict scores for the missing cases, the regression equation in the random sample was used to predict missing scores for cases with data still missing.

The results for the correlation analyses appear in table 7. As shown, there are no biases for the correlation associated with prevention program effects if there is no differential attrition. This is true even with an inaccessible missing data mechanism.

When there is differential attrition and an accessible mechanism, the standard zero-order correlation analysis based on complete cases is biased because it does not take the cause of missingness into account. In this same situation, the zero-order correlations based on the EM algorithm are unbiased.

When there is differential attrition with an *in*accessible missing data mechanism, both standard complete cases and EM algorithm analyses produce biased estimates of program effects. However, note that the correction to the EM algorithm based on a random sample of previously missing cases produces an unbiased estimate of the correlation corresponding to the program effect.

The results for the ANCOVA appear in table 8. As with the zero-order correlation analyses, there are no biases for the regression weights associated with prevention program effects if there is no differential attrition. This is true even with an inaccessible missing data mechanism.

When there is differential attrition and an accessible mechanism, the regression coefficient based on complete cases is not biased because it does take the cause of missingness into account. In fact, in this particular situation (missing data only for posttest drug use), the complete cases ANCOVA analysis is equivalent to the analysis based on the EM algorithm.

TABLE 7. *Attrition study: Program effect results based on correlation coefficients*

	Deviations from actual values (standard errors in parentheses)				
	Missing data mechanism				
	Accessible		Inaccessible		
Differential attrition	Complete	EM	Complete	EM	EM_c
Yes	-0.09	0	-0.16	-0.11	-.01
	(.01)	(.01)	(.01)	(.00)	(.01)
No	-0.02	0	-.00	-.00	
	(.01)	(.01)	(.01)	(.01)	

KEY: Complete = listwise deletion; EM = EM algorithm; EM_c = correction to the EM estimates based on the sample of previously missing cases.

SOURCE: Graham, J.W., and Donaldson, S.I. Evaluating interventions with differential attrition: The importance of nonresponse mechanisms and use of follow-up data. *Journal of Applied Psychology* 78:119-128, 1993; Copyright (1993) by the American Psychological Association. Reprinted by permission.

TABLE 8. *Attrition study: Program effect results based on ANCOVA (betas)*

	Deviations from actual values (standard errors in parentheses)				
	Missing data mechanism				
Differential attrition	Accessible		Inaccessible		
	Complete	EM	Complete	EM	EM$_c$
Yes	0	0	-0.25	-0.25	-.01
	(.01)	(.01)	(.01)	(.01)	(.03)
No	-0.01	-0.01	-.00	-.00	
	(.01)	(.01)	(.02)	(.02)	

KEY: Complete = listwise deletion; EM = EM algorithm; EM$_c$ = correction to the EM estimates based on the sample of previously missing cases.

SOURCE: Graham, J.W., and Donaldson, S.I. Evaluating interventions with differential attrition: The importance of nonresponse mechanisms and use of follow-up data. *Journal of Applied Psychology* 78:119-128, 1993; Copyright (1993) by the American Psychological Association. Reprinted by permission.

When there is differential attrition with an *in*accessible missing data mechanism, both standard complete cases ANCOVA and ANCOVA based on the EM algorithm produce biased regression estimates of program effects. However, as with the zero-order correlation analysis, the correction to the EM algorithm based on a random sample of previously missing cases produces an unbiased estimate of the regression weight corresponding to the program effect.

Followup to Attrition Study. In order to illustrate the fact that pairwise deletion is not a general solution to missing data problems, one cell of the previous attrition study was reanalyzed. In particular, the cell with differential attrition and accessible missing data mechanism was examined. Table 9 presents the results of this brief simulation in which five new data sets were generated with differential attrition and the accessible missing data mechanism. The data sets were analyzed with standard complete cases analyses (i.e., listwise deletion), the EM algorithm, and pairwise deletion.

As before, the complete cases correlation for the program effect is biased. However, the complete cases estimate of correlation R_{12} also is substantially biased, and this produces an unbiased estimate of the regression weight corresponding to the program effect. Also as before, all correlation and regression estimates based on the EM algorithm are unbiased. Finally, the estimates corresponding to program effects based on pairwise deletion are seriously biased both for the correlation analysis and the ANCOVA analysis.

Analysis of the Process Model

The analysis of the process model will be used to illustrate the utility of the multiple-group structural equation-modeling procedure (e.g., Allison 1987) with empirical data. In this case, there were 1,977 cases with complete data for pretest drug use (seventh grade), program variables, process data relating to the normative education curriculum, and posttest (eighth

TABLE 9. *Attrition study: Comparisons of estimates based on various procedures*

Mean deviations
from true parameter values
(accessible missing data mechanism)

	True	EM	Estimation procedure	
			pairwise	listwise
R21	-.025	.000	.000	-.175
R31	-.096	-.009	-.108	-.108
R32	.598	-.008	-.006	-.006
b1	-.18	-.021	-.241	-.021
b2	.596	-.007	-.006	-.007

grade) drug use. However, there were only 925 cases for the immediate posttest measure of behavioral resistance skills. Thus, a complete cases analysis of the entire process model was undesirable because it would produce a substantial loss of statistical power for certain parts of the model and would be based on a rather small subset of the total sample. Because there were just two major patterns of missingness, the multiple-group procedure would be ideal for analysis with missing data. The annotated control statements for running the appropriate LISREL model can be obtained from John Graham.

The process model tested is shown in figure 1. The results for complete cases analysis and analysis using the multiple-group procedure are presented in table 10. It can be seen by inspection of table 10 that results for the parts of the model not related to the resistance-training measure

TABLE 10. *Results of process model with complete cases and multiple-group LISREL procedure*

	Cause	RT	RT	NORM	NORM	NORM
	Effect:	Behav	Alc 3	NotOK	Prev	Alc 3
Complete	b	0.297	0.122	0.116	-0.27	-.036
Cases	SE	0.05	0.049	0.05	0.056	0.049
N = 925	z	5.94	2.48	2.34	4.92	0.73
Allison	b	0.299	0.098	0.159	-0.25	-.024
Procedure	SE	0.05	0.033	0.034	0.037	0.032
+N = 1052	z	5.98	2.99	4.69	6.8	0.73

	Cause	Behav	NotOK	Prev
Analysis	Effect:	Alc 3	Alc 3	Alc 3
Complete	b	-0.082	-0.199	0.108
Cases	SE	0.032	0.033	0.029
N = 925	z	2.58	6.09	3.71
Allison	b	-0.075	-0.209	0.097
Procedure	SE	0.03	0.021	0.019
+N = 1052	z	2.47	9.77	5.00

KEY: RT = resistance training program dummy variable; NORM = normative education program dummy variable; Behav = measure of resistance skills; NotOK = beliefs about acceptability of adolescent alcohol use; Prev = perceptions of adolescent drug use prevalence; Alc 3 = alcohol use at time 3 (8th grade).

are rather different for the complete cases and multiple-group procedures. Note that the parameter estimates are comparable for the two procedures but the standard errors for the multiple-group procedure are considerably smaller for parameters not directly involving the measure of behavioral skills. Z-values for these estimates are shown in bold in table 10.

Note that the parameter estimates and standard errors for parameters directly involving the measure of behavioral resistance skills are virtually unchanged for the two models. This makes sense in that these estimates are based on the smaller sample size (N = 925). Also note that the parameter estimate, NORM --> ALC 3, was not significant for the complete cases analysis (N = 925) and also was not significant when the remaining data were added, bringing the effective sample size to N = 1977.

Substantive Conclusions. The data for this example come from the AAPT study (Hansen and Graham 1991). Based on these analyses, it is reasonable to conclude that the normative education (NORM) curriculum had significant effects on the mediating variables: perceptions of prevalence of peer drug use and perceptions of acceptability of peer alcohol use. In turn, these mediating variables have a significant effect on alcohol use at the eighth grade (all analyses controlled for alcohol use at seventh grade). That is, there was a significant indirect effect of the NORM program on eighth-grade alcohol use, which was mediated by perceptions of prevalence and acceptability of peer alcohol use.

The resistance training (RT) curriculum had a significant effect on the mediating variable, behavioral resistance skills, which in turn had a modest but significant effect on eighth-grade alcohol use. There was a significant indirect effect of the RT program on eighth-grade alcohol use, which was mediated by behavioral resistance skills. However, there also was a significant, direct, counterproductive effect of the RT program on eighth-grade alcohol use. Although there is no firm evidence to explain this direct effect, Donaldson and colleagues (submitted) have shown that the effect could be due to the unexpected effect of RT increasing perceptions of peer prevalence of drug use offers.

Analysis of Program by Grade of Intervention Interaction

The missing data problem with the analysis of the program by grade interaction was introduced in table 3. In the AAPT study (Hansen and Graham 1991), programs were implemented in the fifth and seventh grades. One of the main questions of interest was whether the programs would have greater effectiveness when implemented earlier or later. One hypothesis was that it is best to intervene in the seventh grade, when students are beginning to feel strong pressures to use various substances. On the other hand, one of the key curricula, normative education, was designed to demonstrate to young adolescents that using drugs at their age is not as common as most kids believe. One might suppose that such a curriculum would be more effective in the fifth grade, when substance use is very low, than in the seventh grade, when at least some adolescents have begun using drugs. It was an easy matter to do a posttest-only analysis of variance using grade, program, and the gradexprogram interaction as effects. The results for the posttest-only analysis are presented in table 11.

Unfortunately, the more sensitive (and, perhaps, more appropriate) ANCOVA could not be used because there was no generally appropriate pretest measure of drug use that could be used as a covariate in the ANCOVA. The fifth-grade measure of drug use was available for those receiving the program as fifth graders, and the seventh-grade measure of drug use was available for those receiving the program in the seventh grade. However, these two measures were not equivalent and could not be used as a single covariate (i.e., pretest use) in the same analysis. If complete case analysis were used, either fifth graders only or seventh graders only would be used. This obviously was no solution.

Fortunately, the authors did include pretest measures at the fifth-grade level for one of the two cohorts receiving the program in the seventh grade (see panel 2 in table 3). However, even with this, if complete cases analysis were used, it would mean discarding data for one entire cohort of subjects receiving the program as seventh graders (120 classrooms). This could bias an interpretation and would reduce statistical power.

TABLE 11. *Analysis of variance*

Dependent variable = 8th-grade alcohol use

Source	Posttest only		Pre-post EM algorithm	
	z	p	z	p
Alc5	--	--	3.68	.0001
Pub5	.51	ns	.92	ns
Pub7	.05	ns	.61	ns
PYear	1.55	.12	.85	ns
NORM	-1.78	.08	-1.98	.048
PYear*NORM	.07	ns	-.42	ns
RT	-.03	ns	.14	ns
PYear*RT	-.82	ns	-.53	ns
NORM*RT	-.90	ns	-1.14	ns
PYear*norm*rt	-.20	ns	.09	ns

KEY: N = 420 classrooms. Alc5 = alcohol use at 5th grade; Oub5 = public (1) versus private (-1) schools (5th-grade interventions); Pub7 = public (1) versus private (-1) schools (7th-grade interventions); PYear = grade of intervention (7th = 1, versus 5th = -1); NORM = NORM (1) versus NoNORM (-1); RT = RT (1) versus NoRT (-1).

One would think that the multiple-group structural equation-modeling procedure would be ideal for this missing data problem in that there were two missing data patterns—those with the fifth-grade pretest and those without it. Unfortunately, because missingness was partially confounded with grade of intervention, the group containing missing data had no variability for the grade of intervention variable.

The solution used here is the EM algorithm. Although missingness was partially confounded with grade for the multiple-group analysis, grade of intervention was well defined for the sample as a whole. EMCOV.EXE, the hybrid EM program, was used for this problem; BMDP AM also would perform well for this type of problem.

The results for the EM algorithm also appear in table 11. For the post-test-only analysis, the program NORM had only a marginally significant effect on eighth-grade alcohol consumption. However, for the ANCOVA using pretest as a covariate, this effect reached statistical significance. Note that none of the interactions involving grade of intervention even approached statistical significance. One can conclude from these findings that: (1) the NORM program has a modest effect on reducing or delaying the onset of alcohol use, (2) the RT curriculum has no overall effects, and (3) fifth- or seventh-grade interventions are equivalent.

The third result should be modified, however, in that those receiving the program in the fifth grade also received a one- to three-session booster in the seventh grade. Thus, the conclusion to be reached here is that receiving the program in seventh grade only is as effective as receiving the program in the fifth grade with a seventh-grade booster.

DISCUSSION

A cross-section of missing data problems has been presented in this chapter. Omissions within a survey, attrition from whole waves of measurement, and planned missingness have been discussed. All of these problems are encountered routinely in drug prevention research.

Attrition Solutions

Two approaches to solving the problem of attrition, perhaps the most insidious problem discussed in the drug prevention literature, were presented. The first solution is to plan the research with attrition in mind, identifying the likely causes of attrition and measuring as many of them possible. If one can include these causes in the analysis, biases associated with attrition can be minimized or eliminated.

The second solution to the problem of attrition is to collect data from a sample of those initially missing. This type of solution may be difficult to implement but may be cost effective in the long run. For some kinds of prevention studies, studies involving parents or other adults, for example, experience shows that the sampling procedure can be successful. For studies involving adolescents, however, use of this procedure may present more of a challenge.

General Missing Data Analysis Solutions

Two general solutions for analysis with missing data, the EM algorithm and a multiple-group structural equation-modeling procedure (e.g., Allison 1987), were discussed. For analysis of continuous data, especially analyses that can be based on a covariance matrix, one of these solutions always should be used.[9] The EM algorithm theoretically is applicable to any missing data problem. In practical terms, however, its ready availability is limited currently to BMDP, which may not be widely available. However, other versions of the EM algorithm (e.g., EMCOV.EXE) are becoming more readily available. Also, current implementations of the EM algorithm do not allow for special problems, such as adjusting the EM estimates for inaccessible missing data mechanisms. The other drawback noted for the general EM algorithm is that correct standard errors are not computed for the parameter estimates of ultimate interest. Fortunately, one can use bootstrapping procedures (Efron 1982) to obtain these standard errors for any problem.

The multiple-group structural equation-modeling solution (e.g., Allison 1987), is an excellent procedure when it is applicable. The main advantages of the procedure are (1) that it provides unbiased and statistically powerful estimates of the model of ultimate interest, and (2) that it provides good estimates of the standard errors for these model parameters. Because of the practical and statistical limitations on the number of missing data patterns that may be present, this procedure often involves discarding a small amount of data. However, experience shows that this loss of data is unimportant compared to the gains that can be made.

Statistical Power

Whenever a researcher has missing data, there are important statistical power issues to be considered. It has been mentioned throughout this chapter that one of the advantages of using the EM algorithm or multiple-group structural equation-modeling procedures is that one makes full use of data that are available. This means that, compared to analyses using only complete cases, one can estimate certain parameters with greater statistical power.

This point was made most clearly in the example of analysis of the process model of prevention program effects (see figure 1 and table 10). Compared to analyses with complete cases, statistical power was boosted substantially for several parameter estimates that were not related to the missing variable.

On the other hand, this same example illustrated very well that these missing data procedures do not give something for nothing. The results in table 10 showed that there was no gain in statistical power for parameter estimates relating directly to the variable with missing data.

Statistical power also is a particularly important issue when research plans specify missing data patterns. For example, although the advantage of using the three-form design is that one can collect data for additional variables without placing too much of a burden on any individual respondent, researchers who use this approach should bear in mind that they are giving up statistical power. With the three-form design (see table 2), correlations between variables within the same block of items are estimated with only two-thirds of the total sample. Correlations between variables across blocks of items are estimated with only one-third of the total sample. Researchers should carefully weigh the loss of statistical power associated with this measurement plan. In most cases, a researcher will have ample power even with a one-third sample. However, for certain key analyses, this could be totally unacceptable.

Limitations

This chapter has not discussed all missing data problems nor presented all solutions. Several important procedures available for dealing with missing data in the continuous variable situation probably were omitted. The authors hope that readers will forgive these omissions. In addition, procedures for categorical data analysis with missing data were presented. Although this certainly is an important area, it is one that goes beyond the scope of the present chapter. Others who have discussed solutions to this problem recently include Little and Rubin (1987), MacKinnon and Graham (1993), Muthen and colleagues (1987), and Rindskopf (1992).

Points To Remember

There are several points made in this chapter that should be reemphasized:

1. Whenever possible, use the EM algorithm (or other maximum likelihood procedure, including the multiple-group structural equation-modeling procedure or, where appropriate, multiple imputation) for analyses involving missing data.

2. If other analyses must be used, keep in mind that they produce biased results and should not be relied upon for final analyses. Recom-

mendations regarding the use of other procedures for *preliminary* looks at the data include:

a. Never use mean substitution, even for preliminary analyses.
b. With minimal missing data, analysis of complete cases may be a reasonable solution.
c. If data are missing completely at random, pairwise deletion or complete cases analysis may be a reasonable solution.
d. If data are not missing completely at random and the cause of missingness has been measured, complete cases may produce unbiased estimates, although it is a generally less powerful approach than the EM algorithm or multiple-group procedure.

3. When data are missing, missing data analysis procedures do *not* generate something out of nothing. Missing data analysis procedures *do* make the most out of the data available, maximizing precision of estimation and eliminating biases.

4. When data are missing, work hard to find the cause of missingness and include the cause in the analysis model. When planning a study, think about what the causes of missingness are likely to be and obtain measures for as many causes as possible.

5. Ultimately, one can never know whether the cause of missingness is fully accessible. So, one solution is to sample the cases with missing data and adjust EM algorithm parameter estimates accordingly.[10]

NOTES

1. Note that, in actual practice, one would expect some amount of nonrandom missingness to be superimposed over top of the random missingness due to the three-form design.

2. Because the missing values are imputed and not real, the standard errors for these analyses will be lower than they should be. In these

cases, other methods (e.g., using bootstrap procedures) must be used to obtain proper estimates of the standard errors.

3. A general version of the EM algorithm also should be available with the next release of SYSTAT.

4. There also is a general version of the EM algorithm available within the Gauss program. However, this may be even less accessible than BMDP. Although the Gauss program undoubtedly will prove to be a very good program, the authors are not prepared to comment on it further at this time.

5. This is true if the causes of missingness are random processes or if they are accessible and are included properly in the analysis.

6. By "more efficient," the authors mean less variability around the *true* parameter value. Other approaches may yield less variability (i.e., lower standard errors) around *biased* parameter estimates.

7. One should be careful in this step to make use of a randomizing procedure that provides a good approximation to true random selection. The simplest approaches (e.g., using the RANDOMIZE TIMER function in BASIC) are known to be flawed. Results based solely on this randomizing procedure will produce standard errors that are incorrect to an unknown degree.

8. Again, by "most efficient," the authors mean the least variability around the true parameter value. Some of the values for average standard error shown in table 6 are smaller than those shown for the EM algorithm. However, these figures refer to variability around the substantially biased parameter estimate.

9. Some missing data problems (e.g., analysis of difference scores) involve continuous data but cannot be analyzed directly with a covariance matrix. Such problems can be handled with multiple

imputation procedures using the EM algorithm (not simple regression) as the basic single imputation method.

10. This suggestion applies especially to the case of attrition but may be of less value for the case of nonrandom omissions.

REFERENCES

Allison, P.D. Estimation of linear models with incomplete data. In: Clogg, C., ed. *Sociological Methodology 1987.* San Francisco: Jossey Bass, 1987. pp. 71-103.

Biglan, A.; Severson, H.; Ary, D.V.; Faller, C.; Gallison, C.; Thompson, R.; Glasgow, R.; and Lichtenstein, E. Do smoking prevention programs really work? Attrition and the internal and external validity of an evaluation of a refusal skills training program. *J Behav Med* 10:159-171, 1987.

Dempster, A.P.; Laird, N.M.; and Rubin, D.B. Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion). *J R Stat Soc* B39:1-38, 1977.

Dent, C.W. Using SAS linear models to assess and correct for attrition bias. In: *International Proceedings of the SAS User's Group 13th Annual Conference*, 1988.

Dixon, W.J., ed. *BMDP Statistical Software.* Rev. print. Berkeley, CA: University of California Press, 1988.

Donaldson, S.I.; Graham, J.W.; Piccinin, A.M.; and Hansen, W.B. "Resistance Skills Training and Alcohol Use Onset: Evidence for Beneficial and Potentially Harmful Effects in Public and Private Catholic Schools." Manuscript submitted for publication.

Efron, B. *The Jackknife, the Bootstrap, and Other Resampling Plans.* Philadelphia: Society for Industrial and Applied Mathematics, 1982.

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*. NIDA Research Monograph 107. DHHS Pub. No. (ADM)91-1761. Washington, DC: Supt. of Docs., U.S. Govt. Print. Off., 1991. pp. 81-109.

Frane, J. Description and estimation of missing data. In: Dixon, W.J., ed. *BMDP Statistical Software Manual*. Vol. 2. Berkeley, CA: University of California Press, 1988.

Graham, J.W., and Donaldson, S.I. Evaluating interventions with differential attrition: The importance of nonresponse mechanisms and use of follow-up data. *J Appl Psychol* 78:119-128, 1993.

Graham, J.W., and Hofer, S.M. "Testing Interactions With Missing Data: An Application of the EM Algorithm." Paper presented at the annual meeting of the Society for Multivariate Experimental Psychology, Cape Cod, MA, October 22-24, 1992.

Graham, J.W., and Hofer, S.M. "EMCOV.EXE Users Guide." Unpublished manuscript.

Graham, J.W.; Hofer, S.M.; and MacKinnon, D.M. "Maximizing the Usefulness of Data Obtained With Planned Missing Value Patterns: An Application of the EM Algorithm and Multiple Imputation." Manuscript submitted for publication.

Graham, J.W.; Rohrbach, L.A.; Hansen, W.B.; Flay, B.R.; and Johnson, C.A. Convergent and discriminant validity for assessment of skill in resisting a role play alcohol offer. *Behav Assess* 11:353-379, 1989.

Hansen, W.B.; Collins, L.M.; Malotte, C.K.; Johnson, C.A.; and Fielding, J.E. Attrition in prevention research. *J Behav Med* 8:261-275, 1985.

Hansen, W.B., and Graham, J.W. Preventing alcohol, marijuana, and cigarette use among adolescents: Peer pressure resistance training versus establishing conservative norms. *Prev Med* 20:414-430, 1991.

Hansen, W.B.; Graham, J.W.; Wolkenstein, B.H.; Lundy, B.Z.; Pearson, J.L.; Flay, B.R.; and Johnson, C.A. Differential impact of three alcohol prevention curricula on hypothesized mediating variables. *J Drug Issues* 18:143-153, 1988.

Hansen, W.B.; Graham, J.W.; Wolkenstein, B.H.; and Rohrbach, L.A. Program integrity as a moderator of prevention program effectiveness: Results for fifth grade students in the Adolescent Alcohol Prevention Trial. *J Stud Alcohol* 52:568-579, 1991.

Heckman, J.J. Sample selection bias as a specification error. *Econometrica* 47:153-161, 1979.

Jöreskog, K.G., and Sörbom, D. *LISREL 7 User's Reference Guide.* Mooresville, IN: Scientific Software, Inc., 1989.

Leigh, J.P.; Ward, M.M.; and Fries, J.F. Reducing attrition biac with an instrumental variable in a regression model: Results from a panel of rheumatoid arthritis patients. *Stat Med* 12:1005-1018, 1993.

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

MacKinnon, D.P., and Graham, J.W. "Analysis of Categorical Data With Missing Values." Unpublished manuscript. (Available from David MacKinnon, Department of Psychology, Arizona State University, Tempe, AZ 85287.)

Muthen, B.; Kaplan, D.; and Hollis, M. On structural equation modeling with data that are not missing completely at random. *Psychometrika* 52:431-462, 1987.

Rindskopf, D. A general approach to categorical data analysis with missing data, using generalized linear models with composite links. *Psychometrika* 57:29-42, 1992.

Rohrbach, L.A.; Graham, J.W.; Hansen, W.B.; Flay, B.R.; and Johnson, C.A. Evaluation of resistance skills training using multitrait-multimethod role play skill assessments. *Health Educ Res* 2:401-407, 1987.

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

AUTHORS

John W. Graham, Ph.D.
Professor
College of Health and Human Development
The Pennsylvania State University
110 South Henderson
University Park, PA 16802-6504

Scott M. Hofer
Graduate Research Assistant

Andrea M. Piccinin
Graduate Research Assistant

Department of Psychology
University of Southern California
Los Angeles, CA 90089-1061

Latent Class Analysis of Substance Abuse Patterns

John S. Uebersax

ABSTRACT

This chapter discusses use of latent class analysis (LCA) as a tool for identifying substance use patterns in cross-sectional data. LCA serves as an exploratory and data reduction tool that helps clarify the nature of substance use and may provide insight concerning effective prevention strategies. LCA is well suited to categorical data such as typically are collected in substance use research. Use of LCA can be divided into three steps: (1) model comparison and selection, (2) assignment of cases to latent classes, and (3) interpretation of the latent classes. Quantitative indices of model fit may assist model comparison and selection. Latent classes can be interpreted by examining probabilities of substance use in each latent class and by examining differences on exogenous variables. Limitations, extensions, and software for LCA are discussed. An example illustrates use of LCA with actual data collected from a current substance abuse prevention study.

INTRODUCTION

Substance abuse is not the same in every case. There are important differences among individuals in terms of the substances abused and the amount, frequency, and social context of use. Recognition and identification of common patterns promote understanding of the psychological determinants of substance abuse and the development of more effective prevention interventions.

Latent class analysis (LCA) is a statistical method for finding groups in data. LCA is related to "mixture model" types of cluster analysis (Day 1969; Wolfe 1970). LCA differs from most forms of cluster analysis, however, in that it is intended mainly for use with categorical data. This is significant because, in substance abuse research, variables typically are measured at the categorical level. This chapter discusses use of LCA for substance abuse prevention research. The focus is practical rather than technical and addresses the question, "How does one actually use LCA in a substance abuse prevention study?"

Latent Class Analysis (LCA)

LCA is attributable mainly to sociologist Paul Lazarsfeld (1950). Lazarsfeld envisioned LCA as a tool to identify respondent groups from survey data. Applications were limited until Goodman (1974) supplied an efficient estimation method. LCA now is used increasingly, especially in psychology, sociology, education, and health research. A book by Lazarsfeld and Henry (1968) remains an important source of information on LCA. A good introduction to the subject is provided by McCutcheon (1987). For technical details on LCA, see Goodman (1974). Langeheine and Rost (1988) discuss current developments in the area.

The LCA model posits the existence of two or more population subtypes or *latent classes*. Each latent class has a set of probabilities for various responses on each observed (*manifest*) variable. In the present context, a latent class corresponds to an ideal substance abuse pattern; response probabilities are the probabilities of various levels of substance use for each latent class.

The model is understood easily with reference to table 1 and figure 1. Table 1 illustrates the concept of a response pattern. s_1, s_2, and s_3 denote responses to three substance use items, coded 0 = not used and 1 = used; there are eight possible response patterns of the form (s_1, s_2, s_3). Table 1 shows the patterns and their hypothetical observed frequencies in a population.

TABLE 1. *Possible response patterns for three dichotomous substance use items*

Pattern	Response pattern*			Observed frequency
	s_1	s_2	s_3	
1	0	0	0	432
2	0	0	1	23
3	0	1	0	31
4	0	1	1	17
5	1	0	0	175
6	1	0	1	84
7	1	1	0	126
8	1	1	1	87

KEY: * Coded as 0 = nonuse, 1 = use

Figure 1 schematically represents the LCA model. Starting at the top, the circle represents a case in the population selected at random. X_1, X_2, and X_3 represent three latent classes. The use of three substances (the same substances for each class) is denoted here by s_1, s_2, and s_3. The numbers represent probabilities. The top set are the probabilities of a randomly selected case belonging to each latent class; these are the *latent class probabilities* of the LCA model. The lower set are the probabilities of substance use given each latent class, or *conditional response probabilities.* The conditional response probabilities shown in figure 1 are the probabilities of substance use; subtracting them from 1 gives the probabilities of nonuse (items with more than two response levels have, correspondingly, several conditional response probabilities for each latent class).

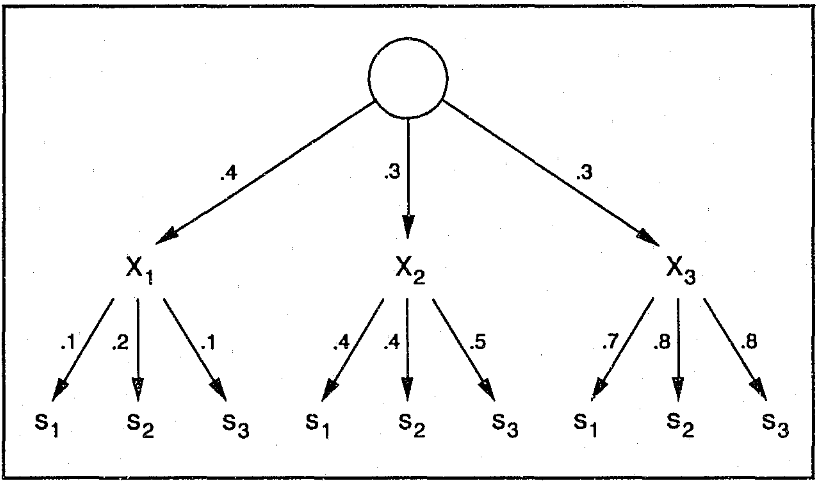


FIGURE 1. *Schematic representation of the latent class model*

The input for LCA consists of observed response pattern frequencies like those in table 1. For each analysis, one also specifies the number of latent classes in the solution. The procedure then determines optimal (maximum likelihood) estimates for the unknown latent class and conditional response probabilities, which form the basis of interpretation of results.

EXAMPLE APPLICATION

Background

The example here uses data from a substance abuse prevention study in Winston-Salem, NC. The study involves middle school and high school students in the Winston-Salem public school system. Reported substance use by high school students in the 1991-1992 academic year is considered here; analysis is limited to 11th- and 12th-grade male students, for whom substance use is highest.

Data were obtained with a 115-item, self-administered survey. The survey contains items on current and lifetime substance use; hypothesized mediating variables (e.g., personality, attitudes towards drugs); and demographic information.

The present analysis considers seven lifetime substance use items: drunkenness, cigarettes, marijuana, cocaine, heroin, amphetamines, hallucinogens, and inhalants. Responses on each item, originally ordered-categorical, were recoded to dichotomies: Students who reported having been drunk once or more were coded positive on the drunkenness variable; those reporting having smoked at least one pack of cigarettes were coded positive on the cigarettes variable; all other variables were coded positive if the student reported at least one lifetime use of the corresponding substance.

Respondents are assured anonymity, and the survey response rate is high overall—over 90 percent. For this analysis, a small number of students who did not respond to every item were eliminated; the total N for the analyses reported here is 855.

Analysis and Results

Use of LCA can be divided into three steps; the analysis here illustrates each of them:

- *Model selection.* One first tests several latent class models and selects one that is optimal in some way. Models differ mainly in the number of latent classes but also may differ in other ways. Various measures of model fit can be used to assist model selection.

- *Assignment of cases.* Once a model is selected, each case is assigned to its most likely latent class based on the model parameter estimates and cases' responses to the manifest variables.

- *Latent class interpretation.* The main procedure for interpretation is to examine the response probabilities of items given each latent class. One also may examine whether latent classes differ on exogenous variables—that is, variables other than those used to estimate the latent classes.

From the raw data—students' responses to the substance use items—observed frequencies for each response pattern were generated using the PROC FREQ feature in SAS, with the LIST option. This supplied the input to the LCA program PANMARK (other programs could be used as well; see **SOFTWARE** section). Six models, with from one to six classes, were tested; the results are summarized in table 2.

The table shows the models, the number of estimated parameters, the degrees of freedom (df), and model fit according to three criteria. The df are equal to the number of possible rating patterns minus 1 (here, 2^7-1 = 127) minus the number of estimated parameters.

TABLE 2. *Results of latent class models of responses to 7 substance use items by 855 male 11th- and 12th-grade students*

Model	Description	No. of parameters	df	G^2	X^2	Normed fit index
M1	1 class	7	120	1,468.42	24,550.34	—
M2	2 classes	15	112	306.29	501.19	.791
M3	3 classes	23	104	90.81	108.93	.938
M4	4 classes	31	96	65.61	81.13	.955
M5	5 classes	39	88	43.08	48.82	.971
M6	6 classes	47	80	31.89	38.91	.978

Deciding the number of latent classes is given much attention in the technical literature. However, substance abuse prevention researchers will do well to note that model preference less often is a statistical than a practical issue. In substance abuse research, one is more likely to view latent classes as a means for data reduction; a solution is sought that captures as much meaningful variation among cases as possible without resorting to an excessive number of classes. Still, although the researchers' judgment should be primary in selecting among models, one should not lose sight of statistical criteria altogether.

The familiar Pearson X^2 statistic, calculated by comparing observed with model-predicted response pattern frequencies, can be used to assess goodness of fit. An alternative is the likelihood-ratio chi square statistic, G^2 (see McCutcheon 1987 for details). Under ideal conditions, G^2 and X^2 both follow the χ^2 distribution and can be used to statistically test departure of the model from observed data. For an acceptable model, both statistics should be close to the df and close to each other; the required conditions are a sufficiently large sample size and data that are not too sparse (sparse data have many response patterns with small frequencies). Prevention studies usually meet the first condition, but the second sometimes is problematic. In the present case, for example, note that in table 2 with models M4-M6, G^2 and X^2 are much lower than the df, the result of sparse data.

More refined model selection criteria have been proposed that are related to G^2 but add a component to penalize models with more parameters (Collins et al., this volume; Sclove 1987). Much work, however, remains to be done in this area.

Table 2 also shows the normed fit index (nfi) (Bentler and Bonett 1980; Clogg 1977) for each model. For a model with *k* latent classes, the nfi is calculated as the G^2 statistic for a 1-class model minus G^2 for the *k*-class model, divided by G^2 for the 1-class model. It can be interpreted informally as the proportion of unexplained variance accounted for by the *k*-class model. Some researchers will find this index, which approaches model fit more from a descriptive than an inferential standpoint, useful. The nfi increases markedly going from two to three classes and little beyond five latent classes. The results, therefore, suggest that a model with from three to five latent classes is best. The solution for model M3 is relatively uninteresting, and one of the latent classes for M5 has a very low prevalence; therefore, focus attention on M4.

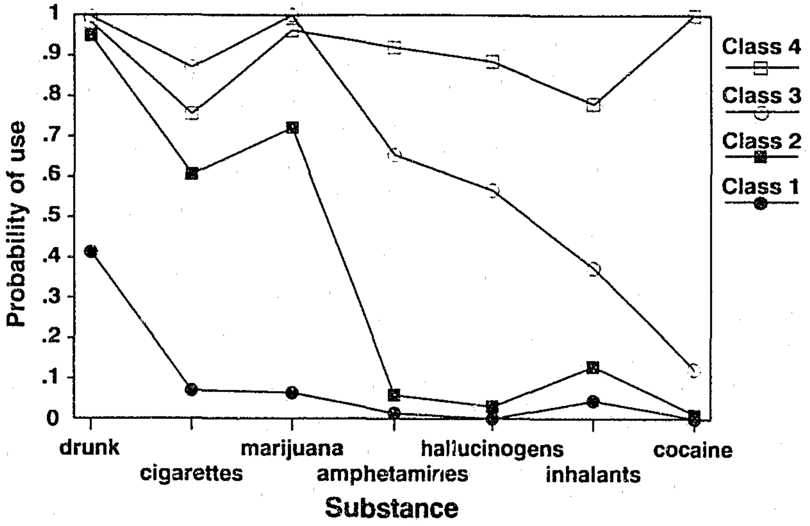


FIGURE 2. *Probabilities of substance use conditional on latent class for Model M4 of table 2*

The four latent classes for M4 have estimated population prevalences of .474 (class 1), .322 (class 2), .063 (class 3), and .141 (class 4). Figure 2 shows estimated probabilities of use of each substance for each latent class.

Class 1, accounting for nearly half the population, is termed the "non-user" group, although, interestingly, even for this group, the probability of at least one episode of drunkenness is above .4. Members of class 2 have a very high probability of reported drunkenness and lower but relatively high probabilities of cigarette and marijuana use; this group is termed "conventional substance users." Members of class 4 have very high or relatively high probabilities of reported use on all items; this group is termed "general substance abusers." For class 3, reported probabilities of use of amphetamines, hallucinogens, inhalants, and cocaine are intermediate between those of conventional substance users and general substance abusers; for drunkenness, cigarette use, and marijuana use, the probabilities are slightly higher than for general substance abusers, although for drunkenness and marijuana, the differences aþpear negligible; this is the "moderate drug use" group.

The four classes correspond to roughly increasing levels of substance use. Often the results will not lend themselves to so simple an interpretation. For example, with the same students, when items on beer, wine, and hard liquor use are added, one sees more crossing of response profiles. This shifts interpretation away from degree of overall substance use more toward different patterns that involve specific substance combinations.

Examination of a graph such as the one in figure 2 may reveal important aspects of substance abuse within a population. The following are representative of the kinds of questions that LCA may suggest:

- Many students in the nonuser group have been drunk but have not used other substances. What does this say about the socialization factors responsible for adolescent substance abuse? Where and with whom do they have the opportunity and motivation to be drunk such that they are not simultaneously exposed to or motivated to use other substances?

- Does the conventional substance use group represent a transitional stage of experimentation with alcohol, cigarettes, and marijuana from which adolescents may move to use of other drugs, or does it represent a terminal pattern that reflects preference for these substances?

- In the general substance abuse group, there still are many students who do not smoke cigarettes. What dissuades these students from smoking cigarettes? If researchers knew this, they could use the information to dissuade them from use of other substances?

- Again, in the general substance abuse group, cocaine use is more common than amphetamine, hallucinogen, and inhalant use. This is not true for the other groups. Do students who use cocaine find the other substances less interesting?

By drawing attention to these types of issues, LCA can provide insights into substance use in a population and refine thinking about prevention intervention strategies.

Another way to interpret a latent class solution is with exogenous variables. As noted above, to do this one first assigns each case to its most likely latent class. LCA provides the probabilities of membership in each latent class given each response pattern, or *recruitment probabilities* (Lazarsfeld and Henry 1968, pp. 36-38). Each case is assigned to the latent class for which its membership probability is highest. (Current LCA programs usually provide the recruitment probabilities but do not perform the actual classification of cases. Case classification can be done with the MERGE feature in SAS, or, for example, as here, with a short BASIC program.)

Once cases are assigned, latent classes can be compared on the exogenous variables. Table 3 summarizes the comparisons of the latent classes of M4 on 13 psychological scales. Each scale is composed of several items given on the same survey as the substance use items. The 13 scales also were factor analyzed using iterated principal factor analysis and orthogonal varimax rotation. The results showed a two-factor solution with two items ("academic orientation" and "assistance-helping") that did not load strongly on either factor.

Each scale was used as the dependent variable in an analysis of variance (ANOVA), with class membership as the independent variable. Results are expressed as R^2, or the proportion of total variation on the scale accounted for by between-class differences. Significance is assessed with the usual F-test. Table 3 also shows how much latent classes differ on each scale after removing the effects of all other scales; this can be interpreted as the unique contribution of each psychological variable to explaining latent class differences. Unique contributions are expressed as squared partial correlations obtained by entering each scale in a stepwise discriminant analysis after entry of all other scales, with latent class as the group variable. Significance is assessed with the F-to-enter statistic. The

TABLE 3. *Association between psychological variables and latent class membership*

Factor/Scale	Partial R^{2a}	R^{2b}
Factor I	0.4161*	0.4016*
Life compatibility	0.3615*	0.0370*
Pledges	0.3282*	0.0963*
Peer use and beliefs	0.2946*	0.0301*
Beliefs about consequences	0.2783*	0.0220*
Resistance skills	0.0964*	0.0059
Factor II	0.0353*	0.0114
Activities/alternatives	0.0398*	0.0021
Decision skills	0.0313*	0.0016
Self-esteem	0.0281*	0.0057
Goal orientation	0.0190*	0.0035
Sociability	0.0146*	0.0121
Stress management	0.0076	0.0018
Academic orientation	0.0230*	0.0017
Assistance-helping	0.0044	0.0092

KEY: [a] Based on univariate ANOVA
[b] Controlling for all other scales or factors in a stepwise discriminant analysis
* $p < .01$

results show clear differences among latent classes on the psychological variables and, in that sense, they validate the latent class solution.

A parallel analysis to the above was conducted using factor scores on the two factors. Factor scores were calculated as the unweighted mean of standardized scale scores on the constituent scales. The variables on factor I, which appear related to values, principles, and normative beliefs, more strongly differentiate classes than the more diverse factor II variables. The scales within each factor also vary in their association with latent classes. For example, the "life compatibility" variable (perceived compatibility of substance use with the student's life goals) is associated more strongly with latent class membership than the "resistance skills" variable (ability to resist peer influence to use substances). The results suggest that students' perceptions of the compatibility of substance use with their personal goals and ideal lifestyles may be an important mediating variable that should receive special attention in designing substance abuse prevention interventions.

The researcher also may wish to consider extensions of this approach to latent class interpretation. For example, with discriminant analysis, one may consider the number of discriminant functions and the amount of variance accounted for by each. Similarly, one may plot the groups relative to the discriminant functions to interpret the differentiating dimensions.

LIMITATIONS AND EXTENSIONS

Limitations

Some potential limitations of LCA are noted below.

Local Maxima. LCA programs use iterative methods for maximum likelihood estimation. Sometimes algorithms converge on a local maximum rather than the global maximum solution; this is true of many statistical procedures. The simplest way to avoid local maximum solutions is to run a program several times using different parameter starting values and to select the best-fitting solution. Use of multiple start values can be included in the LCA software, making this process largely invisible to the user.

Identification. With an unidentified model, different parameter values account for the data equally well. The situation is analogous to having more unknowns than equations, resulting in an infinite number of solutions. LCA model identifiability requires that the number of possible rating patterns minus 1 is greater than or equal to the number of estimated parameters. This restricts the number of latent classes one can estimate for a given number of variables and rating levels. For example, given dichotomous items, a two-latent class model requires at least three items (even then, no df remain to assess model fit, so a more realistic minimum requirement in this case is four items). Unusual patterns of observed data sometimes may cause nonidentifiability; again, if this occurs, the main consequence is to limit the number of latent classes one can consider. Some LCA programs include the option to check model identifiability.

Number of Variables. With many variables and response levels, the number of possible response patterns can be very large. For example, with 10 items and 3 response levels each, over 59,000 response patterns are possible. Because of this, some LCA programs allow only a limited number of variables. The problem can be avoided or minimized if the estimation algorithm considers only rating patterns that actually are observed—usually far fewer than the number possible. This approach greatly extends the number of variables that can be used in an analysis.

Multiple Indicators. LCA assumes *conditional independence* of manifest variables. This stipulates that variables are independent within each latent class. For example, it requires that, within a given latent class, alcohol use is as common among those who use marijuana as among those who do not use marijuana. This assumption sometimes is difficult to justify, especially if two items are similar, such as, "Have you used marijuana in the last week?" and "Have you used marijuana in the last month?" LCA should produce useful results despite moderate violations of this assumption, although model fit may be decreased. Future versions of LCA may address this limitation.

Extensions

Extensions of the basic LCA approach, which some researchers may wish to consider, include multiple-group LCA, located latent class models, and mixed-mode measurement.

Multiple-Group LCA. As with structural equation modeling, one can estimate a latent class model simultaneously across two or more groups. By comparing models where one or more parameters are held constant across groups with models in which the parameters are free to vary, one can investigate group differences. For example, it might be useful to know if schools in different areas have the same basic latent classes but different proportions of students belonging to each.

Located Latent Class Models. Many recent authors have discussed located latent class models (Formann 1992; Lindsay et al. 1991; Rost 1988; Uebersax 1993). These models view latent classes as located on one or more underlying continua. With this approach, one can examine, for example, whether different latent classes correspond to increasing levels of overall substance use. Located latent class models also can help reduce the number of parameters that require estimation.

Mixed-Mode Measurement. With continuous measures, the counterpart of LCA is latent profile analysis (Lazarsfeld and Henry 1968). Latent class analysis for problems with mixed-mode measurement (e.g., combinations of dichotomous, ordered categorical, and continuous measures) is an area of ongoing research (Everitt and Merette 1990; Uebersax 1992).

The discussion here has assumed cross-sectional data. For discussion of extensions of LCA appropriate for longitudinal data, see Collins and colleagues (this volume).

SOFTWARE

At present, no major statistical package includes LCA. However, several stand-alone computer programs are available; most are written for personal computers. These programs include MLLSA (Clogg 1977), PANMARK (van de Pol et al. 1989), LAT and D-Newton (Haberman 1979), CGAGS (Hagenaars 1990), and LT-CLASS (Andersen 1990). The LTA program for latent transition analysis (Collins et al. 1992, this volume) also can be used to estimate the standard latent class model. Any of these programs will serve well for basic analyses.

Researchers considering more advanced or extensive use of LCA may wish to consider some of the following options in selecting software: (1) how many variables are allowed; (2) what input data formats are possible; (3) if model identifiability is checked; (4) if some parameters can be assigned fixed values or set equal to one another; (5) if multiple-group analysis is possible; (6) if standard errors of parameter estimates are calculated; (7) if recruitment probabilities are calculated; and (8) if variable and value labels are permitted.

CONCLUSIONS

In conclusion, LCA can be a useful data analysis tool for substance abuse prevention research. Its function is to assist the broader goal of developing a theoretical understanding of substance abuse and designing and implementing effective interventions. It is important not to reify the latent classes; they are best regarded as abstractions that help clarify variation in substance abuse in a population.

REFERENCES

Andersen, E.B. *The Statistical Analysis of Categorical Data.* Berlin: Springer-Verlag, 1990.

Bentler, P.M., and Bonett, D.G. Significance tests and goodness of fit in the analysis of covariance structures. *Psychol Bull* 88:588-606, 1980.

Clogg, C.C. "Unrestricted and Restricted Maximum Likelihood Latent Structure Analysis: A Manual for Users." Working Paper No. 1977-09. University Park, PA: The Pennsylvania State University, Population Issues Research Center, 1977.

Collins, L.M.; Wugalter, S.E.; and Rousculp, S.S. *LTA User's Guide*. Los Angeles: University of Southern California, J.P. Guilford Laboratory for Quantitative Psychology, 1992.

Day, N.E. Estimating the components of a mixture of normal distribution. *Biometrika* 56:463-474, 1969.

Everitt, B.S., and Merette, C. The clustering of mixed mode data: A comparison of possible approaches. *J Appl Stat* 17:283-297, 1990.

Formann, A.K. Linear logistic latent class analysis for polytomous data. *J Am Stat Assoc* 87:476-486, 1992.

Goodman, L.A. Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* 61:215-231, 1974.

Haberman, S.J. *Qualitative Data Analysis.* Vol. 2, *New Developments.* New York: Academic Press, 1979.

Hagenaars, J.A. *Categorical Longitudinal Data.* Newbury Park, CA: Sage, 1990.

Langeheine, R., and Rost, J., eds. *Latent Trait and Latent Class Models.* New York: Plenum, 1988.

Lazarsfeld, P.F. The logical and mathematical foundation of latent structure analysis. In: Stouffer, S.A.; Guttman, L.; Suchman, E.A.; Lazarsfeld, P.F.; Star, S.A.; and Clausen, J.A., eds. *Measurement and Prediction.* Princeton, NJ: Princeton University Press, 1950. pp. 362-412.

Lazarsfeld, P.F., and Henry, N.W. *Latent Structure Analysis.* Boston: Houghton Mifflin, 1968.

Lindsay, B.; Clogg, C.C.; and Grego, J. Semiparametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *J Am Stat Assoc* 86:96-107, 1991.

McCutcheon, A.C. *Latent Class Analysis.* Beverly Hills, CA: Sage, 1987.

Rost, J. Rating scale analysis with latent class models. *Psychometrika* 53:327-348, 1988.

Sclove, S. Application of model-selection criteria to some problems in multivariate analysis. *Psychometrika* 52:333-343, 1987.

Uebersax, J.S. "A Framework To Unify Latent Class Analysis and Multivariate Mixture Estimation." Working Paper No. 92-01. Winston-Salem, NC: Bowman Gray School of Medicine, Department of Public Health Sciences, 1992.

Uebersax, J.S. Statistical modeling of expert ratings on medical treatment appropriateness. *J Am Stat Assoc* 88:421-427, 1993.

van de Pol, F.; Langeheine, R.; and de Jong, W. *PANMARK User Manual.* Voorburg, The Netherlands: Central Bureau of Statistics, 1989.

Wolfe, J.H. Pattern clustering by multivariate mixture analysis. *Multivariate Behav Res* 5:329-350, 1970.

ACKNOWLEDGMENT

This research was supported in part by National Institute on Drug Abuse grant R01 DA7030.

AUTHOR

John S. Uebersax, Ph.D.
Associate Professor
Department of Public Health Sciences
Bowman Gray School of Medicine
Wake Forest University
Winston-Salem, NC 27157-1063

Latent Transition Analysis and How It Can Address Prevention Research Questions

*Linda M. Collins, John W. Graham,
Susannah Scarborough Rousculp, Penny L. Fidler,
Jia Pan, and William B. Hansen*

ABSTRACT

The objective of this chapter is to introduce latent transition analysis (LTA) to the substance use prevention research community. LTA is a new methodological technique for testing stage-sequential models, such as models of substance use onset. LTA estimates several different sets of parameters. One of these sets is the transition probability matrix, which contains information about the probability of movement between stages in the model. LTA can be used to evaluate the effectiveness of prevention intervention programs by comparing the transition probability matrices of the program and control groups. If the prevention program is successful, the transition probability matrices will indicate that the probability of moving to a more advanced stage of drug use is lower for the program participants than for the control group. An advantage of taking a stage-sequential approach is that examining the transition probability matrix reveals how effective a program is for individuals entering the program with different levels and types of substance use experience.

In this chapter, LTA is used to evaluate a variety of models of the early onset process separately for Anglo, Latino, and Asian-American adolescents, measured in seventh grade and again in eighth grade. Although somewhat different models are found to fit the three ethnic groups best, the differences likely are due to differences in the overall amount of substance use experience. Based on these results, it is suggested that, to be most effective, prevention programs should take place earlier for

Anglos and Latinos, and later, followed by boosters, for Asian Americans.

INTRODUCTION

A thorough understanding of the substance use onset process, and of diversity in this process, is important if prevention efforts to delay or halt onset are to be successful. One useful way to view the substance use onset process is as a stage sequence of substance use experiences (e.g., Yamaguchi and Kandel 1984). Methodology has existed for some time to test models of onset based on event history data, for example, reports of when a substance was tried. However, most school-based prevention researchers do not collect this kind of data because doing so is too labor intensive and because drug use data collected this way from adolescents are not very accurate (Collins et al. 1985). Instead, most school-based prevention efforts use longitudinal panel designs, in which data are collected at regular intervals and the emphasis is on the present, the recent past, or general lifetime use.

This chapter illustrates latent transition analysis (LTA), a methodology for estimating and testing stage-sequential models in longitudinal panel studies. The LTA model will be used to examine the nature and extent of ethnic group differences in early substance use prevalence and onset. Using LTA, it is possible to estimate the prevalence of the various stages in a model in a given sample and also to estimate the incidence of transitions between stages. These estimates are adjusted for measurement error, resulting in a more accurate picture of the onset process.

LATENT TRANSITION ANALYSIS (LTA)

The LTA model will be presented relatively briefly here; for a more complete presentation, see Collins and Wugalter (1992) and Graham and colleagues (1991).

LTA is a latent variable model for longitudinal panel data. By the term "latent variable model," researchers mean that they are measuring a theoretically error-free latent variable by means of fallible observed variables. In this study, the latent variable is substance use onset. It has been measured in seventh grade and again in eighth grade by four fallible observed variables: an alcohol item, a tobacco item, a drunkenness item, and an item indicating advanced use. In the LTA procedure, the latent variable has two important special features. First, it is *dynamic*; that is, individuals exhibit growth on this latent variable over time. Second, it is conceptualized as a *sequence of stages.* In LTA terminology, stages are referred to as "latent statuses."

Figure 1 depicts a substance use onset process discussed by Collins and colleagues (in press-*a*). This is an example of a dynamic stage-sequential latent variable. The latent statuses correspond to substance use experience and are denoted in the circles. In this model, individuals may begin their substance use experience by passing through any of a number of stage sequences, as depicted by the arrows in figure 1. For example, according to figure 1, some individuals begin their substance use experience with alcohol followed by either tobacco or an experience with drunkenness, while others begin with tobacco followed by alcohol. Only certain latent statuses will appear in a given model. There are eight latent statuses consistent with the model depicted in figure 1: "no use;" "tried alcohol;" "tried tobacco;" "tried alcohol and tobacco;" "tried alcohol, been drunk;" "tried alcohol, been drunk, advanced use;" "tried alcohol, tried tobacco, advanced use;" and "tried alcohol and tobacco, been drunk, advanced use." LTA models the transitions between latent status memberships across time.

The LTA Mathematical Model[1]

Suppose there are two occasions of measurement, with the first taken at Time t and the second at Time t+1. Further suppose there are four manifest indicators: item 1, with $i,i' = 1,\ldots I$ *response categories; item 2, with* $j,j' = 1,\ldots J$ *response categories; item 3, with* $k,k' = 1,\ldots K$ *response categories; and item 4, with* $l,l' = 1,\ldots L$ *response categories, where* $i, j, k,$

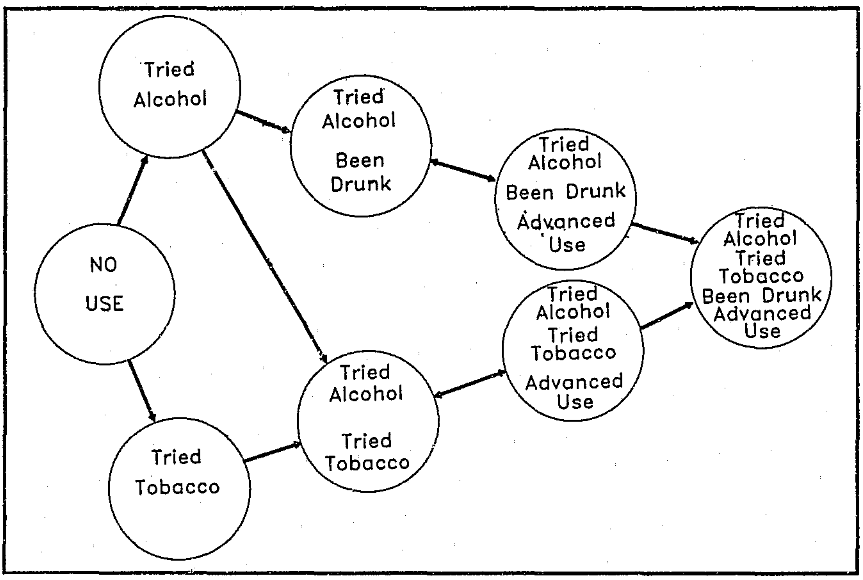

Tried Alcohol
Tried Alcohol Been Drunk
Tried Alcohol Been Drunk Advanced Use
Tried Alcohol Tried Tobacco Been Drunk Advanced Use
NO USE
Tried Alcohol Tried Tobacco Advanced Use
Tried Alcohol Tried Tobacco
Tried Tobacco

FIGURE 1. *Stage-sequential model of substance use onset discussed in Collins and colleagues (in press)*

and l *refer to responses obtained at Time* t, *and* i', j', k', and l' *refer to responses obtained at Time* $t+1$. *(For example, in the substance use research that will be described here, the following manifest indicators were used: an alcohol use item, a tobacco use item, an item asking about drunkenness, and an advanced use item that was a composite of several substance use items. Data were collected in seventh grade and again in eighth grade.) The extension to more than two occasions, fewer than four indicators, or more than four indicators is direct. There are* $a,b = 1,...S$ *latent statuses, with* a *denoting a latent status at Time* t *and* b *denoting a latent status at Time* $t+1$.

Let $Y = \{i,j,k,l,i',j',k',l'\}$ *represent a "response pattern," a vector of possible responses made up of a single response to the manifest indicator of the exogenous variable and responses to the four items at Times* t *and* $t+1$. *Then the estimated proportion of a particular response pattern,* $P(Y)$, *is expressed as equation (1) on the following page.*

$$P(Y) = \sum_{a=1}^{S} \sum_{b=1}^{S} \delta_a \rho_{i|a} \rho_{j|a} \rho_{k|a} \rho_{l|a} \tau_{b|a} \rho_{i'|b} \rho_{j'|b} \rho_{k'|b} \rho_{l'|b} \qquad (1)$$

Parameters Estimated in the LTA Model

In the LTA models discussed in this chapter, three different types of parameters are estimated:

δ_a represents the proportion in latent status *a* at Time *t*; in other words, this parameter is the estimated proportion of subjects in each latent status at the first occasion of measurement. Using the latent variable in figure 1, an example would be the estimated proportion of individuals who at Time *t* have used tobacco only.

$\tau_{b|a}$ is a transition probability representing the probability of membership in latent status *b* at Time *t*+1, conditional on membership in latent status *a* at Time *t*. These parameters represent the probability of moving to a particular latent status at the second occasion of measurement, conditional on latent status membership at the first occasion. In figure 1, one example of a transition probability would be the probability of moving to the "alcohol and tobacco" latent status at the second occasion, given membership in the "alcohol only" latent status at the first occasion. The transition probability matrix is latent, that is, adjusted for error in the observed items. The transition probabilities usually are arranged in a matrix like the one below:

$$\begin{bmatrix} \tau_{1|1} & \tau_{2|1} & \tau_{3|1} & \cdots \\ \tau_{1|2} & \tau_{2|2} & \tau_{3|2} & \cdots \\ \tau_{1|3} & \tau_{2|3} & \tau_{3|3} & \cdots \\ \vdots & \vdots & \vdots & \end{bmatrix}$$

where $\tau_{b|a}$ represents the probability of membership in stage b at the end of the interval, given membership in stage a at the beginning of the interval. Because the elements of the matrix are conditional probabilities, each row of this matrix sums to unity.

LTA also estimates measurement parameters. $\rho_{i|a}$ represents the probability of response i to item 1 at Time t, conditional on membership in latent status a at Time t; $\rho_{i'|b}$ represents the probability of response i' to item 1 at Time t+1, conditional on membership in latent status b at Time t+1; etc. In other words, these parameters assess the degree of error in each observed item. The ρ's play two roles in LTA models. First, they map the manifest items onto the latent statuses in much the same way that factor loadings map variables onto factors. For example, if the probability of responding no to each of the substance use items is high for a particular latent status, this would be interpreted as a "no substance use" latent status. If, in another latent status, the probability of responding yes is high for the alcohol item while the probability of responding no is high for the remaining items, this latent status would be interpreted as "tried alcohol only." The second role that the ρ's play is in reflecting measurement precision. If measurement is error free, each manifest response is determined completely by latent status membership, and all the ρ's are 0 or 1. In general, the closer these parameters are to 0 or 1 for a particular item, the closer the relationship between latent status membership and manifest responses.

Comparison of LTA and Covariance Structure-Modeling

There are many analogies between LTA and covariance structure-modeling (Jöreskog and Sörbom 1989). Both are latent variable models where fallible observed variables serve as indicators of error-free unmeasured variables. Both procedures involve a measurement model that maps the observed variables onto the latent variables. In covariance structure models, the latent variable is continuous and usually is measured by continuous indicators, whereas LTA involves discrete latent variables and indicators. In covariance structure models, factor loadings provide the link between observed and unmeasured variables; in LTA, the measurement parameters serve this purpose. However, the LTA measurement parameters cannot be interpreted in exactly the same way as factor loadings. With factor loadings, a large absolute value is a strong loading, while a value close to 0 indicates no relationship, or a very weak relationship, between a variable and a factor. In contrast, LTA measurement parameters are estimates of probabilities, so a value near 0 or near unity indicates "sureness," or a strong relationship between a measured variable and a latent variable. A value close to *1/J,* where *J* is the number of response alternatives, indicates no relationship between a measured variable and an observed variable. Negative values are impossible.

USING LTA TO INVESTIGATE ETHNIC DIFFERENCES IN ADOLESCENT SUBSTANCE USE ONSET

There is a growing body of evidence that ethnic differences in adolescent substance use prevalence are genuine, reliable, and substantial. Perhaps the most compelling evidence comes from Bachman and colleagues (1991), who conduct the Monitoring the Future project. This project has surveyed nationally representative samples of high school seniors yearly since 1975. The surveys have revealed consistently that Native Americans have the highest prevalence rates for most substances, followed by Anglos; that Latinos show intermediate prevalence rates; and that Asian Americans show the lowest substance use rates, with African Americans showing only slightly higher use.

This general finding has been replicated in a variety of settings by numerous other studies. Oetting and Beauvais (1990) found results remarkably similar to those reported in Bachman and colleagues (1991) in their American Drug and Alcohol Survey, which is based on a nationwide nonrandom sample. Both Welte and Barnes (1987), based on a large random sample of junior high and high school students from New York, and Brannock and colleagues (1990), based on a smaller sample from two high schools and one college in southern California, found results consistent with those found by Bachman and colleagues (1991). Grady and colleagues (1986), using a sample of New England seventh and eighth graders, found that Anglos showed greater use of tobacco, alcohol, and marijuana than African Americans. Graham and colleagues (1990) followed three successive cohorts of southern California students from seventh grade through eighth grade. Their results were consistent with those found by Bachman and colleagues (1991) and also suggested that their substance use prevention program was less effective for Anglos than it was for minorities. There is a considerable body of older research that is consistent with these findings (e.g., Engs 1977; Humphrey and Friedman 1986; Humphrey et al. 1983; Kandel et al. 1976; McIntosh et al. 1979; Walfish et al. 1981; Wechsler and McFadden 1979) despite the documented changes in overall trends in adolescent substance use over the last decade.

Because an individual arrives at a level of substance use experience after going through an onset process, the finding that there are ethnic differences in substance use prevalence raises the important question of whether there are ethnic differences in this substance use onset process as well. Such differences may take one of two forms. One possibility is that the onset process essentially is the same across ethnic groups, but onset begins earlier and/or the process is accelerated for certain groups. Alternatively, the onset process itself may be qualitatively different for different ethnic groups. If so, there may be differences in time and rate of onset, but direct comparisons between groups at best can be limited when the process itself differs.

The Substance Use Onset Process

The stage-sequential point of view on substance use onset was pioneered by Yamaguchi and Kandel (1984), who examined the onset process from tenth grade through early adulthood. They found that use of alcohol and/or cigarettes preceded marijuana use and that marijuana use was a necessary precursor to use of other illicit drugs. Graham and colleagues (1991) used a longitudinal panel design to test several models of early substance use onset. Their subjects were in seventh grade at the first wave of data collection and eighth grade at the second wave. Graham and colleagues (1991) found that the best-fitting model was one in which most subjects initiated their substance use with alcohol followed by tobacco, but an important minority of subjects initiated their substance use with tobacco followed by alcohol. Next was a first experience with drunkenness, followed by advanced use (defined as regular use of alcohol, regular use of tobacco, or any experience with marijuana).

In the present study, the researchers tested five models of substance use onset using a larger sample of which the Graham and colleagues (1991) sample is a subset. Because the researchers were interested in ethnic differences in onset, the models were tested separately for Anglos, Latinos, and Asian Americans.

METHODS

Subjects

The subjects for this study completed a drug use survey as seventh graders in either fall 1987 or fall 1988 and again as eighth graders 1 year later as part of the Adolescent Alcohol Prevention Trial (Graham et al. 1989; Hansen and Graham 1991; Hansen et al. 1988). The study participants were those Anglos, Latinos, or Asian Americans who had complete data for relevant measures on both pretest and posttest; the participants were taken from a sample of seventh graders (N = 5,242) who completed the survey at pretest. The subsample used in this study contains 1,443 Anglos, 1,185 Latinos, and 498 Asian Americans.

Measures

The measures used in this study included lifetime alcohol use (How many drinks of alcohol have you had in your whole life?); lifetime cigarette use (How many cigarettes have you smoked in your whole life?); and lifetime drunkenness (How many times have you ever been drunk?). The alcohol item was coded 0 if the subject reported "no use" or "sips for religious services" and was coded 1 for "sips (not for religious services)" or more in his or her lifetime. The cigarette item was coded 0 for "never tried" and 1 for "one puff" or more in his or her lifetime. The drunkenness item was coded 0 for "never been drunk" and 1 for "been drunk once" or more.

Several other measures were used in the analyses reported in this chapter, including alcohol use in the previous month and previous week, tobacco use in the previous month and previous week, and lifetime marijuana use. Models involving these items separately showed considerable instability. It appeared that much of the instability stemmed from the fact that these were young adolescents with very low levels of use. Thus, these items tapping greater involvement with various substances were combined into a single composite item reflecting advanced use. The combined item was scored 0 if the subject had engaged in no alcohol use and no tobacco use in the previous week and the previous month and had never used marijuana; otherwise, it was coded 1.

Models Under Consideration

In this study, the researchers specified five models to be tested using LTA. Figure 2 depicts all of these models, with different types of arrows indicating which path is featured in a particular model. All of the models specify that the onset process may begin with alcohol or with tobacco followed by alcohol. Model 1, the model depicted in figure 1, is the most parsimonious of the five models. This model suggests that for those in the "tried alcohol, tried tobacco" latent status and those in the "tried alcohol, been drunk" latent status, the next transition is into a "tried alcohol, tried tobacco, been drunk" latent status. This model suggests an orderly progression of increasing involvement where alcohol, tobacco, and then drunkenness occur before advanced use. Model 2 eliminates the "tried alcohol, tried tobacco, been drunk" latent status, involving instead transitions to a "tried alcohol, been drunk, advanced use" latent status or a "tried alcohol, tried tobacco, advanced use" latent status. Model 3 adds a latent status to model 1, suggesting the existence of a "tried alcohol, been drunk, advanced use" latent status. This allows for the possibility of engaging in advanced use (of alcohol or marijuana) before having tried

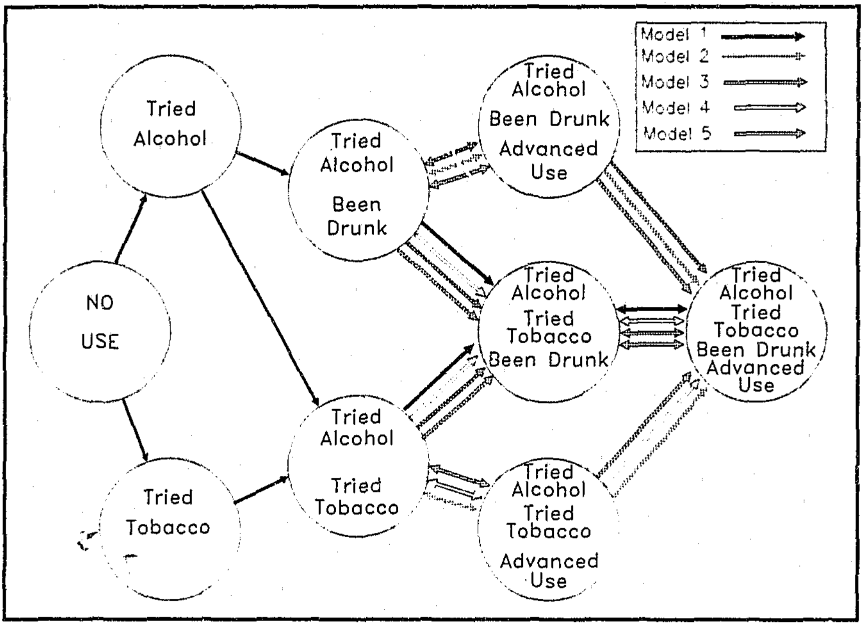

Model 1
Model 2
Model 3
Model 4
Model 5
NO USE
Tried Alcohol
Tried Tobacco
Tried Alcohol Been Drunk
Tried Alcohol Tried Tobacco
Tried Alcohol Been Drunk Advanced Use
Tried Alcohol Tried Tobacco Been Drunk
Tried Alcohol Tried Tobacco Advanced Use
Tried Alcohol Tried Tobacco Been Drunk Advanced Use

FIGURE 2. *Models considered in the present study*

SOURCE: Collins, L.M.; Graham, J.W.; Long, J.; and Hansen, W.B. Crossvalidation of latent class models. *Multivariate Behavioral Research,* in press.

tobacco. Model 4 includes the "tried alcohol, tried tobacco, been drunk" latent status and the "tried alcohol, tried tobacco, advanced use" latent status. Both model 2 and model 4 suggest that it is possible to proceed to advanced use without having been drunk. Finally, model 5, the most complex of the five models, includes all of the paths and latent statuses involved in models 1, 2, 3, and 4.

Evaluating the Models

Typically the fit of LTA models is evaluated using the likelihood ratio statistic, G^2. For fixed degrees of freedom, a smaller G^2 indicates a better fit of the model being tested to the data. Hypothesis-testing can be used to aid in model selection. However, it is well known that the *p*-values associated with G^2 are very inaccurate for models like LTA (Collins et al.

1993; Holt and Macready 1989; Read and Cressie 1988). As an alternative to relying on these *p*-values, the authors have taken a cross-validation approach (Collins et al., in press-*b*; Cudeck and Browne 1983). They split the sample randomly into two samples that will be referred to as sample A and sample B and fit each model in sample A, estimating all relevant parameters. In order to assess goodness of fit, the authors computed G^2 for the fit of each sample A model *in the sample B data.* They then reversed the process, fitting each model in sample B and then computing G^2's based on sample A. This is known as double cross-validation. Ideally, this procedure will point clearly to a single model that has a low cross-validation G^2 in both samples; in practice, the results usually are not so clear cut. When the results were ambiguous in this study, the authors chose the most parsimonious models.

RESULTS

Statistical analyses were performed using the software LTA (Collins et al., in press-*a*). In order to achieve model identification, some parameters were constrained to remain equal to each other where it made conceptual sense to do so. The LTA program requires the user to input initial parameter estimates to be used as "start values" to begin the estimation procedure. If a model is identified, the choice of start values usually has little or no impact on the final solution. As is consistent with good practice when estimating latent class models, two very different sets of start values were used for each model in this study. In 25 out of 30 analyses, the results were virtually identical. Small differences between the two solutions occurred in model 4 for both subsamples of Anglos and both subsamples of Latinos and in model 3 for one of the Latino subsamples.

Model Selection

Table 1 shows the cross-validation G^2's for each of the LTA models that was estimated in each subsample. For Anglos, model 4 cross-validates best in one sample, but model 5 cross-validates best in the other sample. Model 2, although it does not cross-validate best in either sample, cross-validates second best in both samples; thus, model 2 is chosen for Anglos. For Latinos, model 5 cross-validates consistently well; it is the best in one sample and the second best in the other. For Asian Americans, models 1 and 2 cross-validate best in both samples; the authors choose model 1 because it is most parsimonious.

Table 2 contains the ρ parameters for the Anglo sample. These parameters represent the probabilities of a yes response, conditional on latent status membership. As discussed above, the values of these parameters are what determines the interpretation of the latent statuses. For those individuals in the first latent status, the probability of responding yes to ANY of the substance use items is extremely low. Thus, the first latent status is interpreted as a "no use" latent status. For those in the second latent status, the probability of responding yes to the alcohol item is large, but the probability of responding yes to any other items is small. Thus, this latent status is interpreted as "alcohol use only." Similarly, the third latent status is interpreted as "tobacco use only," the fourth as "alcohol and tobacco," the fifth as "alcohol and drunkenness," the sixth as "alcohol, drunkenness, and advanced use," the seventh as "alcohol, tobacco, and advanced use," and the last as "alcohol, tobacco, drunkenness, and advanced use."

The overall structure of these parameters cross-validates well; in other words, the same interpretation of the latent statuses is indicated in both samples. Also, in general, these parameters are above .75 or below .25, indicating a strong relationship between the items and the latent statuses. Where the manifest items are dichotomous, as they are here, a parameter estimate close to .5 suggests that the item in question is not a good indicator of latent status membership. The weakest relationship in these data between an item and latent statuses is the relationship between the advanced use indicator and the last three latent statuses.

Table 3 shows the transition probability matrix for the Anglo sample. Sample A estimates are in the first line in each row, and sample B estimates are in the second line. The elements on the diagonal of each matrix represent probabilities of being in the same latent status in both

TABLE 1. *Results of applying five models to ethnic subsamples*

Model fitted to sample A, G^2 on sample B

	Ethnicity		
Model	Anglos	Latinos	Asians
1	198.2	225.1	190.6
2	177.5	212.5	196.2
3	202.8	234.1	213.6
4	174.9	201.5	192.0
5	178.5	207.8	206.4

Model fitted to sample B, G^2 on sample A

	Ethnicity		
Model	Anglos	Latinos	Asians
1	222	253.3	131.5
2	196.7	239.9	127.5
3	205.5	243.8	164.4
4	209.1	241.2	137.6
5	196.2	229.3	138.9

seventh grade and eighth grade, and the elements on the off-diagonal represent probabilities of transitioning to the column latent status, conditional on membership in the row latent status. For example, for

TABLE 2. *Measurement parameters (ρ's) for Anglos*

Latent status		Probability of responding yes to these items, conditional on latent status membership			
	Sample	Ever Tried Tobacco ?	Ever Tried Alcohol ?	Ever been drunk ?	Any advanced use ?
No use	A	.03	.00	.02	.01
	B	.03	.03	.02	.02
Alcohol Use Only	A	.03	.97	.02	.01
	B	.03	.98	.02	.02
Tobacco Use Only	A	.97	.00	.02	.01
	B	.93	.00	.02	.02
Alcohol+Tobacco	A	.97	.97	.02	.01
	B	.93	.98	.02	.02
Alcohol+Drunkenness	A	.03	.97	.79	.01
	B	.03	.98	.90	.02
Alcohol, Drunkenness, Advanced Use	A	.03	.97	.79	.66
	B	.03	.98	.90	.62
Alcohol, Tobacco, Advanced	A	.97	.97	.02	.66
	B	.93	.98	.02	.62
Alcohol, Tobacco, Drunkenness, Advanced	A	.97	.97	.79	.66
	B	.93	.98	.90	.62

those Anglos who start out in the "no use" latent status in seventh grade in sample A, it is estimated that the probability is .58 (in sample B, .55) of being there in eighth grade.

In estimating model parameters, the authors chose to estimate full transition probability matrices, as opposed to fixing the lower triangle (all except the transition from the most extreme latent status back to "alcohol, tobacco, and drunkenness") to 0's. The rationale for fixing the lower triangle to 0's would be that these transitions are impossible in theory. For example, it is impossible to transition from having tried alcohol to having never tried alcohol. On the other hand, although these transitions are impossible, subjects nevertheless respond as if they were possible. Estimating these transitions can give a very useful picture of the kinds of response biases that are operating in a sample to produce these kinds of responses. In several cases, fairly large lower-triangle elements were estimated. However, in general, these parameter estimates did not cross-validate well in these data.

Table 4 shows the estimates of the ρ parameters for the Latino sample, and table 5 shows the transition probability matrix. Table 4 shows that the model that cross-validated best for the Latino sample is similar to the model selected for the Anglo sample, with the addition of an "alcohol, tobacco, drunkenness" latent status.

Table 6 contains the estimated ρ parameters for the Asian-American sample. The parameter estimates for the first four latent statuses based on the Asian-American sample lead to the same interpretation as their counterparts in the Anglo and Latino samples. However, the ρ parameters suggest very different interpretations for the last three latent statuses. For the Asian-American sample, the fifth latent status essentially is similar to the second latent status, and the sixth latent status essentially is similar to the fourth latent status. The only difference is that the fifth and sixth latent statuses involve a somewhat higher probability of responding yes to the drunkenness item, although a no response to this item still is more likely than a yes. The last latent status involves alcohol, tobacco, and advanced use only.

These results illustrate why it is very important to examine the ρ parameter estimates carefully when interpreting and labeling the latent statuses. LTA is a confirmatory procedure in the sense that the user must specify certain important aspects of a model like the number of latent statuses and any constraints on parameter estimates. Generally, a user who specifies

TABLE 3. *Transition probability matrix for Anglos*[1]

Latent status	Sample	Latent status							
No use	A	**.58**	**.30**	**.00**	**.03**	**.01**	**.00**	**.03**	**.04**
	B	.55	.31	.02	.05	.01	.00	.04	.02
Alcohol Use Only	A	**.07**	**.74**	**.00**	**.10**	**.01**	**.03**	**.01**	**.04**
	B	.08	.74	.00	.03	.03	.00	.05	.07
Tobacco Use Only	A	**.04**	**.00**	**.34**	**.53**	**.00**	**.00**	**.00**	**.09**
	B	.00	.00	.22	.40	.00	.00	.19	.18
Alcohol+Tobacco	A	**.00**	**.00**	**.00**	**.63**	**.00**	**.00**	**.04**	**.32**
	B	.00	.00	.00	.70	.00	.00	.14	.14
Alcohol+Drunkenness	A	**.00**	**.00**	**.00**	**.00**	**.48**	**.08**	**.00**	**.45**
	B	.00	.00	.00	.00	.22	.21	.00	.55
Alcohol, Drunkenness, Advanced Use	A	**.00**	**.00**	**.00**	**.00**	**.00**	**.69**	**.00**	**.30**
	B	.00	.00	.00	.00	.32	.66	.00	.00
Alcohol, Tobacco, Advanced Use	A	**.00**	**.00**	**.00**	**.20**	**.00**	**.00**	**.79**	**.00**
	B	.00	.00	.00	.09	.00	.00	.54	.35
Alcohol, Tobacco, Drunkenness, Advanced Use	A	**.00**	**.00**	**.00**	**.00**	**.00**	**.00**	**.00**	**.99**
	B	.00	.00	.00	.00	.00	.00	.00	.97

NOTE: [1] Some rows do not sum to 1 because of rounding.

the number of latent statuses will have particular values of the ρ parameters in mind. However, it is important to examine the estimates of the ρ parameters that result from an LTA analysis, because these estimates may be different from what is hypothesized and may lead to different interpretations of the latent statuses. In the present study, although a solution involving seven latent statuses cross-validated the best for the Asian-American sample, the model as estimated is different from the model 1 depicted in figure 2. Rather than emerging as conceptually distinct latent statuses as depicted in figure 2, the fifth and sixth latent

TABLE 4. *Measurement parameters (ρ's) for Latinos*

Latent status	Sample	Probability of responding yes to these items, conditional on latent status membership			
		Ever Tried Tobacco ?	Ever Tried Alcohol ?	Ever been drunk ?	Any advanced use ?
No use	A	**.00**	**.00**	**.00**	**.01**
	B	.04	.04	.02	.02
Alcohol Use Only	A	**.00**	**.95**	**.00**	**.02**
	B	.04	.97	.02	.02
Tobacco Use Only	A	**.98**	**.00**	**.00**	**.02**
	B	.97	.04	.02	.02
Alcohol+Tobacco	A	**.98**	**.95**	**.00**	**.02**
	B	.97	.97	.02	.02
Alcohol+Drunkenness	A	**.00**	**.95**	**.77**	**.02**
	B	.04	.97	.85	.02
Alcohol, Drunkenness, Advanced Use	A	**.00**	**.95**	**.77**	**.91**
	B	.04	.97	.85	.83
Alcohol, Tobacco, Advanced Use	A	**.98**	**.95**	**.00**	**.91**
	B	.97	.97	.02	.83
Alcohol, Tobacco, Drunkenness	A	**.98**	**.95**	**.77**	**.02**
	B	.97	.97	.85	.02
Alcohol, Tobacco, Drunkenness, Advanced Use	A	**.98**	**.95**	**.77**	**.91**
	B	.97	.97	.85	.83

statuses conceptually are very similar to the second and third latent statuses, respectively. Moreover, the seventh latent status as estimated in the Asian-American subsample does not involve a high probability of

TABLE 5. *Transition probability matrix for Latinos*

Latent status	Sample	Latent status								
No use	A	**.46**	**.26**	**.07**	**.10**	**.03**	**.01**	**.00**	**.05**	**.04**
	B	.64	.20	.00	.06	.01	.01	.03	.01	.05
Alcohol Use Only	A	**.15**	**.61**	**.01**	**.11**	**.03**	**.00**	**.02**	**.03**	**.05**
	B	.04	.66	.02	.13	.02	.00	.05	.03	.05
Tobacco Use Only	A	**.13**	**.01**	**.40**	**.26**	**.00**	**.00**	**.00**	**.13**	**.06**
	B	.00	.02	.33	.34	.06	.00	.06	.09	.21
Alcohol+Tobacco	A	**.01**	**.01**	**.01**	**.65**	**.01**	**.01**	**.18**	**.03**	**.09**
	B	.02	.02	.02	.61	.02	.02	.15	.05	.11
Alcohol+Drunkenness	A	**.01**	**.01**	**.01**	**.01**	**.37**	**.20**	**.01**	**.22**	**.15**
	B	.02	.02	.02	.02	.35	.05	.02	.16	.36
Alcohol, Drunkenness, Advanced Use	A	**.01**	**.01**	**.01**	**.01**	**.35**	**.20**	**.01**	**.38**	**.00**
	B	.02	.02	.02	.02	.47	.32	.02	.13	.00
Alcohol, Tobacco, Advanced Use	A	**.01**	**.01**	**.01**	**.45**	**.01**	**.01**	**.15**	**.00**	**.33**
	B	.02	.02	.02	.36	.02	.02	.16	.06	.34
Alcohol, Tobacco, Drunkenness	A	**.01**	**.01**	**.01**	**.01**	**.01**	**.01**	**.01**	**.53**	**.38**
	B	.02	.02	.02	.02	.02	.02	.02	.47	.43
Alcohol, Tobacco, Drunkenness, Advanced Use	A	**.01**	**.01**	**.01**	**.01**	**.01**	**.01**	**.01**	**.08**	**.84**
	B	.02	.02	.02	.02	.02	.02	.02	.00	.89

having experienced drunkenness. This leaves the interpretation of the last three latent statuses unclear. A partial transition probability matrix for the Asian-American sample appears in table 7. Because the meaning of the last three latent statuses is unclear, transitions involving these latent statuses are difficult to interpret, so they are omitted from the table.

Table 8 shows the estimates of the δ parameters, which are the proportions in each latent status in seventh grade. For the Asian-American sample, the δ estimates for the second and fifth latent statuses and for the fourth and sixth latent statuses are collapsed because of their similarity.

TABLE 6. *Measurement parameters (ρ's) for Asian Americans*

Latent status		Probability of responding yes to these items, conditional on latent status membership			
	Sample	Ever Tried Tobacco ?	Ever Tried Alcohol ?	Ever been drunk ?	Any advanced use ?
No use	A	**.00**	**.10**	**.01**	**.01**
	B	.00	.06	.00	.01
Alcohol Use Only	A	**.00**	**.97**	**.01**	**.01**
	B	.00	.88	.00	.01
Tobacco Use Only	A	**.98**	**.10**	**.01**	**.01**
	B	.87	.06	.00	.01
Alcohol+Tobacco	A	**.98**	**.97**	**.01**	**.01**
	B	.87	.88	.00	.01
Alcohol (+Drunkenness)	A	**.00**	**.97**	**.39**	**.01**
	B	.00	.88	.48	.01
Alcohol, Tobacco (+Drunkenness)	A	**.98**	**.97**	**.39**	**.01**
	B	.87	.88	.48	.01
Alcohol, Tobacco, Advanced Use (+Drunkenness)	A	**.98**	**.97**	**.39**	**1.00**
	B	.87	.88	.48	1.00

The parameter estimates are very close across sample A and sample B for Anglos and Latinos, indicating good cross-validation. The estimates based on the Asian-American sample do not cross-validate as well, although the general pattern of results is consistent across the two subsamples. The results show that, as expected, Anglos are the least likely to be abstainers, even in this early phase of onset. However, they

TABLE 7. *Partial transition probability matrix for Asian Americans*

Latent status	Sample	Latent status						
No use	A	**.83**	**.09**	**.04**	**.04**	**.00**	**.00**	**.00**
	B	.78	.03	.01	.11	.01	.04	.03
Alcohol Only	A	**.00**	**.54**	**.03**	**.14**	**.29**	**.00**	**.00**
	B	.06	.73	.00	.15	.02	.04	.00
Tobacco Only	A	**.38**	**.03**	**.34**	**.16**	**.00**	**.00**	**.10**
	B	.20	.00	.71	.00	.00	.08	.00
Alcohol+Tobacco	A	**.03**	**.03**	**.03**	**.61**	**.03**	**.23**	**.03**
	B	.00	.00	.00	1.00	.00	.00	.00

are least likely by only a small margin. Only approximately 28 percent of Anglos have never tried alcohol or tobacco, as opposed to 31 and 33 percent for Latinos. This difference seems to be due mostly to the relatively large percentage of Anglos who have tried alcohol but have engaged in no further experimentation. The probability of having gone no further than trying a single substance can be obtained by summing the probabilities of membership in the "no use," "alcohol only," and "tobacco only" latent statuses. This shows that the probability of having gone no further than trying a single substance is .65 for Anglos and .77 and .80 for the Asian-American subsamples but is .58 for Latinos. Thus, although the Latino sample contains a slightly higher proportion of abstainers than does the Anglo sample, those Latinos who have tried a substance are likely to have engaged in comparatively more experimentation.

DISCUSSION

Using LTA, the authors have tested several stage-sequential models of the early substance use onset process in three different ethnic groups. Each of these models represented the onset process as a dynamic latent variable measured by four manifest variables. LTA was used to identify

TABLE 8. *Estimates of proportions in each latent status at first occasion*

Latent status	Anglo sample		Latino sample		Asian American sample	
	A	B	A	B	A	B
No use	.28	0.28	.31	0.33	.59	.49
Alcohol Only	.35	0.34	.22	0.22	.12	.22
Tobacco Only	.02	0.03	.05	0.03	.09	.06
Alcohol+Tobacco	.13	0.13	.18	0.16	.18	.19
Alcohol+Drunkenness	.03	0.02	.04	0.02		
Alcohol, Drunkenness, Advanced Use	.02	0.01	.01	0.01		
Alcohol, Tobacco, Advanced Use	.03	0.06	.04	0.07	.04	.04
Alcohol, Tobacco, Drunkenness			.06	0.06		
Alcohol, Tobacco, Drunkenness, Advanced Use	.09	0.13	.10	0.09		

the latent statuses in each model and to provide estimates of the probabilities of membership in each latent status in seventh grade and the conditional probabilities of transitions between latent statuses between seventh grade and eighth grade. These probabilities are adjusted for measurement error occurring in the manifest variables.

Upon first examination, the results of this study suggest that somewhat different onset processes may be operating in Anglo, Latino, and Asian-American samples. The authors found that a model involving nine latent statuses was necessary for Latinos; that a slightly less complex model, omitting the "alcohol, tobacco, drunkenness" latent status, was sufficient to represent the data collected on Anglo subjects; and that, although the simplest model tested here fit the Asian-American sample best, even that proved too complex when two sets of two latent statuses emerged as virtually identical.

However, upon closer examination it seems that the similarities among these ethnic groups in the onset process outweigh the differences. The models are virtually identical in the early phases of the onset process. According to all three models, most individuals initiate their substance use experience with alcohol. However, a small but significant proportion initiate their experience with tobacco. Graham and colleagues (1991) found that this latter group of individuals was on an accelerated onset trajectory compared to those who start with alcohol. That finding seems to hold here for Anglos and Latinos. For Asian Americans, sample A estimates are consistent with this, but the finding does not replicate in sample B. The question of whether Asian Americans who start the onset process with tobacco are on an accelerated onset trajectory is an important one because, according to these results, Asian Americans are more likely to begin the onset process with tobacco than are Anglos or Latinos.

Another interesting feature shared by all three models is the important role that tobacco plays in the remainder of the onset process. These results indicate that in the Anglo and Latino samples, relatively few individuals went on to advanced use without trying tobacco and that, in the Asian-American sample, trying tobacco was an integral part of the early onset process. Drunkenness plays a major role in the onset process for both Anglos and Latinos. Drunkenness is not a major part of the onset model that represents the Asian-American sample in this study.

Although it is possible that the differences in onset process models among ethnic groups reflect real qualitative differences, in this case there is an alternative explanation. The differences that have emerged among the ethnic groups may have to do primarily with how advanced the onset process is. In any stage-sequential process, differentiation among stages cannot take place until enough subjects have passed through the stages. The authors' results indicate that the Asian-American subsample had considerably less substance use experience at the first observation than the other two subsamples. The results also show that the Asian-American subsample advanced through the onset process at a considerably slower rate, as reflected in the transition probability matrix. This may account for the lack of involvement of drunkenness in the Asian-American onset process—too few of the Asian Americans in the sample had arrived at that point in the onset process at that time. Perhaps, in an Asian-American sample with more substance use experience, a model more like model 2 or model 5 would be necessary to represent the onset process. Although the Anglo group has the smallest proportion of abstainers at the outset, Anglos who have initiated the onset process tend to have somewhat less substance use experience than Latinos who have initiated the process. It may be that in the Latino sample sufficient subjects had engaged in various onset activities for the authors to differentiate nine latent statuses. The Anglo and Latino subsamples are advancing through the onset process at comparable rates. Perhaps, if a little bit more time were allowed to elapse, the additional latent status would emerge in the Anglo sample. This seems likely, given that the more complex model, model 5, cross-validated nearly as well as model 2 in the Anglo sample.

Implications for Prevention

The degree and kind of ethnic differences found in this study have implications for planning prevention curricula. The result that the onset process essentially is comparable across groups, although there are some differences, offers hope that a single prevention curriculum can be effective for Anglos, Latinos, and Asian Americans. However, the comparability of the onset process across ethnic groups does not guarantee that the psychosocial factors prompting transitions between latent statuses also are comparable. If these factors are different, this will have to be taken into account in prevention programs.

Ideally, a prevention intervention should occur just before onset is expected. The results of this study suggest that the optimal timing of an intervention may vary according to the ethnic composition of the target population. Results indicate that 72 percent of Anglo seventh graders and 67-69 percent of Latino seventh graders already have initiated the onset process. Thus, for these ethnic groups, interventions probably should start earlier than seventh grade. In contrast, considerably fewer Asian seventh graders have started the onset process, and, furthermore, the process seems to be slower for this ethnic group. The results of the study by Graham and colleagues (1990), which showed a trend for stronger program effects among Asian-American students, indicate that perhaps seventh grade is a good time for beginning interventions on this subpopulation. The present study suggests that interventions should start earlier for Anglo and Latino students. Because the onset process is slower for Asian Americans, taking place over a long timespan, periodic boosters may be needed particularly with this group.

Limitations of This Study

An obvious limitation of this study is the lack of African-American and Native-American subjects. There were no Native Americans in this sample and far too few African Americans (fewer than 75) to test the models of interest in this study. A second important criticism of this study and, by implication, many other studies that have looked at ethnic differences in substance use, is the way in which the authors and most researchers measure ethnicity. As Cheung (1991) has pointed out, ethnicity is a multidimensional construct that cannot be captured well in a single variable. Furthermore, many observed "ethnic" differences undoubtedly are due to differences on a constellation of other variables, such as attitudes, educational levels, and socioeconomic status, for which ethnicity serves as a rough proxy. Yet, in most studies (including this one), ethnicity is measured by a single manifest variable. This approach obviously cannot capture the complexity and richness of ethnicity. Where ethnicity is measured poorly, some ethnic differences will be obscured, and observed differences will be subject to misinterpretation. At the very least, understanding the culture and social norms operating in various ethnic groups and how they relate to substance use onset is far more important than merely noting ethnic differences. There is a need for further research on ethnicity and early substance use onset using more sophisticated measures of ethnicity.

More About Latent Transition Analysis (LTA)

There are several features of LTA that have not been discussed in the present chapter. Where data have been collected on three or more occasions, second-order models can be tested in which transitions between latent statuses depend not only on latent status membership at the immediately previous time but on membership at two times previous as well. The LTA approach can incorporate a discrete exogenous grouping variable. This means it can be used to test multiple-groups models, in which the grouping variable either is manifest or latent. For more information, refer to Collins and Wugalter (1992) and to the *LTA User's Guide* (Collins et al., in press-*a*).

LTA's capability to incorporate a discrete exogenous grouping variable is a useful feature for researchers wishing to test the effectiveness of a prevention intervention program. By treating a dummy variable representing program versus control group membership as the exogenous grouping variable, the researcher can compare ρ, δ, and τ parameters across groups. If the prevention program is successful, the transition probability matrices will indicate that the probability of moving to a more advanced stage of drug use is lower for the program participants than for the control group. An advantage of taking a stage-sequential approach is that examining the transition probability matrix reveals how effective the program is for individuals entering with different levels and types of substance use experience. For example, Graham and colleagues (1991) found that a prevention program that was successful overall was not successful for individuals who had entered the prevention program having tried tobacco but not alcohol.

Although LTA is a promising technique that offers the researcher a unique look at the onset process, it has some serious shortcomings. Two shortcomings stem from sparseness, which can occur when there are many indicators and relatively few subjects and/or when the measurement parameters are extreme. One of these shortcomings is the problem of goodness-of-fit testing, discussed earlier in this chapter; the other is large standard errors for some of the parameters, particularly the transition probabilities. Despite these problems, Collins and Wugalter (1992) concluded based on an extensive simulation that the addition of indicators is a benefit to most latent transition models as long as the indicators belong in the model. The procedure also has some limitations. For example, LTA currently does not have a missing data procedure, so listwise deletion of subjects must be used. Also, the procedure currently cannot incorporate continuous exogenous predictors, such as grade point average. The authors are working on expanding the capability of LTA in both of these areas.

CONCLUSIONS

The analyses done in this study illustrate the benefits of the LTA approach for analysis of substance use data. LTA allows the researcher to test and compare a variety of models of the substance use onset process. In this example, the authors assessed whether several ethnic groups can be represented by the same general model. LTA can be used for many other types of research questions, including testing the effectiveness of drug abuse prevention interventions. Much information in an LTA is contained in the transition probability matrix, which shows the probabilities of transitions among stages, for instance, among stages in the drug use onset process. Furthermore, in LTA the transition probability matrix is latent, which means that error in the observed variables is taken into account when the matrix is computed. This produces a more meaningful picture of the patterns of substance use onset.

NOTE

1. This section may be skipped without loss of continuity.

REFERENCES

Bachman, J.G.; Wallace, J.M., Jr.; O'Malley, P.M.; Johnston, L.D.; Kurth, C.L.; and Neighbors, H.W. Racial/ethnic differences in smoking, drinking, and illicit drug use among American high school seniors, 1976-1989. *Am J Public Health* 81:372-377, 1991.

Brannock, J.C.; Schandler, S.L.; and Oncley, P.R., Jr. Cross-cultural and cognitive factors examined in groups of adolescent drinkers. *J Drug Issues* 20:427-442, 1990.

Cheung, Y.W. Ethnicity and alcohol/drug use revisited: A framework for future research. *Int J Addict* 25:581-605, 1991.

Collins, L.M.: Fidler, P.L.; Wugalter, S.E.; and Long, J.D. Goodness-of-fit testing for latent class models. *Multivariate Behav Res* 28:375-389, 1993.

Collins, L.M.; Graham, J.W.; Hansen, W.B.; and Johnson, C.A. Agreement between retrospective accounts of substance use and earlier reported substance use. *Appl Psychol Meas* 9:301-309, 1985.

Collins, L.M.; Graham, J.W.; Long, J.D.; and Hansen, W.B. Cross-validation of latent class models of early substance use onset. *Multivariate Behav Res,* in press-*b*.

Collins, L.M., and Wugalter, S.E. Latent class models for stage-sequential dynamic latent variables. *Multivariate Behav Res* 27:131-157, 1992.

Collins, L.M.; Wugalter, S.E.; and Rousculp, S.S. *LTA User's Guide*. Los Angeles: J.P. Guilford Laboratory of Quantitative Psychology, in press-*a*.

Cudeck, R.A., and Browne, M.W. Crossvalidation of covariance structures. *Multivariate Behav Res* 18:147-167, 1983.

Engs, R. Drinking patterns and drinking problems of college students. *J Stud Alcohol* 38:2144-2156, 1977.

Grady, K.; Gersick, K.E.; Snow, D.L.; and Kessen, M. The emergence of adolescent substance use. *J Drug Educ* 16:203-220, 1986.

Graham, J.W.; Collins, L.M.; Wugalter, S.W.; Chung, N.K.; and Hansen, W.B. Modeling transitions in latent stage-sequential processes: A substance use prevention example. *J Consult Clin Psychol* 59:48-57, 1991.

Graham, J.W.; Johnson, C.A.; Hansen, W.B.; Flay, B.R.; and Gee, M. *Prev Med* 19:305-313, 1990.

Graham, J.W.; Rohrbach, L.; Hansen, W.B.; Flay, B.R.; and Johnson, C.A. Convergent and discriminant validity for assessment of skill in resisting a role play alcohol offer. *Behav Assess* 11:353-379, 1989.

Hansen, W.B., and Graham, J.W. Preventing alcohol, marijuana, and cigarette use among adolescents: Peer pressure resistance training versus establishing conservative norms. *Prev Med* 20:414-430, 1991.

Hansen, W.B.; Graham, J.W.; Wolkenstein, B.H.; Lundy, B.Z.; Pearson, J.L.; Flay, B.R.; and Johnson, C.A. Differential impact of three alcohol prevention curricula on hypothesized mediating variables. *J Drug Educ* 18:143-153, 1988.

Holt, J.A., and Macready, G.B. A simulation study of the difference chi-square statistic for comparing latent class models under violation of regularity conditions. *Appl Psychol Meas* 13:221-232, 1989.

Humphrey, J.A., and Friedman, J. The onset of drinking and intoxication among university students. *J Stud Alcohol* 47:455-458, 1986.

Humphrey, J.A.; Stephens, V.; and Allen, D.F. Race, sex, marijuana use and alcohol intoxication in college students. *J Stud Alcohol* 44:733-738, 1983.

Jöreskog, K.G., and Sörbom, D. *LISREL 7 User's Reference Guide.* Chicago: Scientific Software, Inc., 1989.

Kandel, D.; Single, E.; and Kessler, R.C. The epidemiology of drug use among New York high school students: Distribution, trends, and change in rate of use. *Am J Public Health* 66:43-53, 1976.

McIntosh, W.; Fitch, F.; Staggs, F.; Nyberg, K.; and Wilson, B. Age and drug use by rural and urban adolescents. *J Drug Educ* 9:129-143, 1979.

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

Read, T.R.C., and Cressie, N.A.C. *Goodness-of-Fit Statistics for Discrete Multivariate Data.* New York: Springer-Verlag, 1988.

Walfish, H.; Wentz, D.; Banzing, P.; Brennan, F.; and Champ, S. Alcohol abuse in a college campus: A needs assessment. *Eval Program Plann* 4:163-168, 1981.

Wechsler, H., and McFadden, M.J.D. Drinking among college students in New England. *J Stud Alcohol* 40:969-996, 1979.

Welte, J.W., and Barnes, G.M. Alcohol use among adolescent minority groups. *J Stud Alcohol* 48:329-336, 1987.

Yamaguchi, K., and Kandel, D.B. Patterns of drug use from adolescence to young adulthood: II. Sequences of progression. *Am J Public Health* 74:668-672, 1984.

ACKNOWLEDGMENTS

This work was supported by National Institute on Drug Abuse grant DA 04111 and by the National Institute on Alcohol Abuse and Alcoholism.

AUTHORS

Linda M. Collins, Ph.D.
Professor
Department of Human Development and Family Studies

John W. Graham, Ph.D.
Professor
College of Health and Human Development

The Pennsylvania State University
110 South Henderson
University Park, PA 16802-6504

Susannah Scarborough Rousculp, M.A.
Research Assistant

Penny L. Fidler, M.A.
Research Assistant

Jia Pan, M.A.
Research Assistant

J.P. Guilford Laboratory of Quantitative Psychology
University of Southern California
Los Angeles, CA 90089-1061

William B. Hansen, Ph.D.
Associate Professor
Department of Public Health Sciences
Bowman Gray School of Medicine
Medical Boulevard
Winston-Salem, NC 27157-1063

Incorporating Trend Data To Aid in the Causal Interpretation of Individual-Level Correlations Among Variables: Examples Focusing on the Recent Decline in Marijuana Use

Jerald G. Bachman

ABSTRACT

Given the close correspondence of several trends beginning in 1979, it is tempting to conclude that increases in perceived risk and disapproval led to the decline in actual use of marijuana. In this chapter, two alternative interpretations are considered, reflecting different hypotheses about individual-level causal processes: (1) changes in use led to the changes in attitudes, or (2) changes in some other factor(s) (e.g., increased "conventionality") caused both changes in use and changes in attitudes.

This chapter documents a series of analyses designed to untangle such issues by incorporating trend data along with individual-level, cross-sectional relationships. One analysis strategy shows that controlling attitudes could "account for" the time trend in marijuana use, whereas the reverse is not true. The second analysis strategy examines how time trends in marijuana use are affected by multivariate controls for attitudes, as well as other individual characteristics, and shows that only the attitude measures can "explain" the time trend in marijuana use. Although these analyses are viewed as helping to explain the recent secular trend downward in marijuana use, as well as the still more recent decline in cocaine use, their most important contribution to prevention intervention research may be that they support a very basic generalization about individual-

level causal processes: *individual attitudes about specific drugs affect individual use of those drugs.*

INTRODUCTION

When two or more trends over time correspond closely with each other, it is tempting to conclude that there is an underlying causal connection. Conversely, a lack of correspondence suggests the absence of causal connection. In the field of drug research, a number of trend patterns have emerged that have potential implications for prevention efforts. Consider, for example, the following findings shown in figure 1, all based on the Monitoring the Future annual surveys of large representative samples of high school seniors:

1. Seniors' beliefs that marijuana is *harmful* began to increase in 1979 and continued to rise throughout the 1980s.

2. Seniors' *disapproval* of marijuana use showed nearly parallel increases beginning in 1980.

3. Seniors' *use* of marijuana *decreased* steadily beginning in 1980.

4. Seniors' perceptions that marijuana is readily *available* has shown little change from the mid-1970s onward.

First, and most simply, the above evidence strongly suggests that recent changes in marijuana use, as well as changes in perceived risk and disapproval, have had little to do with (perceived) availability of marijuana; this implies that the "supply side" strategy for prevention of marijuana use has not been very effective. That is not the only possible conclusion, of course, but it is surely the most parsimonious.

Second, given the close correspondence among the other trends, it is tempting to conclude that the increases in perceived risk and disapproval have contributed to the decline in actual use of marijuana. Here,

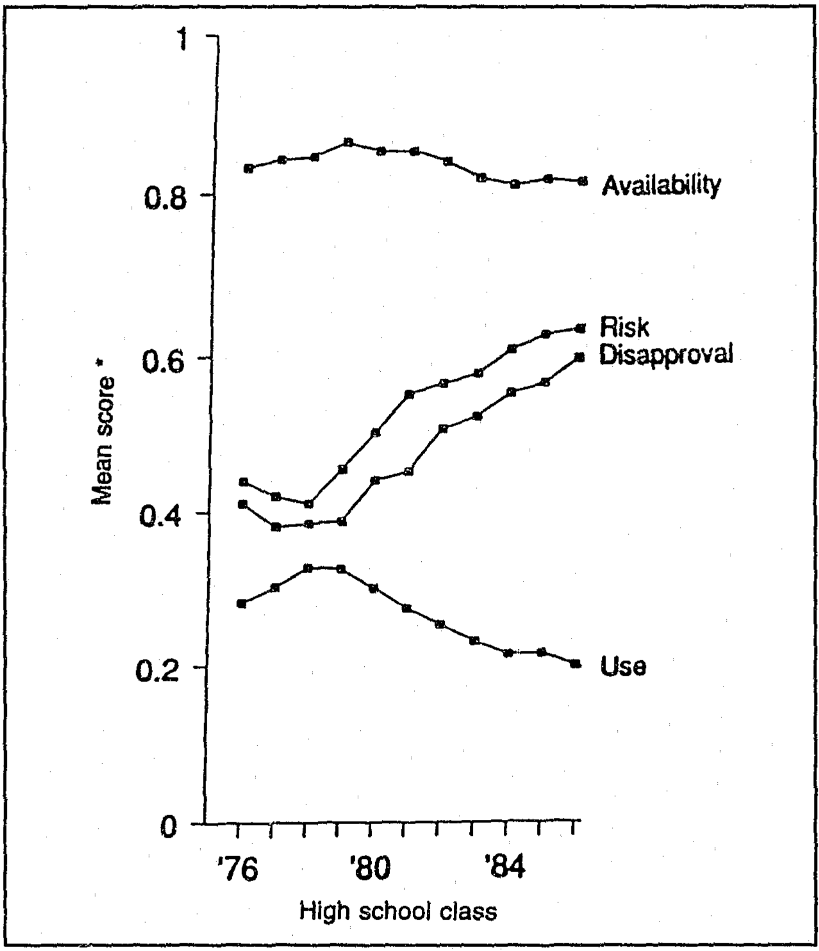


FIGURE 1. *Trends in annual marijuana, perceived availability, perceived risk, and disapproval: High school seniors, 1976-1986*

KEY: * All items were scaled with the minimum possible score set equal to 0 and the maximum possible score set equal to 1.

SOURCE: Adapted from Bachman, J.G.; Johnston, L.D.; and O'Malley, P.M. Explaining the recent decline in cocaine use among young adults: Further evidence that perceived risks and disapproval lead to reduced drug use. *J Health and Soc Behav* 31:173-184, 1990.

however, the argument becomes much more complicated. When faced with parallel (or opposite) trends, the question remains whether A (attitudes) causes B (behaviors), B causes A, or C (one or more other variables, perhaps unmeasured) causes both A and B. Moreover, there also is the problem of going from the aggregate level (reflected by the trend data) to the individual level (the level at which the causal hypotheses often are formulated). This chapter documents a series of analyses undertaken in the hope of untangling some of these issues, at least with respect to the recent changes involving marijuana attitudes and use. Several earlier papers have focused on the substantive findings with respect to drug use (Bachman et al. 1986, 1988, 1990). The present chapter incorporates key findings from these earlier papers but now focuses on the analysis strategy *per se*. Because the earlier papers were developed over a period of several years, the first major section of this chapter covers the interval from 1976 to 1985, and the second discusses the period from 1976 to 1986.

STATEMENT OF THE PROBLEM: DID CHANGES IN ATTITUDES DURING THE 1980s CAUSE THE DECLINE IN MARIJUANA USE AMONG YOUTH?

As suggested above, one straightforward interpretation of the marijuana findings is that the changes over time in attitudes (A) caused the changes in behaviors (B). Indeed, early reports of findings from the Monitoring the Future surveys of high school seniors stated a clear preference for that kind of "A causes B" interpretation (Johnston et al. 1981). Johnston (1985) later expanded the argument, noting additional trends (e.g., rising proportions of marijuana quitters who listed physical and/or psychological risks as their reasons for quitting), all consistent with the notions that individuals' use of marijuana is influenced by their attitudes about marijuana and that changing attitudes about marijuana (in response to various historical changes in such factors as information about the drug) may have led to a reduction in demand.

Jessor (1985) found Johnston's argument plausible but "not yet compelling," pointing out that aggregate trend data are not sufficient to establish causal order. Jessor spelled out two alternative hypotheses. The first of these is that B causes A: "It remains quite possible that regular use of marijuana declined and beliefs about its harmfulness subsequently increased rather than the other way around" (Jessor 1985). The second alternative is that C (conventionality) causes both A and B:

> It is possible to entertain an equally plausible alternative hypothesis to account for both the increased perception of harm from regular use and the actual decline in regular use, namely, that there has been an increase in the general conventionality of adolescents during this same historical period. Such an increase in conventionality would lead to less motivation to use marijuana or to seek its effects, and would also imply greater receptivity to messages from authorities about the harmfulness of drug use (p. 259).

Jessor's comments clearly articulate the problem faced by those who would draw conclusions from correspondences among trends: In the absence of additional data, it is virtually impossible to sort out cause and effect. Fortunately, the Monitoring the Future surveys do include an additional key ingredient: the fact that the several trends are based on the same annual samples of high school seniors rather than from completely independent sources permits analyses that incorporate individual-level, cross-sectional relationships. This ingredient is crucially important because the various hypotheses illustrated above all are based (implicitly, if not explicitly) on individual-level causal interpretations.

ANALYSIS ISSUES AND STRATEGIES: INCORPORATING INDIVIDUAL-LEVEL, CROSS-SECTIONAL RELATIONSHIPS AMONG VARIABLES ALONG WITH TREND DATA

It will be useful here to distinguish two analysis strategies, both of which require the examination of individual-level, cross-sectional relationships among the attitude and behavior measures. The first strategy focuses on whether the trend data can be explained by one of the two simplest interpretations: A causes B (operationalized as prediction 1 below) or B causes A (operationalized as prediction 2). The second strategy expands the scope of inquiry to consider whether some other factor(s), perhaps C, cause(s) both A and B.

First Analysis Strategy: Examining How Time Trends in Behaviors Are Affected by "Holding Constant" Attitudes, and Vice Versa

Samples and Measures. This section summarizes analyses reported in detail by Bachman and colleagues (1986). The data are derived from the Monitoring the Future surveys of high school seniors taken from 1976 to 1985. Each of these nationally representative annual surveys included 5 different questionnaire forms, with 3,000 or more cases per form each year. Although items on marijuana use appeared in all five forms, questions on perceived risk appeared only in form 5, and questions on disapproval appeared only in form 3. In more recent surveys, key questions on perceived risk and disapproval appear on several forms, thus permitting additional correlational analyses not possible with the earlier surveys.

Because the different forms involve random subsets of the total annual samples, there are very slight differences in marijuana use trends, depending upon whether the analysis is based on the form 5 subsample, which cross-tabulates marijuana use with perceived risk, as shown in figure 2, or on the form 3 subsample, which links marijuana use with disapproval, as shown in figure 3. All such differences are trivially small and do not affect the conclusions discussed here.

Recall that throughout the 1980s survey of each successive class of high school seniors showed higher rates of perceived risk and disapproval associated with marijuana and also lower levels of (self- reported) use. It seems most likely that the increased negative attitudes strongly contributed to the decline in marijuana-using behavior. Specifically, it is likely that an individual's attitudes about marijuana strongly influenced actual use of the drug and that changes over time in information about marijuana led to changed attitudes and, in turn, changed behavior. However, a plausible alternative interpretation is that seniors who did not use marijuana themselves were, as a consequence, more likely to feel and express negative views about marijuana. This distinction was operationalized in the form of two competing predictions.

Prediction 1: With attitudes held constant, marijuana use will show no change from one year to another. The underlying hypothesis is that individuals generally behave in accordance with their attitudes and that perceived risk and disapproval inhibit the use of marijuana. Therefore, as the proportions of young people holding these negative attitudes about marijuana increased each year, the numbers willing to use marijuana *consequently* declined. According to this argument, if this were the sole basis for the relationship between the two trends, then, after (statistically) "holding constant" the attitudes at any particular level, no decline in usage rates from one year to the next should have been observed within that attitude category.

Figure 2 is one example of the initial findings, based on analyses extending from 1976 to 1985. The figure shows that monthly marijuana use rates consistently were close to 70 percent among those who saw "no risk" in occasional marijuana use and less than 10 percent among those who saw "great risk." The trends within these two subgroups clearly were consistent with prediction 1; specifically, the fluctuations from year to year seemed largely random with no clear evidence of a trend upward *or* downward. Among those perceiving slight or moderate risk, the percentages of monthly marijuana users actually rose somewhat, prompting the comment that ". . . these data suggest that if it were not for the sharp increases in perceived risk since 1978, marijuana use for seniors

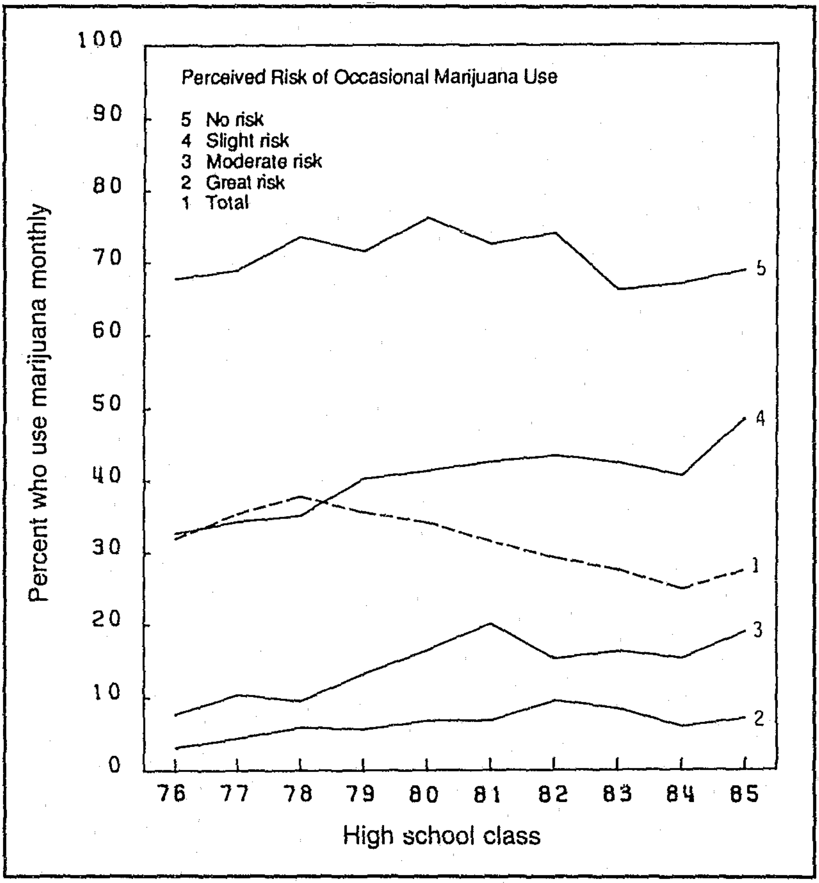


FIGURE 2. *Trends in monthly marijuana use by level of perceived risk of occasional marijuana use*

as a whole might have risen rather than declined" (Bachman et al. 1986, p. 12).

Figure 3 provides another example, again for the period from 1976 to 1985. It shows that monthly marijuana use rates consistently were about 60 percent among those who reported they "don't disapprove" of occasional marijuana use and 3 percent or lower among those who indicated they "strongly disapprove," both fully consistent with prediction 1. The usage rates for those in the intermediate category who said they

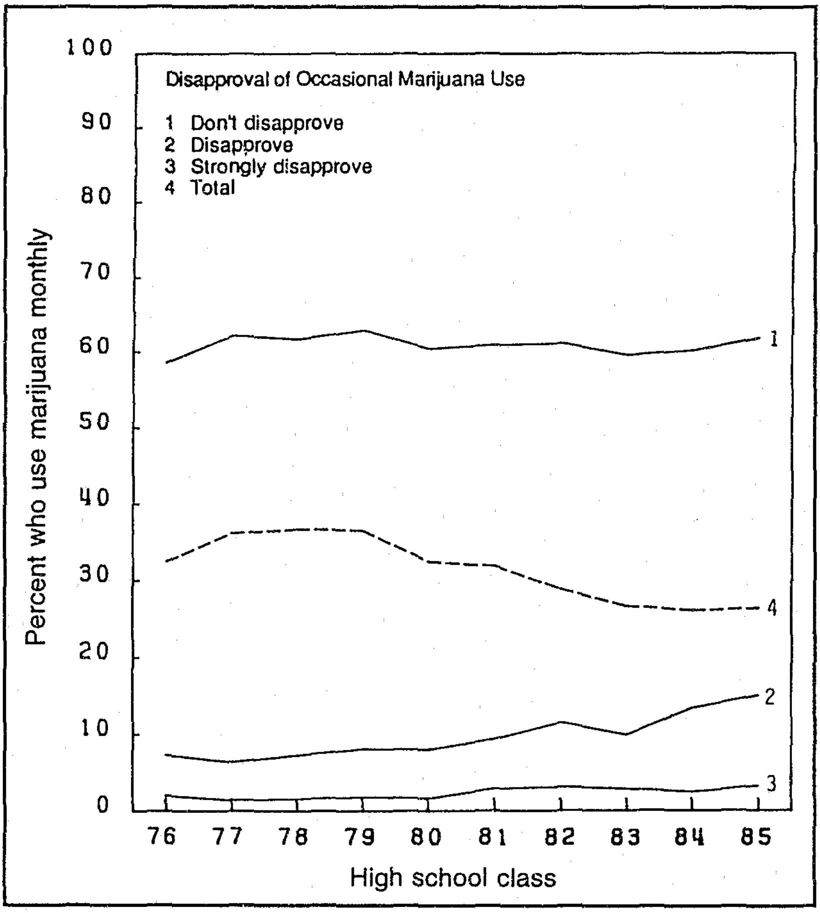


FIGURE 3. *Trends in monthly marijuana use by level of disapproval of occasional marijuana use*

"disapprove" rose from about 7 percent to 15 percent. These findings are similar to the findings for the intermediate levels of perceived risk.

The evidence thus far is largely supportive of prediction 1 but, before reaching any conclusions, the data from the reverse perspective, as stated in prediction 2, should be examined.

Prediction 2: With marijuana use held constant, attitudes about marijuana will not change from one year to another. The underlying hypothesis here is that individuals bring their attitudes into conformity

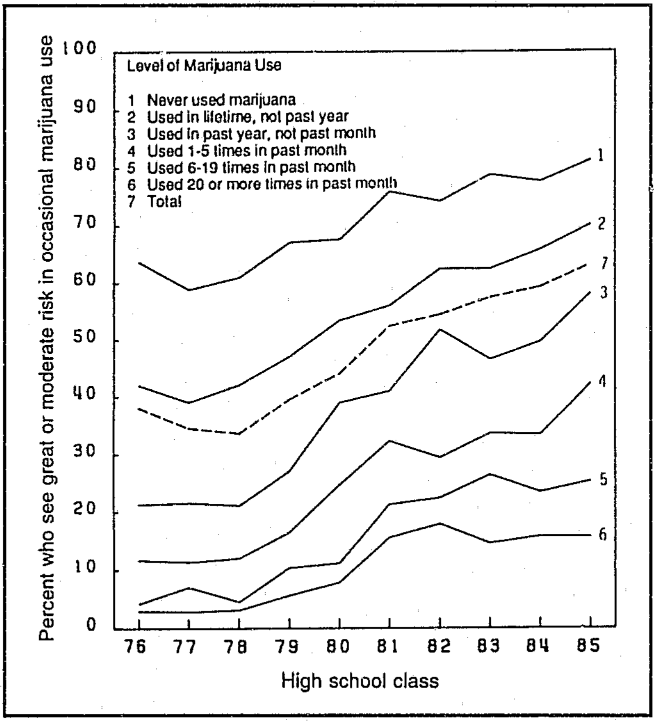


FIGURE 4. *Trends in perception of great or moderate risk in occasional marijuana use by level of marijuana use*

with their behaviors. According to this perspective, the only reason marijuana *attitudes* changed on average during the 1980s simply is that the proportions of individuals actually *using* marijuana grew progressively smaller. If that explanation is correct, then looking separately at subgroups who use marijuana frequently, those who seldom used it, and those who did not use it reveals little or no upward trend in disapproval or perceived risks.

In fact, as exemplified in figures 4 and 5, the data led to a very different conclusion:

> In sum, contrary to Prediction 2, we find that controlling for the behavior of marijuana does nothing to reduce

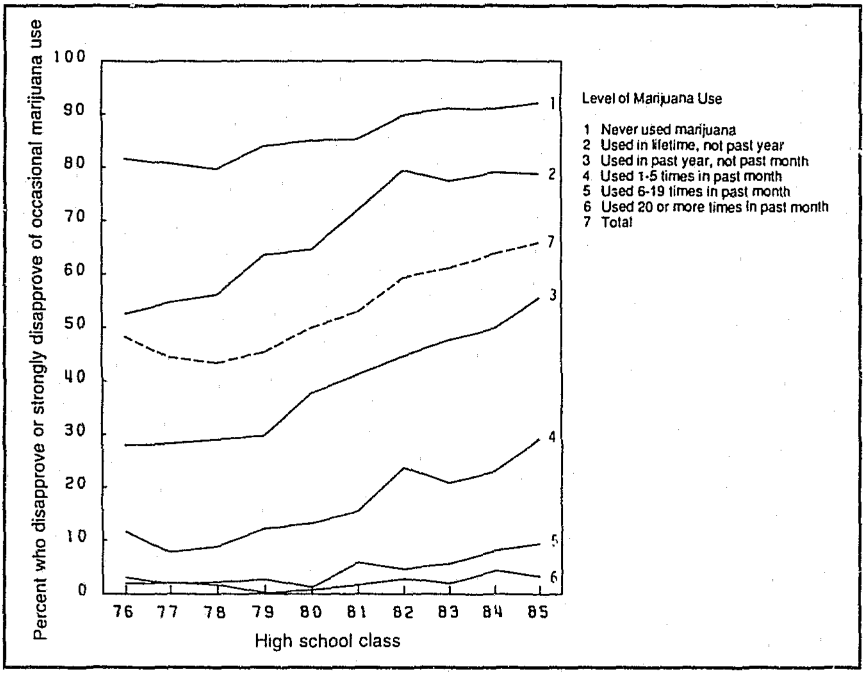


FIGURE 5. *Trends in disapproval or strong disapproval of occasional marijuana use by level of marijuana use*

> or "explain away" the upward trend from 1978 through 1985 in negative attitudes about marijuana. Subgroups consisting of frequent users, infrequent users, and non-users, all show substantial increases in the proportions wio disapprove of marijuana use and perceive that such use is risky (Bachman et al. 1986, p. 14).

Methodological Observations on the Technique of Examining One Trend While Holding Another Constant. The analyses just summarized really are quite elementary from a statistical standpoint—indeed, all tabulations are in the form of simple percentages. Instead of percentages, of course, mean rates of marijuana use in figures 2 and 3, mean perceived risk in figure 4, and mean disapproval in figure 5 could have been plotted with virtually identical results (e.g., Bachrnan et al. 1988; figure 1, this chapter). However, percentages are preferable, whenever possible, because of their ease of interpretation by broader audiences.

This lack of statistical complexity surely is one of the chief advantages of this technique. If controlling levels of A eliminates (or reverses) trends in B, and if the converse is *not* the case (i.e., controlling B does not eliminate trends in A), that would seem to be fairly straightforward and persuasive evidence. (However, a first journal submission based solely on this technique was *not* sufficiently convincing to the journal's reviewers. Perhaps simplicity also must be counted as a *dis*advantage!)

A major limitation of the technique just presented is that it is bivariate. It treats only two trends at a time and is limited to exploring whether A seems to cause B to a greater extent than whether B causes A (or vice versa, or neither). The problems of multiple causes (A1, A2, A3 . . .) or "third variable" causes (C), require more sophisticated methods, such as the next method.

Second Analysis Strategy: Examining How Time Trends in Behavior Are Affected by Multivariate Controls for Attitudes and Other Individual Characteristics.

As an extension of the first analysis strategy, the second analysis strategy incorporated several additional variables reflecting "lifestyle" factors that also could be considered as positive or negative indicators of conventionality. These analyses were carried out somewhat later than those described above. Thus, data from the 1986 cohort were added, extending the span to the period from 1976 to 1986.

Change and Stability in Lifestyle Factors Linked to Drug Use. Before incorporating lifestyle factors into multivariate analyses including time trends in drug use, it was important to address the stability of such factors and the consistency of their relationships with drug use (specifically, marijuana use). Overall, the level of consistency was rather high. The factors most important in predicting marijuana use during the late 1970s also were very important in the early 1980s. More specifically, marijuana use was more frequent among those who did poorly in school (those exhibiting low grades or frequent truancy), those frequently away from home in the evenings, those with high earnings and long hours committed to part-time work, those with low commitments to religion, and those describing their political views as very liberal or radical (Bachman et al. 1988). Background factors, such as race, parental education, number of parents in the home, urbanicity, and region, added relatively little in regression analyses when combined with the above factors; accordingly, these factors were not included in the multivariate analyses described below.

Given that these lifestyle factors remained important predictors of marijuana use throughout the 1976-1986 period, it was important to consider whether any of these factors showed sufficient change to account for the downward trend in marijuana use during the 1980s. Although political views moved in a conservative direction (which appears consistent with the decline in marijuana use), there also was a reduction in religious commitment (which would, if anything, lead one to expect an increase in marijuana use). Each of eight factors was examined separately following the first analysis strategy as illustrated in figures 2 and 3. Quite clearly, no single lifestyle or conventionality factor could "account for" the declining trend in marijuana use, whereas both perceived risk and disapproval *were* able to do so (Bachman et al. 1988). These preliminary analyses provided a great deal of useful detail; however, they also were somewhat cumbersome and lacked the ability to examine multiple factors simultaneously.

Pooling Data From Multiple Years. The strategy for providing multivariate controls was to employ straightforward multiple regression techniques, but applied to a somewhat unusual data set. Specifically, this employed analysis files that combined data from all 11 cohorts of seniors (in graduating classes of 1976-1986; total N per form was approximately 33,000). One advantage of such a pooling is that it simplifies reporting. Correlations between marijuana use and each of the other variables already studied already showed little or no change during the 1976-1986 period, so there was no need to continue reporting separate correlations for each of the 11 cohorts.

"Cohort Mean" Marijuana Use as a Measure of Secular Trend. By pooling the data across all 11 cohorts, it was possible to assign a new variable to each individual, consisting of the "cohort mean" for marijuana use. Specifically, each respondent was assigned the mean annual marijuana use score for his or her graduating cohort. This permitted calculation of correlations between individual marijuana use and the mean level of marijuana use among all seniors for that year. In other words, this made it possible to compute the extent to which the total variance in individual marijuana use throughout the period in question (1976-1986) was explainable simply in terms of which year the individual graduated—that is, the overall secular trend in use.[1] This new variable can be referred to as a measure of the secular trend in marijuana use. Confidence in treating this as a secular trend rather than as cohort differences resulted from a variety of other analyses that showed the secular trend interpretation is by far the most parsimonious in accounting for year-to-year changes in seniors' use of marijuana (O'Malley et al. 1984, 1988).

Here is how this assignment of scores actually worked. Annual marijuana use is reported on a 7-point scale, with the following values:

1 = 0 occasions
2 = 1-2 occasions
3 = 3-5 occasions
4 = 6-9 occasions
5 = 10-19 occasions
6 = 20-39 occasions
7 = 40 or more

The mean score on that scale for seniors in 1976 was about 2.7; accordingly, all respondents from 1976 were assigned 2.7 as their value on the new "marijuana secular trend" variable. For the class of 1977, the mean was about 2.8, so that value was assigned to all of them. For the classes of 1978 and 1979, the mean had reached about 3.0, so that value was added to all of their files. Thereafter, use declined gradually; by 1986 (the last year used in the analyses summarized here), the mean was down to about 2.2, so that was the value assigned to all members of the class of 1986. In other words, the marijuana secular trend variable rose from 2.7 to 3.0 and then declined to 2.2 during the interval studied. Extrapolating from the 7-point scale, these figures mean that marijuana use among seniors dropped by roughly half from 1979 (mean of about four uses per year) to 1986 (mean of about two uses per year).

The shift in cohort means across the 1976-1986 period is substantial; however, it does not begin to match the wide range of individual variation within each year—or across all years. Thus, the correlation between individual use and the marijuana secular trend variable necessarily was limited; the actual product-moment correlation was about 0.12, meaning that this substantial secular trend accounts for about 1.5 percent of the total variance in individual marijuana use during the period in question. As described in the first report:

> This finding serves as a useful reminder that although year-to-year variations in marijuana use over the past decade are important and interesting, such variations remain small in comparison to the wide range of variability among seniors within each year of the study (Bachman et al. 1988, p. 105)[2].

Nevertheless, that secular trend in marijuana-using behavior was viewed as quite important, given that annual use rates dropped by about half from 1979 to 1986. It is this very importance that prompted the exploration of whether the decline might be explainable in terms of such attitudinal factors as perceived risk and/or disapproval (i.e., A causes B) or in terms of changes in conventionality (C causes both A and B).

Regression Analyses Contrasting Different Sets of Predictors. Table 1[3] displays a portion of the regression analysis findings, those that include the disapproval measures. Three sets of variables, examined separately and then in combination, are treated as "predictors"[4] of indi-

TABLE 1. *Multiple regression analyses predicting annual marijuana use from (A) lifestyle variables, (B) disapproval of marijuana use, and (C) mean marijuana use per year*

Predictor	r	A	B	C	A+B	A+C	B+C	A+B+C
A) Lifestyle variables								
Grades	-.206	-.091			-.045	-.094		-.045
Truancy	.362	.239			.135	.233		.135
Hours worked per week	.117	.038			.011	.021		.012
Average weekly income	.131	.028			.038	.047		.037
Religious commitment [a]	-.269	-.171			-.051	-.176		-.051
Political beliefs [b]	.170	.090			.025	.089		.025
Evenings out per week	.315	.219			.111	.213		.111
Gender (M = 1, F = 2)	-.114	-.030			-.018	-.030		-.018
B) Disapproval of regular marijuana use	-.677		-.677		-.573		-.680	-.574
C) Mean marijuana use per year	.120			.120		.105	-0.15	-.003*
R		.497	.677	.120	.713	.507	.678	.713
R^2		.247	.459	.015	.508	.257	.459	.508

KEY: * $p > .05$

[a] Mean of two items: How often do you attend religious service? (1 = Never . . . 4 = About once a week or more); How important is religion in your life? (1 = Not important . . . 4 = Very important)

[b] Single item: How would you describe your political beliefs? (1 = Very conservative . . . 5 = Very liberal . . . 6 = Radical).

NOTE: Entries in the first column are product-moment correlations coefficients (r); entries in the bottom rows are multiple correlation coefficients (R and R^2) adjusted for degrees of freedom. All other table entries are standardized regression coefficients.

SOURCE: Adapted from Bachman, J.G.; Johnston, L.D.; O' Malley, P.M.; and Humphrey, R.H. Explaining the recent decline in marijuana use: Differentiating the effects of perceived risks, disapproval, and general lifestyle factors. *J Health Soc Behav* 29:92-112, 1988.

vidual seniors' self-reported amounts of marijuana use during the past year:

- Set A includes seven lifestyle dimensions plus gender;
- Set B is personal disapproval of regular marijuana use; and
- Set C is the marijuana secular trend measure (i.e., the nationwide mean of marijuana use—by seniors—during the year when the individual graduated).

The lifestyle variables in set A show a multiple correlation of .50 with annual marijuana use, explaining fully 25 percent of its variance. This contrasts with the much smaller correlation of .12 with set C, the secular trend measure, representing only 1.5 percent of the variance in marijuana use (as noted earlier). Clearly, if one wished to account for a senior's use of marijuana, then religiosity, truancy, and frequency of evenings out would provide much more explanatory power than knowing the year of graduation. However, a slightly better result is obtained by using both; indeed, set A+C accounts for fully 1.0 percent more variance than set A alone. Additionally, the regression coefficient for the secular trend measure is changed very little when the set A variables are added to the equation (the coefficient for C changes from .120 to .105). Thus, very little of the secular trend can be "explained away" by the lifestyle variables included as potential indicators of conventionality.

What about attitudes as an alternative approach to explaining the secular trend? Set B, disapproval of regular marijuana use, accounts for fully half of the variance in individual marijuana use. More importantly, the addition of the secular trend measure provides no increase at all in predictive power. The variance explained by set B+C is identical to that explained by set B alone, and the coefficient for C changes from .120 to -.015 when set B is included as a predictor. Thus, this part of the analysis leads to the same conclusion as the earlier, simpler approach: If there is a control for the attitude measure, the secular trend "effect" essentially disappears. (Indeed, the small but significant *negative* coefficient for C when set B is included among the predictors is consistent with the slight *upward* trends in lines 2 and 3 in figure 3, opposite to the *downward* trend for the total sample shown in line 4.) The same general finding can be seen occurs when the set A variables are included; comparing set A+B with set A+B+C shows again that, once the attitude measure is included in the equation, the secular trend measure does not explain any additional variance—and the coefficient for set C goes to -.003.

Although the data are not reproduced here, the findings were comparable when the attitude measure was perceived risk of regular marijuana use (Bachman et al. 1988). The conclusion was drawn from regression analyses that:

> . . . the secular trend in marijuana use cannot be "explained" in terms of the lifestyle measures included in Set A, but the trend *can* be "explained" either by the measure of perceived risk or by the measure of disapproval (p. 105).

A Replication and Extension: Explaining the Recent Decline in Cocaine Use. In May 1986, basketball star Len Bias died as a result of cocaine use; a few weeks later, football star Don Rogers also died because of cocaine. The following spring, the 1987 Monitoring the Future survey of high school seniors showed marked increases in perceived risk and disapproval associated with cocaine use, along with a substantial decline in self-reported use but no decrease in perceived availability of the drug. When it became clear that each of these trends continued into 1988, it seemed worthwhile to conduct multivariate analysis of the cocaine trends and to see if the pattern of results in some respects replicated those obtained in the earlier analyses of marijuana trends. Although the relationships were weaker with respect to cocaine (as would be expected, given the much lower rates of usage for this drug), the analyses again showed that, whereas the lifestyle factors could not "explain" the recent decline in cocaine use, the attitudes—either perceived risk or disapproval—could (Bachman et al. 1990). Those analyses were based on data through 1988, but more recent tabulations have shown that the trends in cocaine attitudes and use continued for several additional years (Johnston et al. 1992).

However, even in the absence of the complex multivariate analyses, the researchers noted that by now the simple trend comparison had become more compelling:

> We would find it hard to argue plausibly that such different secular trends in the use of these two drugs [marijuana and cocaine] could have been caused by some *general* trend among young people toward becoming more "conservative" or less "trouble-prone" in recent years. . . . Changes in *drug-specific* factors, on the other hand, correspond clearly to the declines in both marijuana use and cocaine use (Bachman et al. 1990, p. 181).

Methodological Observations on the Multivariate Analysis Strategy. The chief advantage of this second of the two lines of analysis simply is that it is multivariate; it permits examining a wide range of predictors simultaneously and exploring the extent to which explained variance is shared (overlapping) or unique while, at the same time, including the secular trend measure as one predictor.

An additional advantage of this multivariate approach is that it places the secular trend "effects" alongside "effects" (i.e., correlations) involving individual differences in lifestyles and attitudes; in the process, it illustrates dramatically that the action is much greater at the individual level. Why, then, bother with the secular trends? One reason is they still are quite large—as noted earlier, marijuana use was cut about in half from 1979 to 1986. The more compelling reason, from the present perspective, is that *the analyses of secular trends may provide some additional leverage in the attempts to sort out causal interpretations at the individual level.* Specifically, the present findings (i.e., that the secular trends in attitudes can "account for" the secular trends in use, whereas the reverse is *not* true) are strongly consistent with the interpretation that *individual attitudes about specific drugs influence individual drug use behavior.*

GENERAL OBS⁻RVATIONS ON THE ANALYSIS STRATEGY OF COMBINING INDIVIDUAL-LEVEL, CROSS-SECTIONAL DATA WITH TREND DATA

The analyses summarized in this chapter were prompted by a desire to learn more about the causal connections between drug-related attitudes and the actual use of drugs. On the one hand, the analyses made use of individual-level correlational data in explaining trends in both attitudes and behaviors with respect to marijuana and later cocaine. On the other hand, and perhaps more importantly, the analyses used the trend data to provide some extra leverage in understanding individual-level causal dynamics.

The researchers interpreted the findings as supporting the initial hypothesis that individuals' attitudes about a drug—specifically, perceived risk and disapproval—are among the primary factors contributing to their use or nonuse of that drug. The multivariate analyses also clearly indicate, however, that these are not the only contributors; other lifestyle factors also appear to have an impact, consistent with findings in much earlier analyses (Bachman et al. 1981).

On reviewing this work, which evolved over several years, it seems that two basic conditions must be met in order for this strategy to lead to clear conclusions. First, it is necessary that the attitude and behavior measures show some correlation at the individual level. That certainly is the case with respect to marijuana; annual use correlated -.57 with perceived risk and -.68 with disapproval. Such correlations clearly indicate the *possibility* that one factor has a direct (and/or indirect) causal impact on the other. The second condition is that the secular trend is stronger for one factor than for the other; specifically, the "between-years variance" (i.e., the variance "explained" by knowing the year of measurement) must be greater for one of the two factors. That also is the case. Figure 1 shows that the rises in perceived risk and disapproval are more pronounced than the corresponding declines in marijuana use. It should be added that, for this purpose, it would be technically correct to scale figure 1 to equalize standard deviations rather than ranges. The latter was chosen for the published report because of its greater intuitive value and because it turned out that the two scalings were mostly similar. The one difference is in line with the researchers' preferred interpretation: The rise in perceived risk is even more pronounced when scaled to equalize standard deviations.

With these two conditions in place, the most *parsimonious* interpretation is that (1) some factors that changed from one year to another led to substantial shifts in attitudes about marijuana, and (2) because such attitudes do affect behavior, there was a smaller shift in marijuana use (it is smaller because the attitude-behavior correlation is less than perfect). It is important to stress that, so long as the correlation between cause and effect is distinctly lower than 1.0, the change on the *outcome* dimensions should be somewhat smaller than the change on the *causal* dimensions. It should be noted that, if the correlation were very close to 1.0, the techniques described in this chapter would not give any leverage in disentangling causes from consequences.

One other methodological observation is that it does not seem strictly necessary to have all data from the same sets of respondents in order to meet the two conditions described above. If one knows the extent to which each of two dimensions have shown aggregate year-to-year changes and can express those changes as proportions of the *individual-level* variance (whether that variance estimate is obtained from the same data or elsewhere), and if one also has a trustworthy estimate of the *individual-level* correlation between the two dimensions (again, whether obtained from the same or different data sets), then one can carry out the kinds of calculations done here—at least with respect to estimating whether A causes B more than B causes A.

Possible Adjustments for Measurement Error

The analyses described above did not take account of issues of measurement error, at least not explicitly. The first strategy, examining time trends in behaviors while holding constant attitudes (and vice versa), is not easily adaptable to adjustments for measurement error. But the second strategy, involving multivariate controls and using cohort means to indicate secular trends, is readily amenable to such adjustments. The simplest approach would be to disattenuate the correlation matrix (i.e., adjust correlations upward to compensate for estimated measurement error) before conducting regression analyses. A more comprehensive approach might be to use structural equation models.

Such adjustments for measurement error were not included in earlier reports because doing so would not have changed the findings substantially and, thus, the additional complexity was not warranted. The researchers reached that conclusion considering carefully the likely effect of adjustments for measurement error. It may be useful to review those considerations here, especially since, in other applications of this approach, it may be appropriate to include such adjustments:

1. *Individual self-reports of drug use.* Earlier analyses documented what appears to be a widespread systematic bias toward under reporting total occasions of drug use over a 12-month interval, compared with a 30-day interval. That bias was attributed largely to failure of recall rather than deliberate distortion (Bachman and O'Malley 1981). Such a bias, however, does not necessarily distort correlations or lower reliability estimates. In fact, fairly high levels of reliability have been estimated consistently in the drug use measures (O'Malley et al. 1983). For example, in other analyses that *did* use disattenuated correlations, the estimated reliability of the annual marijuana use measure was .90 (Bachman et al. 1984).

2. *Cohort means as measures of secular trends in drug use.* Each graduating cohort of seniors is represented by a sample of approximately 16,000 cases. With these numbers of cases, the *sampling* error is vanishingly small. Accordingly, it seems that no adjustment for measurement error would be needed for this variable.

3. *Individual measures of drug-related attitudes.* Assessments of reliability and stability have focused primarily on measures of drug use rather than measures of drug-related attitudes. Nevertheless, it is

likely that reliabilities are lower for the attitude measures since these items involve 3-point or 4-point response scales with large majorities of respondents sometimes clustered in a single category.

4. *Other measures used in the regression analyses.* Some of the measures listed as lifestyle variables in table 1 can be assumed to have fairly low to very low error (e.g., grades, hours worked, income, and gender), while others (e.g., truancy, religious commitment, political beliefs, and evenings out) may have moderate error.

Likely Effects of Adjustments for Measurement Errors. Suppose the above sorts of measurement errors were taken into account and appropriate adjustments were made so as to dissattenuate the correlation matrix underlying the calculations shown in table 1. The result would have been slightly larger estimates of the relationships in table 1, but there would have been no important change in overall patterns or conclusions. That judgment is based on the specific considerations discussed below.

First of all, the reliability estimate of .90 for the dependent variable measure, individual-level annual marijuana use, would lead to adjusting virtually all coefficients upward to a slight extent. Specifically, for a simple correlation with a second measure judged to be error free, such as mean marijuana use per year, the estimate would be the original correlation multiplied by the reciprocal of the square root of the estimated reliability (in this case, 1÷.949 = 1.054). The result would be that the correlation of .120 in table 1 would be adjusted upward to .126.

Second, the large negative correlation between disapproval and marijuana use would be enhanced by the above adjustment and also by a (probably larger) adjustment reflecting the measurement error in the disapproval measure. After such adjustments, it would remain true that, when predictor set B+C was used, the coefficient for C would be close to 0, and the joint prediction would not be any improvement over the use of the attitudes (set B) alone. In other words, the changing attitudes would continue to "account for" the secular trend in marijuana use.

Finally, the further adjustment in the lifestyle variables (set A) would heighten their overall contribution, but that would not change the story appreciably with respect to the *marginal* contribution of the secular trend measure (set C). Overall R-squared values would rise, of course, but the purpose in these analyses was not to seek a precise estimate of those values; rather, the purpose was to see whether some factors might "account for" or "explain" the secular trend in marijuana use.

IMPLICATIONS OF EPIDEMIOLOGICAL TREND STUDIES FOR PREVENTION INTERVENTION

The first journal article reporting the analyses summarized here suggested that one of the implications for those concerned with prevention is that " . . . realistic information about risks and consequences of drug use, communicated by a credible source, can be persuasive and can play an important role in reducing demand, which ultimately must be the most effective means of reducing drug use" (Bachman et al. 1988, p. 108-109).

It must be emphasized that the conclusion quoted above reflects an inference about *individual-level causal processes*—an inference developed by exploiting trend data coupled with some individual-level data. It also should be stressed that it is the individual-level interpretation that is likely to have the most important implications for prevention intervention.

An important question remains about what caused the overall trends during the 1980s in attitudes about marijuana. The interpretation was offered earlier that *some factors that changed from one year to another led to substantial shifts in attitudes about marijuana.* What were those factors? The important factors very likely included increasingly persuasive research findings on physical and psychological consequences, more extensive and effective coverage in the media, and firsthand observation of some schoolmates (virtually no school was immune) who did indeed fit the reports about marijuana-using "burnouts." Were some of those factors "prevention intervention?" That is, perhaps, a matter of definition.

In any case, the point here is that these trend analyses do not tell us which among a myriad of societal forces were most dominant in producing the year-to-year changes in perceived risks and disapproval associated with marijuana use. In a previous National Institute on Drug Abuse research monograph on prevention intervention research, Johnston (1991) made the same point quite clearly:

> Epidemiological studies . . . provide outcome data on the aggregate impact of all the forces in society that influence drug use—whether they are *labeled* as prevention programs, whether they are *intended* to prevent or promote drug use, and whether they are *organized programs* (p. 74).

The trend studies and analyses can be very useful, in other words, but they remain only one part of the prevention intervention research puzzle.

NOTES

1. An alternative strategy, if the researchers had been willing to assume that any secular trend was strictly linear, would have been to assign to individuals numerical values of 1 through 11 (or 1976 through 1986) corresponding to their graduating class and then consider the extent to which *those* values correlated with individual use (again using the pooled individual data from all 11 classes). Such an approach, however, would not have captured the curvilinear trend in marijuana use during the period in question. On the other hand, it would have avoided any tendency to capitalize on chance fluctuations from year to year—not much of a problem when the annual means are based on thousands of cases, but potentially a problem with smaller samples.

2. Later analyses (Bachman et al. 1990) extending from 1976 to 1988 showed the trend continuing, thus explaining more variance (the product-moment correlation rose to about .16, accounting for about 2.5 percent of the total variance).

3. For further details and comparable data on perceived risk, see Bachman et al. (1988), from which table 1 was adapted.

4. The terms "predictor" and "variance explained" are used here because they are the familiar ones used in describing regression analyses. In fact, the author does not assume single directions of causation for some of the lifestyle dimensions. Moreover, the secular trend "correlation" is recognized as merely a different way of expressing the proportion of overall individual differences in marijuana use related to overall year-to-year changes during the decade studied.

REFERENCES

Bachman, J.G.; Johnston, L.D.; and O'Malley, P.M. Explaining the recent decline in cocaine use among young adults: Further evidence that perceived risks and disapproval lead to reduced drug use. *J Health Soc Behav* 31:173-184, 1990.

Bachman, J.G.; Johnston, L.D.; O'Malley, P.M.; and Humphrey, R.H. "Changes in Marijuana Use Linked to Changes in Perceived Risks and Disapproval." Occasional Paper No. 19. Ann Arbor, MI: Institute for Social Research, 1986.

Bachman, J.G.; Johnston, L.D.; O'Malley, P.M.; and Humphrey, R.H. Explaining the recent decline in marijuana use: Differentiating the effects of perceived risks, disapproval, and general lifestyle factors. *J Health Soc Behav* 29:92-112, 1988.

Bachman, J.G., and O'Malley. P.M. When four months equal a year: Inconsistencies in students' reports of drug use. *Public Opin Q* 45:536-548, 1981.

Bachman, J.G.; O'Malley, P.M.; and Johnston, L.D. Drug use among young adults: The impacts of role status and social environments. *J Pers Soc Psychol* 47:629-645, 1984.

Jessor, R. Bridging etiology and prevention in drug abuse research. In: Jones, C.L., and Battjes, R.J., eds. *Etiology of Drug Abuse: Implications for Prevention.* National Institute on Drug Abuse Research Monograph 56. DHHS Pub. No. (ADM)85-1335. Washington, DC: Supt. of Docs., U.S. Govt. Print. Off., 1985.

Johnston, L.D. The etiology and prevention of substance use: What can we learn from recent historical changes? In: Jones, C.L., and Battjes, R.J., eds. *Etiology of Drug Abuse: Implications for Prevention.* National Institute on Drug Abuse Research Monograph 56. DHHS Pub. No. (ADM)85-1335. Washington, DC: Supt. of Docs., U.S. Govt. Print. Off., 1985. pp. 155-177.

Johnston, L.D. Contributions of drug epidemiology to the field of drug abuse prevention. In: Leukefeld, C.G., and Brakoslu, W. J., eds. *Drug Abuse Prevention Research: Methodological Issues.* NIDA Research Monograph 107. DHHS Pub. No. (ADM)91-1761. Washington, DC: Supt. of Docs., Govt. Print Off., 1991. pp. 57-80.

Johnston, L.D.; Bachman, J.G.; and O'Malley, P.M. *Highlights From Student Drug Use in America, 1975-1980.* DHHS Pub. No. (ADM)81-1066. Washington, DC: Supt. of Docs., U.S. Govt. Print. Off., 1981.

Johnston, L.D.; O'Malley, P.M.; and Bachman, J.G. *Smoking, Drinking, and Illicit Drug Use Among American Secondary School Students, College Students, and Young Adults, 1975-1991.* Vol. 1, *Secondary Students.* DHHS Pub. No. (ADM)92-1920. Washington, DC: Supt. of Docs., U.S. Govt. Print. Off., 1992.

O'Malley, P.M.; Bachman, J.G.; and Johnston, L.D. Reliability and consistency of self-reports of drug use. *Int J Addict* 18:805-824, 1983.

O'Malley, P.M.; Bachman, J.G.; and Johnston, L.D. Period, age, and cohort effects on substance use among American youth. *Am J Public Health* 74:682-688, 1984.

O'Malley, P.M.; Bachman, J.G.; and Johnston, L.D. Period, age, and cohort effects on substance use among young Americans: A decade of change, 1976-1986. *Am J Public Health* 78:1315-1321, 1988.

ACKNOWLEDGMENTS

This work was supported by grant R01 DA 01411 from the National Institute on Drug Abuse.

This chapter builds upon earlier work done in collaboration with Lloyd Johnston, Patrick O'Malley, and Ronald Humphrey. Patrick O'Malley and John Schulenberg reviewed and commented upon earlier drafts of this chapter. All of these contributions are gratefully acknowledged.

AUTHOR

Jerald G. Bachman, Ph.D.
Research Scientist and Program Director
Institute for Social Research
University of Michigan
426 Thompson Street
Ann Arbor, MI 48106-1248

Multilevel Models for Hierarchically Nested Data: Potential Applications in Substance Abuse Prevention Research

Ita G.G. Kreft

ABSTRACT

This chapter reports on an application of a multilevel analysis. A multilevel analysis is a data analysis that uses variables that are measured at different levels of the hierarchy. A hierarchy can have many levels, such as student level, class level, school level, and State or country level, where students are nested within classes, classes are nested within schools or school districts, and school districts can be nested within towns, States, or countries. As soon as one pays attention, hierarchies are present in all data. In large-scale prevention research, researchers usually have information about two or more levels involved, for instance, variables describing individuals (such as achievement, drug use, gender, and measures of socioeconomic status or home environment); variables describing schools (such as school environment, urban versus rural, and type of treatment administered); and perhaps variables describing districts, States, or countries. It is well known that the analysis of variables (i.e., measures at different levels of the hierarchy) on any of these levels separately can be misleading, as will be shown in this chapter. It is more satisfactory to construct a model and technique that simultaneously take information on all levels into account.

This chapter introduces such a multilevel model for hierarchically nested data by evaluating the effect of a drug prevention program, Normative Education (NORM), wherein data are collected on students nested within

schools. The model is a linear regression model. The difference between this model and the traditional linear regression model is that it takes the intraclass correlation into account and treats variables measured at different levels of the hierarchy in a more appropriate way.

INTRODUCTION

Methodological problems are present whenever real-life experimentation is the object of study. This chapter deals with one of them: how to analyze data that are collected over students in existing schools, where the treatment consists of drug prevention programs. From the hierarchical structure of the data, where students are nested within classrooms, classrooms are nested within schools, and schools are nested within programs, it follows that measurements can be obtained from all levels of this hierarchy. If measurements are at different levels, a question that remains is, "What should be taken as the unit of analysis, the student or the school?" A way to make this choice is by asking another question: "What is the unit of interest?"

It seems that the effect of the drug prevention programs on individual students is the main object of interest, and a logical choice would be an analysis of covariance (ANCOVA). However, ANCOVA has its problems in this situation. The first problem is that observations in groups are correlated. This statistical problem cannot be solved in the traditional ANCOVA framework. ANCOVA analysis also lacks the ability to answer important questions, such as: "What are the effects of drug prevention programs on special groups of students, for instance, high-risk students or boys versus girls?" This chapter will argue that analyzing this type of data with any of the traditional linear techniques, either ANCOVA or regression (i.e., regression executed at the school, the class, or the student level) is not satisfactory. To analyze those data, a model that can handle the clustered and hierarchical structure of the data in an appropriate way is needed.

To analyze data that have a hierarchical structure and contain measurements from different levels of the hierarchy (i.e., multilevel measurements), techniques that are based on assumptions that are in agreement with the data structure are needed. The next paragraphs summarize the main problems that have to be dealt with when analyzing multilevel data. The concepts defined here are intraclass dependency, hierarchical nesting, random effects, cross-level interaction, and different sources of variation in unbalanced data.

Intraclass Dependency

Observations that are close in time, space, or both are assumed to be more similar than observations far apart in time, space, or both. Intraclass correlation is defined as the degree to which individuals share common experiences due to closeness in space, time, or both. In traditional linear models, the effect of omitted variables is summarized in the error term. The assumption is that individual errors are unrelated to each other because the omitted variables have a random, instead of a specific, effect.

However, since observations in the same group or *context* share some omitted variables (i.e., the ones related to the shared context), a covariance of individual error terms can be observed in such situations. For example, in evaluation of drug prevention programs, existing school classes are used. Students in the same class have a lot in common: They share the same school environment and have the same teachers. The omitted variables in analysis models represent nonrandom influences of the same school climate and the same peer pressure for students in the same school. The degree of covariance in the error terms of individuals sharing the same school or class can be expressed in a correlation coefficient, that is, defined as the ratio of between-school variation to total variation in the dependent variable (Cochran 1977). This correlation is known as the *intraclass correlation.*

Intraclass correlation is associated in the literature with an increase in Type I errors (Barcikowski 1981; Cochran 1977; Crits-Christoph and Mintz 1991; Murray and Hannan 1990). If intraclass correlation occurs, as it will when clustered data are sampled, the assumption of independent observations in the traditional linear model is violated. For instance, the 30 students in the same class are not 30 independent observations. The degree of intraclass correlation determines how many independent observations there really are. Since tests of significance lean heavily on the number of independent observations involved, the existence of intraclass correlation makes the test of significance too liberal when using traditional linear models. Based on research by Barcikowski (1981), it can be shown that even a small intraclass correlation (like $r = 0.01$) can inflate the alpha level from the assumed level of 0.05 to 0.17 under specific circumstances (see table 1 later in this chapter).

Random Effects versus Fixed Effects

In fixed effects analysis of variance (ANOVA), the factor (or treatment) is said to be fixed if all possible treatments in which the researcher is interested are present. In research that uses real-life groups, this assumption can hardly ever be made. For instance, in drug prevention research, the treatment is administered to groups, such as school classes, that are a random sample from all possible school classes. Students are nested within these groups. An effect of a certain treatment in real-life experimentation has to be redefined as random instead of fixed because the groups are not formed by randomization of, for instance, students over treatment groups. Treatment effects have to be viewed as random effects because the effect may differ randomly from group to group or from school to school. In drug use prevention programs, schools are sampled from a large population of schools. Even when this sampling is convenient rather than strictly random, random effects are assumed.

The following distinction between random and fixed models can be made: fixed effects models focus on differences between means, while random effects models focus on variances. It has to be kept in mind that the way the data are obtained affects the inferences that can be made. Random effects can be used as the basis for making inferences about populations from which the samples are drawn. In other words, "in endeavoring to decide whether a set of effects is fixed or random, the context of the data, the manner in which they were gathered and the environment from which they came are the determining factors" (Searle et al. 1992, p. 16).

Hierarchical Nesting

Students are nested within classes, classes are nested within schools and neighborhoods, and schools are nested within States or countries. Once researchers know that hierarchies exist, they can see them everywhere. When samples of real-life groups are used in an experiment, such as school classes, classes instead of individuals are said to be assigned to treatments. The concept of "group," in the context of multilevel analysis, should not be confused with the concept of treatment or treatment group.

Groups

The hierarchy of nesting in drug prevention research usually is students nested in schools and schools assigned to programs or treatments. In the multilevel literature, the lower level (the students) is referred to as "micro-level," while the highest level (schools and programs) is referred to as "macro-level." Measurements obtained at all levels of a hierarchy can be analyzed simultaneously in multilevel modeling. For instance, student measurements, such as gender, race, poverty, level of rebelliousness, and level of risk, are analyzed in relation to school-level measurements, such as rebelliousness level, drug use level, and environmental risk factors of the school.

An example of hierarchically nested data and problems related to analyzing such nested data is in Bachman (this volume). In Bachman's chapter, the question is raised whether the negative relationship between perceived risk and drug use resulting from an aggregated analysis over the years can be related to an individual relationship, where perceived risk causes a lowering of marijuana use. The analysis results show that an aggregate trend over years is different from trends observed when separate groups of individuals are studied (Bachman, this volume, figures 1-5). This well-described phenomenon is called the ecological fallacy or

Robinson effect, named after the author of the first article (Robinson 1950) that showed that aggregated models can measure different things from individual models and, hence, can lead to different conclusions (also see Kreft and De Leeuw 1988). Bachman's chapter is a nice illustration of the complications a researcher faces when studying trends over years with the intention of finding causal relationships between individual attitudes and individual marijuana use.

Cross-Level Interactions

Cross-level interactions are interactions between context and student. This type of interaction was first mentioned in the educational research literature (Cronbach and Webb 1975). The assumption made in Cronbach and Webb (1975) is that some teachers interact better with certain types of students than with others. If certain teachers are, for instance, more effective with bright students than with others, it means that the relationship between an individual student's aptitude and achievement is strengthened. Such a teacher is said to have a meritocratic teaching style. If, on the contrary, a teacher is more effective with slow learners, the relationship between aptitude and achievement may be reduced. The teacher is said to have an egalitarian teaching style. The first type of teacher widens the gap between high and low performers, while the second type of teacher narrows this gap. In the educational literature this is called an aptitude/treatment effect. In theory, the same can happen in drug prevention programs. While some programs widen the gap between high-risk and low-risk students, others may narrow the gap between these two groups of students.

Collins and colleagues (this volume) and Uebersax (this volume) show potential applications of models that test for cross-level interactions. In both chapters, techniques are presented for classification of students according to certain patterns of drug use. After labeling students according to type of drug use, a subsequent multilevel analysis can test if drug prevention programs are more effective for certain types of students than for others. Defining again students as "micro" and prevention programs as "macro," such a cross-level interaction is a micro-macro interaction.

The program strengthens or reduces the relationship between type of student and drug use. The stronger the micro-macro interaction, the stronger the effect of the program for that specific type of student in either direction.

Unbalanced Data and Sources of Variation

When dealing with multilevel data, researchers deal with a nested design, where existing schools are nested within drug prevention programs and where schools may have unequal numbers of students. As illustrated in table 1, a nested design gives rise to two sources of variation: a variation between individuals within groups and a variation between groups within each treatment. To analyze such multilevel data, the analysis model needs to provide for a separation of the total variation in the dependent variable into different sources. The variance components are associated with the larger unit (such as the school or the B's in table 1) and with the smaller units (such as the students or the O's in table 1) within each treatment (the A's in table 1). As a result of the way the data are structured, there is more than one source of variability at the group level: a variation between the groups within the same treatment and a variation between treatments.

THE ANALYSIS OF DRUG PREVENTION DATA: THREE TRADITIONAL STRATEGIES

There are three traditional strategies for analyzing multilevel data: (1) ANCOVA, (2) a means-to-means regression approach, and (3) a "slopes as outcomes" approach. ANCOVA is straightforward and probably what most people do. The means-to-means regression approach involves group means as the unit of analysis in regression to avoid the problem of intraclass correlation in clustered sample designs. Burstein (1980) suggested a new and better approach to the analysis of multilevel data, the slopes as outcome approach. This last approach is considered to be a multilevel technique because it is based explicitly on the fact that

TABLE 1. *Hierarchy of 2 treatments (A), 4 groups (B), and 22 observations (O)*

A_1		A_2	
B_1	B_2	B_1	B_2
O_1	O_7	O_{11}	O_{18}
O_2	O_8	O_{12}	O_{19}
O_3	O_9	O_{13}	O_{20}
O_4	O_{10}	O_{14}	O_{21}
O_5		O_{14}	O_{22}
O_6		O_{16}	
		O_{17}	

observations are collected at different levels and are clustered. The slopes as outcomes approach ("separate models" in table 2), however, has, its own disadvantages and problems. Table 2 illustrates the differences among the two traditional linear models and the two multilevel linear models in relation to the modeling of context effects.

The new approach mentioned in table 2, multilevel random models, conceptually is close to the slopes as outcomes approach and will be introduced later in this chapter. In this section, each of the fixed effects approaches will be reviewed, and the strengths and weaknesses of each will be discussed.

Basic Equations and Assumptions of Fixed Linear Models

The basic equation for all models mentioned in table 2 is in the equation at the top of the next page, where the convention of underlining random variables is used.

$$\underline{Y}_{ij} = a_j + b_j \underline{X}_{ij} + \underline{e}_{ij}$$
$$a_j = \text{Intercept(s)} \qquad (1)$$
$$b_j = \text{Slope(s)}$$

TABLE 2. *Assumptions of two traditional linear models compared to two multilevel models*

	Intercepts	Slopes
Traditional linear regression	equal	equal
ANCOVA	unequal	equal
Multilevel fixed models	unequal	unequal
Multilevel random models	unequal	choice (either equal or unequal)

Subscripts refer to i for individual and j for group; e_{ij} is the usual individual error term, with a mean of 0 and a variance of σ^2. The first three models in table 2 are fixed effects linear models. Within the fixed models, the choices are that intercepts are equal:

$$a_0 = a_1 = \ldots = a_m \qquad (2)$$

or unequal:

$$a_1 \neq a_2 \neq \ldots \neq a_m \qquad (3)$$

Equation (2) applies to the total regression model, not to the ANOVA model or the separate models for separate contexts. ANOVA models assume equation (3), while the separate models for separate contexts approach also assumes this by definition.

The fixed effects models in table 2 also may differ in their assumptions in relation to the slope coefficients. Slopes can be assumed to be equal over contexts as in equation (4), which is an assumption on which the ANOVA model is based:

$$\mathbf{b}_1 = \mathbf{b}_2 = \ldots = \mathbf{b}_m \tag{4}$$

or unequal:

$$\mathbf{b}_1 \neq \mathbf{b}_2 \neq \ldots \neq \mathbf{b}_m \tag{5}$$

Equation (4) states that slopes are equal for all contexts, which is an assumption of most fixed effects models, with the exception of the separate models analysis. The last assumes by definition that all contexts differ in their parameters since, for each context, a separate model is fitted. In multilevel models (the fixed, separate models for separate groups, as well as the random model), assumption (5) is made when it is expected that the relationship between the dependent and independent variable is different over contexts. In separate models, slopes and intercepts are different by definition, whereas in multilevel random effects models intercepts are assumed to be different and differing slopes are given as an option.

ILLUSTRATION OF THE FIXED EFFECTS LINEAR MODELS

The next paragraph illustrates several analyses by using a data set (Hansen and Graham 1991) of 12 schools, with 120 classes, 2,069 students, and 2 treatment situations. Measurements at the micro-level are student prealcohol use and student postalcohol use. Macro-level characteristics are the drug prevention program (NORM) versus something else and mean alcohol level of the school. NORM is short for "Normative Education." Mean preprogram alcohol level of schools is used here as proxy for laws and norms favorable towards alcohol and drug use by peers, siblings, and parents.

Analysis of Covariance (ANCOVA)

An assumption of ANCOVA is that each covariate (here, prealcohol use) has the same relationship with the dependent variable (here, postalcohol use) within each school. The regression coefficient of pretest on posttest in ANCOVA is the pooled within-regression coefficient. The ANCOVA model is applied to the data using schools as the macro-level, students as the micro-level, prealcohol use as the covariate, and postalcohol as the dependent variable.

The equation for the ANCOVA is:

$$\underline{Y}_{ij} = \alpha_j + b_W X_{ij} + \underline{e}_{ij} \tag{6}$$

where the Greek letter for the intercept (α_j) refers to the different estimates for each school. The best estimate for the slope (b) is the pooled within-group estimate, b_W. The estimate of α_j, which is different for every school, is $\overline{Y}_{.j} - b_w \overline{X}_{.j}$. The dot replaces the subscript i in $\overline{Y}_{.j}$ and $\overline{X}_{.j}$ since the pretest score (X) and the posttest score (Y) are summarized over individuals (i) in each school (j) separately. $\overline{Y}_{.j}$ and $\overline{X}_{.j}$ represent the school means for these variables.

The solution for schools obtained by ANCOVA is:

$$\hat{Y}_{ij} = \alpha_j + \mathbf{0.51} X_{ij}$$

The *F*-test for differences between α's is:

$$\boldsymbol{F(11, 2056) = 3.15, p = 0.000}$$

and the value for the pooled within-regression coefficient is 0.51 and equal to a coefficient that would be obtained by a regression equation over all students irrespective of their schools (see summary table 5 later in this chapter). The *F*-test indicates that some or all schools differ significantly in their mean alcohol level when corrected for pretreatment alcohol use. Remember, however, that the Type I error rate is inflated significantly when intraclass correlation is present. A check for intraclass correlation shows that r = 0.01 for these data, which brings the Type I error rate to at least 0.17, according to the table produced by Barcikowski (1981) and summarized in table 3 on the following page. In table 3, it can be seen that, based on the large number of observations per school (most schools have more than 100 observations), even small intraclass correlations may lead to high Type I errors.

More About the Effect of Intraclass Correlation

In studies using existing groups, as in the present example, not students but schools are randomly assigned to treatments. Students within the same school share many experiences (among them, the group dynamics during the treatment) that make them in certain ways more similar to each other than students in different schools. This violates the assumption of independency of observations in linear models and results in an intraclass correlation between the error terms in the linear model. Intraclass correlation reduces the number of independent observations compared to the observed number of observations, enhancing Type I error probability, depending on the number of observations in a school and the magnitude of the intraclass correlation. As shown in table 3, a small intraclass correlation of r = 0.01 in schools with 100 students inflates the Type I error rate from the assumed 0.05 to an observed 0.17 for an ANOVA, while a large intraclass correlation of 0.20 enhances the alpha level to 0.28 (instead of the assumed 0.05) in small schools with only 10 observations per school. Table 3 shows how much the observed alpha levels (in the body of the table) differ from the nominal alpha level (alpha = 0.05) for different values of intraclass correlation and different numbers of observations within groups.

A next logical step would be to see why schools differ significantly in their intercepts. ANOVA models do not show if some of the differences between schools can be attributed to macro-level characteristics. The observed differences between schools in this model may be the result of the drug prevention program NORM in combination with other factors

TABLE 3. *The inflation of the alpha level in the presence of intraclass correlation (Barcikowski 1981, p. 270)*

N per group	Intraclass correlation		
	0.01	0.05	0.20
10	0.06	0.11	0.28
25	0.08	0.19	0.46
50	0.11	0.30	0.59
100	0.17	0.43	0.70

NOTE: The values in the body of the table are the observed alpha levels.

(for instance, the preprogram mean alcohol level of a school), but testing such effects is beyond the limits of this fixed effects model. ANOVA shows limitations and, although the data have been analyzed at the correct level for making inferences about individual students, the fact that the Type I error is inflated poses real problems for inferences. To avoid the danger of Type I errors, using group means as the unit of analysis instead of individual observations is considered by some (Barcikowski 1981; Murray and Hannan 1990) to be more appropriate.

The Aggregate Model

Using the school as the unit of analysis will solve the problem of intraclass correlation. The aggregate model is a between-school model (see equation [7]), where mean prealcohol level is used to predict mean postalcohol level. In the present data, this analysis is based on N = 12. Note on the following page that the subscript "dot j" means that the observations are summarized over individuals (i), and only the subscript (j) for group remains.

$$\underline{\overline{Y}}_{.j} = a + b_B \overline{X}_{.j} + \underline{e}_{.j}$$

If the $\underline{e}_{.j}$ are assumed to be independent, with variance $n_j^{-1}\sigma^2$ (i.e., variance weighted by the number of observations within groups), it follows that the best estimate of b is $b_{Between}$. In the aggregated model, the distinction between individual and contextual effects disappears. The results are (with standard scores between parentheses):

$$\hat{\overline{Y}}_{.j} = \underset{(z=-0.13)}{-0.04} + \underset{(z=1.66)}{0.59}\,\overline{X}_{.j}$$

Both coefficients are nonsignificant. The correlation between mean pretest and mean posttest is r = 0.47.

The aggregate model has several disadvantages, which range from loss of power to loss of interpretation. In this example, with 12 schools, the number of observations dropped from 2,069 to 12. As a result of this loss of power, the conclusion is that prealcohol use is unrelated to postalcohol use. Moreover, inferences to the student level based on these results could be incorrect and, later in the chapter, they will be shown to be incorrect (see table 4).

This illustrates the most serious problem with aggregated models: They answer the wrong questions. Questions about how schools behave are not equivalent to questions about how students behave. Drug prevention research targets students and effects of drug prevention programs on individual students, as well as on certain types of students. Hawkins and colleagues (1992) reached a similar conclusion in their review of the literature—that the overall effect of the program is important as an effect on individual students. Reasons are given why cross-level interaction may exist between the student (micro-level) and programs (macro-level). Questions can be raised, such as, "What is the effect of drug prevention programs on high-risk students versus low-risk students?" where "at risk" may be defined at all levels of the hierarchy, including the individual student level, the social environment level, and the school level. In Hawkins and colleagues (1992), several descriptors of risk factors defined at different levels of the hierarchy are given based on the literature, such as individual student risk factors, environmentally based risk factors, and family-based risk factors. Individual student risk factors are physiological (such as hyperactivity and attention deficit), academic (failure and lack of commitment), or family oriented (high levels of conflict in the family and laws and norms favorable towards alcohol and drug use by peers, siblings, and parents). Environmental risk factors are described as extreme economic deprivation and poverty, neighborhood disorganization, and availability of drugs.

The conclusion is that the aggregate model does solve the Type I error rate problem, but at the cost of a serious loss of power. More importantly, it is off the mark conceptually since it cannot address the question of whether special cross-level interaction effects exist. The following model, first proposed by Burstein and colleagues (1978), offers opportunities for testing the cross-level interaction effects.

Separate Models for Separate Schools

A more suitable analysis than ANCOVA for the hierarchically nested data that still is within the traditional fixed effects linear model framework is fitting a separate model within each school. In the next table, table 4, the result of fitting 12 models within 12 schools is shown where student prealcohol use (X_{ij}) predicts student postalcohol use (Y_{ij}) in each school. The estimates for intercepts and slopes show to be different across schools. The intercepts are nonsignificant, except two, which is almost contradictory to the earlier reported results of the ANCOVA analysis, where the *F*-test results show highly significant differences among (at least some) intercepts or alphas [F(11, 2056) = 3.15, p = 0.000]. The widely differing slopes (from 0.36 to 0.71) contraindicate the fitting of a pooled within slope as is done in ANCOVA.

TABLE 4. *Regressions of prealcohol use on postalcohol use over 12 schools*

Schools	Intercept	(SE)	Slope	(SE)	R	N
1	-0.08	(0.05)	0.41*	(0.07)	0.41	190
2	0.00	(0.06)	0.49*	(0.11)	0.33	161
3	0.12	(0.06)	0.71*	(0.09)	0.53	205
4	0.01	(0.06)	0.57*	(0.09)	0.46	164
5	0.20*	(0.07)	0.74*	(0.08)	0.58	156
6	-0.02	(0.05)	0.56*	(0.06)	0.53	195
7	0.01	(0.05)	0.55*	(0.06)	0.55	192
8	0.07	(0.06)	0.41*	(0.05)	0.47	213
9	-0.16*	(0.04)	0.36*	(0.04)	0.52	185
10	-0.11	(0.06)	0.39*	(0.06)	0.50	118
11	-0.12	(0.06)	0.55*	(0.05)	0.69	118
12	-0.03	(0.06)	0.52*	(0.06)	0.54	172
Total	-0.003	(0.02)	0.51*	(0.02)	0.51	2,069

KEY: * is significant at $p < 0.01$

Comparing results over the separate models in table 4 and ANCOVA, it may look as if the *F*-test in the ANCOVA model is based on two schools (#5 and #9), the only schools that differ significantly from 0. Doing that may be misleading since the two analysis models are incomparable. By allowing the slopes to differ in the last model, the intercepts are different from the ones obtained by ANCOVA and so will be the significance of the ANCOVA *F*-test. A more substantive discussion of this difference between fixed effects ANCOVA models and multilevel models can be found in Aitkin and Longford (1986).

Table 4 shows that the strengths of the correlation between pretest and posttest vary in a similar fashion as the slope coefficients do, meaning that the differences in schools mainly are in their relationships between prealcohol and postalcohol use and not in their intercepts (see column "R" in table 4). Comparing the separate models in each school with the overall or individual model over all observations (see "total" row of table 4) shows again that schools differ from the total model, mainly in their slope and correlation coefficients.

Separate models for separate schools, wherein a student-level micromodel is fitted within each school, reflect the conceptual idea behind multilevel modeling. However, the separate models approach is not very parsimonious. In this simple example, with only one predictor, three parameters per school are estimated: the parameter for the intercept, for the slope, and for the individual-level error variance, which brings the total number of parameters for the 12 schools to 36 in this first step. Sometimes, even more parameters are required, as will be shown next. The same random effects model needs only six parameters to obtain comparable results, which will be explained later in this chapter.

Separate analyses are the first step in testing for separate school effects. The next step is checking for cross-level interactions between school and student. Researchers know from the literature that the environment of a student can have a moderating effect on individual drug use. Brook and colleagues (1990), for instance, found that the effect of drug-using peers was moderated by a strong attachment or bond between parents and adolescents. Rutter (1985) found that resilient children display a repertoire of social skills and belief in their own self-efficacy. Hawkins and colleagues (1992) express the need for research that studies interactions between student characteristics and the characteristics of the environment (i.e., drug prevention programs): "It is not known how children who come from poor managed families, who have failed in school, who are aggressive, or who lost commitment to school respond to 'Just Say No' or other anti-drug messages in the media or in their personal social environments. . . . Additional research is needed on the effectiveness of school policies in preventing or reducing the use of drugs other than tobacco and on the effects of such policies on those at highest risk for drug abuse" (Hawkins et al. 1992, p. 89).

Assuming that characteristics of schools can function as moderators, a model is fitted next with an interaction effect between the mean alcohol level of schools and the student alcohol use. The question of interest here is: "Can a characteristic of a school inhibit or enhance the substance abuse of high-risk students, where 'high-risk' is defined as students with a high level of alcohol use?" From the first step, the separate analyses, it is known that the schools show different magnitudes of the relationship between prealcohol and postalcohol use of students. The next step is to test if mean school alcohol level is related to these observed school differences. In the second step, the values of the slopes constitute the dependent variable, which is predicted by the school mean. This approach appropriately is called the slopes as outcomes approach in the first article that used this procedure (Burstein et al. 1978). Although Burstein and colleagues (1978) used the values of the slopes as the dependent variable in a second step, the values for the intercepts can be used also as the dependent variable in another macro-level regression, where the same macro-variables (e.g., the school alcohol mean) may be used again as the predictors.

Equation (5) shows the macro-level regression analysis with schools as the unit of analysis, slopes as the values for the dependent variable, and school alcohol mean, $\overline{X}_j$, as the predictor. The research question is: "Does a cross-level effect exist between the mean alcohol level of a school and the student-level relationship between prealcohol and postalcohol use?" The research hypothesis is nondirectional, meaning that it does not predict in what direction this effect will be. Either the outcome is that a school with a high mean level of alcohol consumption has a negative (lowering) effect on the positive relation between prealcohol and postalcohol use of the student or the opposite, a positive (strengthening) effect:

$$\underline{b}_j = \alpha+\beta\overline{X}_{.j}+\underline{e}_{.j} \qquad (8)$$

The macro-level equation in model (8) again is an aggregated model, with the difference, compared to the aggregated model reported earlier, that the dependent variable is produced by a statistical model (i.e., a linear that the dependent variable is produced by a statistical model (i.e., a linear regression) instead of being a simple average. Model (8) relates the slope parameters obtained in the micro-models in step 1 (refer to table 4), to a macro-level regressor, which is the mean prealcohol use over schools. The slope as outcomes model again is a fixed effects model. The solution for equation (5), where $\overline{X}_j$ is the mean alcohol level for each school is (with standard scores in parentheses):

$$\mathbf{slopes = 0.52 + 0.16\,\overline{X}_j}$$

$$\text{(z = 13.0) (z = 0.40)}$$

The intercept is the main effect of the slopes, representing the effect of prealcohol use on postalcohol use. The results are, for that reason, close to results obtained earlier for the slope coefficient in ANCOVA and in the individual regression model, where the magnitude of the slope coefficient is 0.51, and significant (with large z-values). The coefficient for $\overline{X}_j$ (representing the cross-level interaction between alcohol mean of the school and student-level alcohol use) is 0.16 but nonsignificant. The correlation between slope coefficients (b's) and school mean is r = 0.12. The coefficient for the slope is 0.16 and not significant (z = 0.40), showing no significant cross-level interaction between student alcohol use and school mean. Since the intercepts in table 4 are almost all close to 0, no "intercepts as outcomes" model is fitted. This lack of variation prohibits any successful further analyses.

The two-step separate models for separate schools approach has as advantages over the traditional models that it treats the observations at the appropriate level and allows for different effects within different schools. On the other hand, the model is ill defined in a statistical sense. For instance, schools are analyzed separately, without any reference to each other; the values of the estimations are taken at face value, without any reference to their reliability; and the error structures at both levels (micro-error, $\underline{e}_{ij}$, and macro-error, $\underline{e}_{.j}$) are not defined. Another disadvantage is the lack of parsimoniousness. Many parameters have to be estimated even in simple analyses such as this one.

Summary

Different models present different answers; the individual student regression shows only a significant slope coefficient, the aggregate model shows no significant results, and the ANCOVA model shows significant differences over intercepts. The results of the individual regression model and the ANCOVA model are questionable because of the existing intraclass correlation. The ANCOVA model erroneously assumes equal slopes for all schools. The results of the aggregate model shows the relation between variables related to schools, which is not the same as a model for students. The two-step separate models for separate schools is statistically ill defined and not very parsimonious. Clearly all three models have their own specific problems for answering questions related to drug prevention programs and their influence on individual students. Would it not be nice to have an approach that allowed inferences at all levels of a problem, produced the correct Type I error rate, did not result in a loss of power, and was parsimonious? Models can answer multilevel questions such as: "If this program is effective, is it equally effective for high-risk students as for low-risk students?"

Such a model will be introduced next. The convention adopted earlier to underline random variables and random coefficients $\underline{a}_j$ and $\underline{b}_j$ of the linear model will be used again. A more extensive discussion of random effects multilevel models follows later in this chapter.

THE MULTILEVEL RANDOM MODEL

A General Introduction

The multilevel random model presented here is a straightforward generalization of the separate models for separate schools approach (again see table 4 and discussion). The basic ideas of the random effects model is the same general approach as in the slopes as outcomes. Again, there is a student-level micro-model, defined separately for each macro-unit (the school). This is a linear model, with an individual-level predictor (prealcohol use) and individual-level dependent variable (postalcohol use). Mason and colleagues (1983), the first to publish an article using this type of multilevel modeling, made the following remarks: "Although its origins are uncertain, the notion of a regression in which the dependent variable consists of regression coefficients from other regressions has long been attractive to social scientists and statisticians" (p. 73).

In the separate equations approach, researchers must decide what exactly they are modeling in the second set of equations. Either the regression coefficients in the within-group models are fixed parameters, or they are random coefficients. If they are fixed, they can be estimated by ordinary within-group regression analysis. However, the distribution of the fixed within-group regression coefficients already is determined in the first step and cannot borrow any strength from other information available in the data set. The slopes as outcomes approach automatically leads to the following question: "Should the regression coefficients in the micro-models be modeled as random variables or as fixed constants?" One answer is that it depends on the contexts and the purpose of the analysis. If contexts (schools) are a random sample of the population of contexts (schools) and the purpose of the analysis is to generalize to this population (to all possible schools in a certain area), researchers may consider a model with random, instead of fixed, coefficients.

Before going into more detail, examine the results of applying a random effects multilevel model to the data. In table 5, the results of the random multilevel model are compared with the results obtained by fixed effects models (i.e., the total individual student model and the aggregate school model). Table 5 shows different symbols for the parameters in fixed effects versus random effects models. The fixed coefficients (a and b) are the symbols used in the fixed models, while in the random models the symbols for the fixed coefficients are the gammas (γ_{00} and γ_{10}) with their respective variances, the omegas (ω_{00} and ω_{11}). Since random effects models have random effects (indicated by the underlining of $\underline{a}_j$ and $\underline{b}_j$ in the equation in table 5), parameters reflecting that randomness are the omegas. The gammas in the random model conceptually are compa-

TABLE 5. *Comparison of parameter estimates between two fixed and one random model*

Individual Model		Aggregate Model		HL Model	
$\underline{Y}_{ij} = a+bX_{ij}+\underline{e}_{ij}$		$\underline{\bar{Y}}_{.j} = a+b\,\bar{X}_{.j}+\underline{\bar{e}}_{.j}$		$\underline{Y}_{ij} = \underline{a}_j+\underline{b}_jX_{ij}+\underline{e}_{ij}$	
$\hat{Y}_{ij} = -0.003+0.51X_{ij}$		$\hat{Y}_j = 0.04+0.59\,X_{.j}$		$\hat{Y}_{ij} = -0.006+0.51X_{ij}$	
parameter estimate	z-test	parameter estimate	z-test	parameter estimate	z-test
a -0.003	-0.17	a -0.004	-0.13	γ_{00} -0.005	
b 0.514*	26.94	b 0.589	1.66	γ_{10} 0.518*	15.83
		variance of the intercept ω_{00}		0.005*	4.32
		variance of the slope ω_{11}		0.008*	4.46

KEY: * is significant at $p < 0.01$

rable to the point estimators or fixed effects in fixed models (like the a and b are), while the variances are the measures of spread or the fluctuation of the schools around the mean estimates for intercept and slopes (the gammas in the random model).

Comparing the values of the fixed effects (a) and (b) of the individual and aggregate models, differences are found again in significance level of the parameters. Comparing the estimated parameters of the fixed and random models shows that the gamma values (and respective standard errors) of the random model are close to the a and b values (and respective standard errors) in the total individual student model. The differences over models clearly are somewhere else. The extra parameters estimated in the random model for variance of schools around the intercept (ω_{00}) and variance of schools around the slope (ω_{11}) for the random model make this model different from the total fixed effects model and, at the same time, more promising.

The concept of separate models for separate schools is introduced here by translating this concept into the freedom the model allows for schools to fluctuate around a mean value for slope and a mean value for intercept. These extra parameters introduce the opportunity to go beyond the student level and find macro-level variables that can explain this variation between schools, much in the same way as was demonstrated in the separate models for separate schools analysis, where slopes as outcomes was used and predicted by school mean alcohol level. Both variances in the random model, the variance for the intercept (ω_{00}) and the variance for the slope (ω_{11}), are significant, with z-values of over 4.00. This result gives reasons to proceed with a model that includes macro-level variables. Macro-variables can model the variances in intercept as well as in slopes. The macro-level variables NORM and mean alcohol use are used in the next paragraphs in an attempt to explain the observed variation among schools (in intercepts as well as in slopes), much in the same way as was done before, when the researchers tried (unsuccessfully) to explain the variation in slopes (see table 4) by the mean alcohol level of schools in the slope as outcomes approach.

The Random Effects Model for Hierarchically Nested Data

The random effects models presented in more detail in this section are comparable to the random effects models found in textbooks such as Searle and colleagues (1992) and Winer (1971). The difference is that these are too general for present purposes. To distinguish the random effects models used in the literature from the one introduced here, the name "random coefficient (RC) model" will be used for the random effects multilevel model for the rest of this chapter. The main difference between the random component models, as discussed in Searle and colleagues (1992) and Winer (1971), is that in RC models more than just variances are estimated since means are estimated along with their variances. The last fact is the reason this model also is known as a "mixed" model (for more details, see Searle et al. 1992).

In equation (9), the RC model is introduced, which has the form of the usual fixed effects linear model (compare for that purpose the model in equation [1]), with a single individual predictor, X_{ij}, but random coefficients, $\underline{a}_j$ and $\underline{b}_j$. The convention of underlining random coefficients is used again here.

$$\underline{Y}_{ij} = \underline{a}_j+\underline{b}_jX_{ij}+\underline{e}_{ij} \tag{9}$$

where $\underline{a}_j$ and $\underline{b}_j$ are random coefficients with a fixed and a random part as in equations (10) and (11):

$$\underline{a}_j = \gamma_0+\underline{\delta}_{0j} \tag{10}$$

and

$$\underline{b}_j = \gamma_1+\underline{\delta}_{1j} \tag{11}$$

As was shown in table 5, the random intercept ($\underline{a}_j$) is estimated as two parameters rather than as one: The first parameter is in the mean intercept over schools, and the second parameter is the variation among schools around that mean. The same is true for the slope ($\underline{b}_j$). Compared to the coefficients estimated in a fixed individual model, two estimates are obtained for each intercept (a) and each slope (b), instead of only one parameter. The fixed parts or means are the gammas (representing the mean values summarized over all schools), and the random parts are the deltas (representing the fluctuation or error of each school around the mean values). Equation (12) shows equation (9), with $\underline{a}_j$ and $\underline{b}_j$ replaced by their two parts, the random part or macro-error ($\underline{\delta}$), and the fixed part or γ:

$$\underline{Y}_{ij} = \gamma_0+\gamma_1X_{ij}+(\underline{e}_{ij}+\underline{\delta}_{0j}+\underline{\delta}_{1j}X_{ij}) \tag{12}$$

Except for the complicated error structure (between parentheses) with macro- and micro-disturbances, the model in equation (12) looks like the usual regression model. The macro-error associated with the intercept while $\underline{\delta}_{1j}$ is the macro-error associated with the slope as well as with the values for X. The macro-level errors are unrelated to the micro-level errors, $\underline{e}_{ij}$. The variance of the intercept is ω_{00}, and the variance of the slope is ω_{11}. The variance of the e_{ij}'s is σ^2.

The random parts of each coefficient are of special interest since this variation can be used to model macro-level effects. The fixed parts of slope and intercept (γ_0 and γ_1) are of interest for the estimation of the micro effects (such as the effects of students prealcohol on postalcohol use), while the random parts are of interest for the estimation of macro effects (as NORM and alcohol mean level of a school). The macro-level errors ($\underline{\delta}_{0j}$ and $\underline{\delta}_{1j}$) are the deviations of schools from intercept and slope estimates, respectively. In analogy to ANOVA, the variance of the student error terms (σ^2) is the within variance, while the between variance is split up in more than one source of macro-level disturbances (ω_{00}, ω_{01}, and ω_{11}). The difference between RC models and ANOVA is that more than one between variance is allowed to exist in an RC model. The present example shows a between-school variance of the intercept, a between-school variance of the slope, and a covariance between slope variance and intercept variance. The difference in definition of the a's and b's in fixed models compared to random models is that the first are conceived as representing the same treatments, whereas the a's and b's in the random case are conceived as random samples of a population of parameters, distributed as (γ_{00}, ω_{00}) for the intercept and as (γ_{10}, ω_{11}) for the slope. Each variance (ω_{00} or ω_{11}) is a variance in its own right and is a component of the variance of Y.

A technical summary of the RC model[1]

In the usual notation for RC models, the random coefficients $\underline{a}_j$ and $\underline{b}_j$ are defined as:

$$\underline{a}_j = \gamma_0 + \underline{\delta}_{0j} \qquad (10)$$

and

$$\underline{b}_j = \gamma_1 + \underline{\delta}_{1j} \qquad (11)$$

where $\underline{\delta}_{0j}$ has variance ω_{00}, $\underline{\delta}_{1j}$ has variance ω_{11}, and $\underline{\delta}_{0j}$ and $\underline{\delta}_{1j}$ have covariance ω_{01}. This extension of the variance component models shows that the total variance, usually divided in a single within and a single between part, now again is divided in a single within part. However, at the between-group level, there are three components, one for each coefficient (the variances of the macro-level errors $\underline{\delta}_{0j}$ and $\underline{\delta}_{1j}$ and a covariance between the two variances of intercept and slope).

In more general notation,

$$\underline{Y}_{ij} = \Sigma X_{ij}\gamma + \Sigma X_{ij}\,\underline{\delta}_j \qquad (13)$$

where γ is defined as all fixed components of the random coefficients in the model, including intercept, while $\underline{\delta}_j$ is defined as all random components of the random coefficients in the model, including the intercept. In summary: The first summation defines the fixed part of the model, and the second summation defines the random part.

This model is based on the assumptions of random school-level slopes, independent from each other but correlated with the random school-level intercepts. Error terms are correlated within contexts, because it is assumed that students in the same school share (unmeasured) characteristics based on their common environment. Because the model allows a variation among schools, it takes the intraclass correlation into account.

Intraclass correlation is defined in the literature (e.g., Cochran 1977; Searle et al. 1992) as the ratio of the between-class variance to the total variance. If the between-class variance (here, for instance, ω_{11} or ω_{00} or both) is equal to 0, the intraclass correlation is equal to 0. In fixed effects models, the omitted variables are summarized in the individual error term ($\underline{e}_{ij}$) only, while in the random models the part that relates to omitted variables based on shared experiences of observations within the *same class is taken out and considered a between-class variation either in intercepts or in slopes.*

Random slope and intercept co-vary only if they belong to the same school. Disturbances are uncorrelated between levels of the hierarchy. Disturbances of $\underline{e}_{ij}$ have the usual structure (IID) and are independent of the macro-errors. The metric of the dependent variable is at least ordinal, although some software for multilevel analysis such as VARCL (Longford 1991) and ML3 (Rasbash et al. 1989) allows for dichotomous dependent variables. The dependent variable is defined at the micro-level of the hierarchy. Observations within the same school have equal coefficients. The choice within RC models is to model all coefficients as random or some as fixed and some as random. In the present example, the choice is between a random intercept model only (and only one macro-error term, $\underline{\delta}_{0j}$) or a model with random intercept and a random slope (with both macro-error terms: $\underline{\delta}_{0j}+\underline{\delta}_{1j}X_{ij}$). A model has been chosen that allows all coefficients to be random. What to model as a fixed coefficient and what to model as a random coefficient ultimately should be decided by a replication of the study.

Estimation of means and variances in the RC model asks for more complicated computational methods than is the case in fixed effects models. In the RC model, the total variance is divided in the usual individual error variance (the micro-error variance of e_{ij}) but also in macro-level variances. Computational methods for the joint estimation of random and fixed parameters by means of the empirical Bayesian estimation methods can be found in the literature as well as in the manuals that accompany the four software packages now available: GENMOD (Mason et al. 1983), HLM (Bryk et al. 1988), ML3 (Rasbash et al. 1989), and VARCL (Longford 1991). For the next analysis, VARCL software will be used (Longford 1991).

Applications of the RC Model

The variables used in the following RC models are a combination of micro- and macro-level variables. At the student level, the variables are prealcohol use and postalcohol use. At the school level, the variables are NORM and mean alcohol level of the school. In the notation of equation (12) for random models, one single subscript was used for the gamma parameters, such as γ_0 and γ_1 (subscript 0 for first parameter or intercept, subscript 1 for the first slope coefficient). To enable indication of a parameter estimate for a macro-level variable, a second subscript is introduced. For instance, γ_{01} is the parameter estimate for the first macro-level variable, and γ_{02} is the parameter estimate for the second macro-level variable, and so on. Cross-level interactions of macro-level variables and student-level variables are treated equally. γ_{11} is used for the effect of the first macro-level variable (NORM, for instance) on the first micro-level variable (prealcohol use). The first 1 in the subscript is the first micro-level variable, and the second 1 in the subscript is the first macro-level variable. The effect of a second macro-variable in the model (mean alcohol level of schools, for instance) on the first micro-level variable (prealcohol use) would be γ_{12}, and so on.

Analysis 1

In the first model, it is assumed that slopes and intercepts differ over schools (see equations [10] and [11]). The research question (see equation [14]) is: "Are slopes significantly different over schools, and is that difference explainable by NORM, the drug prevention program?" If NORM has a negative effect on intercepts, it could be concluded that NORM has an overall lowering effect on student postalcohol use. In all RC models reported in this chapter, both coefficients, the intercept, and the slope of the student or micro-model are defined as random. Although both coefficients are allowed to fluctuate among schools, in this first analysis an attempt is made only to explain the variation in the intercepts (not the variation in the slope) by introducing the macro-level variable NORM in equation (14) and figure 1:

$$\underline{\mathbf{Y}}_{ij} = \gamma_{00}+\gamma_{10}\mathbf{X}_{ij}+\gamma_{01}\mathbf{Norm}+(\underline{\delta}_{0j}+\underline{\delta}_{1j}\mathbf{X}_{ij}+\underline{\mathbf{e}}_{ij}) \qquad (14)$$

In equation (14), γ_{01} is the effect of NORM on the intercept (see the arrow in figure 1 that passes through the intercept, and equation [15] in the next technical [italicized] section). This effect of NORM on postalcohol use is the main effect. The complicated error term between parentheses reflects the fact that, next to the individual error, $\underline{e}_{ij}$, two macro errors are present in this model, $\underline{\delta}_{0j}$ for the intercept and $\underline{\delta}_{1j}X_{ij}$ for the slope. Note that the error related to the slope is associated with values for X and, as a result, has different values for different levels of the predictor. Figure 1 is based on equation (14).

Figure 1 (and all following figures) is organized as follows: the squares represent macro-level variables (here, NORM and the intercept), and the circles represent micro-level variables (here, prealcohol and postalcohol). If an arrow leaving a macro-level variable passes through a square representing the intercept, it shows a direct effect as a function of the intercept ($\underline{a}_j$'s in equation [15]). The parameter estimates for equation (14) are on the following page (with standard scores between parenthesis).

$$\hat{Y}_{ij} = \mathbf{-0.01} + \underset{(z\,=\,16.18)}{\mathbf{0.52X_{ij}}} \underset{(z\,=\,2.24)}{\mathbf{-0.04NORM}}$$

Again, prealcohol use is significantly related to postalcohol use (z = 16.18), and the drug prevention program NORM has a signif-

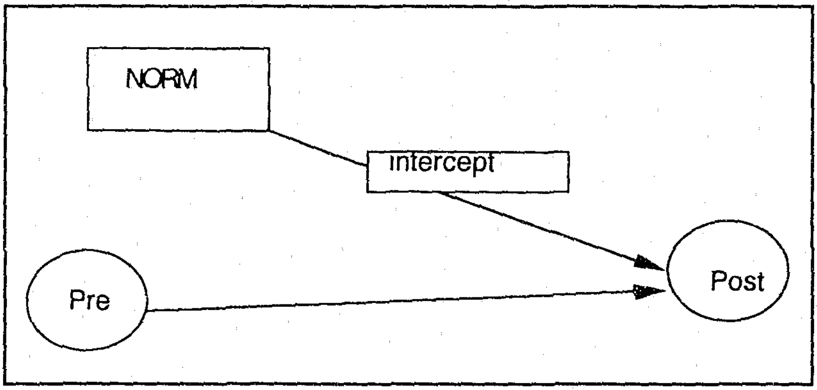


FIGURE 1. *The effect of the drug prevention program NORM*

icant and negative effect on postalcohol use, showing that NORM has an overall lowering effect on alcohol use of students. The analysis results for the macro-level variances (the variances of the macro-level errors, $\underline{\delta}_{0j}$ and $\underline{\delta}_{1j}$, in equations [15] and [16]) show significant differences in slopes and intercepts across schools, which may be explainable by macro-level variables. For more explicit details concerning values of macro-level variances and their respective z-tests, refer to summary table 6 at the end of this section.

A technical summary of the RC model in equation (14)

The model fitted in figure 1 is based on the micro-equation in equation (9), repeated here:

$$\underline{Y}_{ij} = \underline{a}_j + \underline{b}_j X_{ij} + \underline{e}_{ij} \qquad (9)$$

Again, $\underline{Y}_{ij}$ is the individual student variable, postalcohol use of student i within school j, while X_{ij} is the individual-level predictor variable, prealcohol use of the same observation. For the individual-level disturbance, $\underline{e}_{ij}$ is used. In this simple example, all coefficients are random.

The next step in the modeling process is to specify the properties of the random slopes and intercepts. The estimates for slope and intercept are divided into a fixed part and a random part (see equations [15] and [16]). The random parts or variance components are school-level disturbances, with expectation 0. The school-level disturbances are assumed to be independent of the student-level disturbances, $\underline{e}_{ij}$. So $\underline{a}_j$ and $\underline{b}_j$ in equation (14) are defined in equations (15) and (16) as:

$$\underline{a}_j = \gamma_{00} + \gamma_{01}\, NORM + \underline{\delta}_{0j} \qquad (15)$$

$$\underline{b}_j = \gamma_{10} + \underline{\delta}_{1j} \qquad (16)$$

For the gammas (γ_{00}, γ_{01}, and γ_{10}) in equations (15) and (16), the sub-script is defined for the first index as the number of the variable at the micro-level, and the second index is the number of the variable at the *macro-level. This means that* γ_{st} *is the effect of the macro-level* $\underline{t}$ *on the regression coefficients of micro-variable* $\underline{s}$*. Zero is the intercept (i.e., the variable with all values equal to +1, either at the micro-level or on the macro-level). For instance,* γ_{01} *is the effect of the drug prevention program NORM on the micro-level coefficient of the intercept (see figure 1). If the decompositions of random intercept and slope in equations (15) and (16) are substituted in micro-model (9), equation (14) is obtained:*

$$\underline{Y}_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}NORM + (\underline{\delta}_{0j} + \underline{\delta}_{1j}X_{ij} + \underline{e}_{ij}) \qquad (14)$$

Equation (14) shows that the variance in Y_{ij} *is decomposed in a fixed part (*$\gamma_{00}+\gamma_{10}+\gamma_{01}$*) and a random part (*$\underline{\delta}_{0j}+\underline{\delta}_{1j}X_{ij}+\underline{e}_{ij}$*). The random part includes two school-level disturbances (*$\underline{\delta}_{0j}+\underline{\delta}_{1j}X_{ij}$*) and one individual-level disturbance (*$\underline{e}_{ij}$*), which is the usual individual error term. The random part contains a disturbance, which is related to the variable* X_{ij}*. This disturbance shows that the covariance structure is more complicated than researchers are used to seeing in fixed effects linear models. Model (14) again resembles the usual linear regression model, only with a complicated error term. The variances of the random components are called the variance components of the model (hence, the name VARiance Component anaLysis [VARCL] for the computer program applied to this body of data).*

Analysis 2

The next RC model replaces the macro-level variable NORM with the macro-level variable alcohol mean ($\overline{X}_j$). The research question is equal to the one used in the separate analysis of the fixed linear model: "Does the mean alcohol level of schools have an effect on the alcohol use of the students, and does a cross-level interaction exist?" Figure 2 is based on equation (17), where the main effect of mean alcohol use is reflected in the γ_{01} for $\overline{X}_j$. The cross-level interaction effect of the same variable with prealcohol use of the student is reflected in the γ_{11} for $\overline{X}_j$ $X_{i.j}$ (for details, see the following technical summary, equations [18] and [19]; both equations contain the school mean).

$$\underline{Y}_{ij} = \gamma_{10}X_{ij}+\gamma_{01}X_{.j}+\gamma_{11}X_{.j}X_{ij}+(\underline{\delta}_{0j}+\underline{\delta}_{1j}X_{ij}+\underline{e}_{ij}) \qquad (17)$$

The results of model (17) and figure 2 are:

$$\hat{Y}_{ij} = \mathbf{0.06} + \underset{(z\,=\,17.20)}{\mathbf{0.78X_{ij}}} - \underset{(z\,=\,0.59)}{\mathbf{0.15\,\overline{X}_{.j}}} - \underset{(z\,=\,8.55)}{\mathbf{0.09X_{ij}}}$$

The effect of prealcohol on postalcohol use is significant as usual (z = 17.20). The effect of the mean alcohol level of schools has no direct effect, with a value of -0.15 and a z-score of 0.59, but its cross-level interaction with student alcohol use is significant, with a value of -0.09 and a z-value of 8.55. The conclusion based on this analysis is that school mean has a negative effect on alcohol use but only as a cross-level interaction. The arrow in figure 2 shows that this negative effect lowers the strength of the relationship between prealcohol and postalcohol use of the student. In other words, the higher the school mean for alcohol use, the lower the magnitude of the coefficient for the prediction of post-alcohol by prealcohol use of students. In the next model, the macro-variable NORM is added to test if the effect of the mean alcohol level is due partly to drug prevention program NORM or is a separate, unrelated effect.

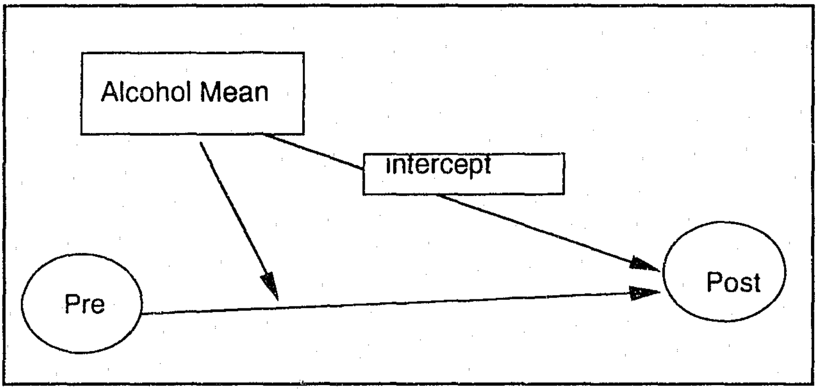


FIGURE 2. *The effect of mean alcohol level of schools*

A technical summary of the RC model in equation (17)

If the same micro-level equation (9) is used as before,

$$\underline{Y}_{ij} = \underline{a}_j + \underline{b}_j X_{ij} + \underline{e}_{ij} \tag{9}$$

where

$$\underline{a}_j = \gamma_{00} + \gamma_{01} \overline{X}_j + \underline{\delta}_{0j} \tag{18}$$

and

$$\underline{b}_j = \gamma_{10} + \gamma_{11} \overline{X}_j + \underline{\delta}_{1j} \tag{19}$$

Substituting equations (18) and (19) in equation (9) and rearranging terms yields equation (17).

Analysis 3

The next model tests the hypothesis that both macro-variables, NORM and school mean (of the prealcohol consumption of students), have an effect on the micro-relation of prealcohol and postalcohol (see figure 3). It is assumed that the drug prevention program NORM lowers the use of alcohol in general, but the mean (pretest) alcohol level of schools is assumed to have a weakening effect on the relationship of prealcohol to postalcohol use in the sense that the higher the mean alcohol level of a school, the less alcohol use at Time 1 predicts alcohol use at Time 2. The RC model used (see equation [20]) again is based on the micro-model in equation (9). In equations (21} and (22) and in figure 3, it is shown that student prealcohol and postalcohol use is related to both school-level variables. Figure 3 shows by way of arrows that the mean alcohol level of schools interacts with prealcohol use of the student but is not related to the intercept. The school mean in this model is related only to the slope and not to the intercept (reflected in equations [21] and [22] in the following technical summary). The final model is on the following page.

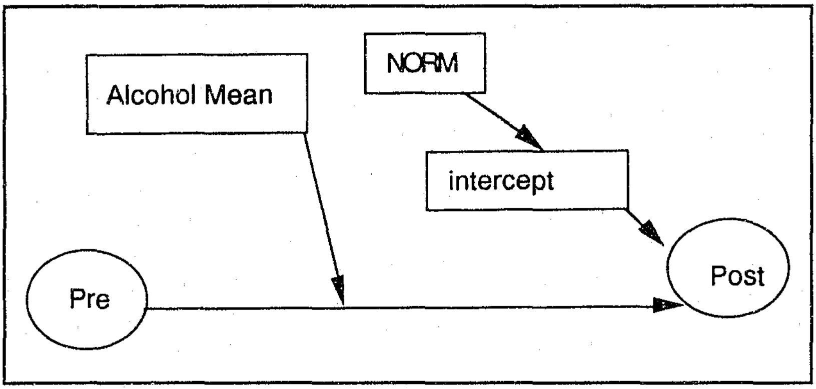


FIGURE 3. *The effect of NORM and mean alcohol level of schools*

$$\underline{Y}_{ij} = \gamma_{00}+\gamma_{10}X_{ij}+\gamma_{01}\text{NORM}+\gamma_{12}\overline{X}_{.j}X_{ij}+(\underline{\delta}_{0j}+\underline{\delta}_{1j}X_{ij}+\underline{e}_{ij}) \tag{20}$$

where γ_{01} is the main effect of NORM and γ_{12} is the cross-level interaction effect of the mean alcohol level ($\overline{X}_{j}$) and the student alcohol use at Time 1, (X_{ij}).

The results based on model (20) and figure 3 are:

$$\hat{Y}_{ij} = \mathbf{0.06}+\underset{(z=17.37)}{\mathbf{0.78X_{ij}}}\underset{(z=2.20)}{\mathbf{-0.04NORM}}\underset{(z=8.35)}{\mathbf{-0.09\,\overline{X}_{j}X_{ij}}}$$

Model (20) shows the familiar solutions, where pretest is significantly related with posttest (z = 17.37), and NORM and the school mean of prealcohol use are significant with, respectively, z = 2.20 and z = 8.35. Both macro-variables have a surpressing (negative) effect on student alcohol use.

A technical summary of the RC model in equation (20)

To understand in more detail how equation (20) was formulated, start again with the basic micro-level equation (9):

$$\underline{Y}_{ij} = \underline{a}_{j}+\underline{b}_{j}X_{ij}+\underline{e}_{ij} \tag{9}$$

In the model are two macro-level variables, NORM and school mean ($\overline{X}_j$). The first has an effect only on the intercept, as is formulated in equation (21):

$$\underline{a}_j = \gamma_{00} + \gamma_{01} NORM + \underline{\delta}_{0j} \qquad (21)$$

while the second macro-variable has an effect only on the slope, as is formulated in equation (22):

$$\underline{b}_j = \gamma_{10} + \gamma_{12}\overline{X}_{.j} + \underline{\delta}_{1j} \qquad (22)$$

Substituting equations (21) and (22) in equation (9) and rearranging terms yields equation (20) on the following page.

$$\underline{Y}_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}NORM + \gamma_{12}\overline{X}_{.j}X_{ij} + (\underline{\delta}_{0j} + \underline{\delta}_{1j}X_{ij} + \underline{e}_{ij}) \qquad (20)$$

Note the notation for the gammas associated with the macro-variable. For the effect of NORM, γ_{01}, the first subscript (0) relates to the intercept, the second subscript (1) indicates the first macro-variable, which is NORM. In the same fashion, the cross-level interaction effect, γ_{12}, has two subscripts; the 1 indicates the first micro-level variable (prealcohol use) and the 2 indicates that $\overline{X}_j$ is the second macro-level variable.

In table 6, the results of the three RC models are summarized. All RC models in this table show that the effect of the micro-level relation pre-alcohol on postalcohol use (γ_{10}) remains equally significant, with z-values around 17.00. The effect of NORM (γ_{01}) in models (8) and (10) is of equal strength, with close to equal z-tests (around 2.20). The cross-level interaction effects in models (9) and (10) also are comparable, with equally strong and equally significant values (both z-values are around 8.55). The macro-error variances (ω's) are close in magnitude over all three models.

The following conclusion is supported by all three analyses: The drug prevention program NORM lowers alcohol use, while the interaction effect between student and school alcohol mean is significant. The last

TABLE 6. *A summary table of three RC models*

Micro-model: $\underline{Y}_{ij} = \underline{a}_j+\underline{b}_jX_{ij}+\underline{e}_{ij}$

NORM MODEL	RC MODELS with CROSS-LEVEL interaction	
Model 14	Model 17	Model 20
$\underline{a}_j = \gamma_{00}+\gamma_{01}\text{NORM}+\underline{\delta}_{0j}$	$\underline{a}_j = \gamma_{00}+\gamma_{01}\overline{X}_{.j}+\underline{\delta}_{0j}$	$\underline{a}_j = \gamma_{00}+\gamma_{01}\text{NORM}+\underline{\delta}_{0j}$
$\underline{b}_j = \gamma_{10}+\underline{\delta}_{1j}$	$\underline{b}_j = \gamma_{10}+\gamma_{11}\overline{X}_{.j}+\underline{\delta}_{1j}$	$\underline{b}_j = \gamma_{10}+\gamma_{12}\overline{X}_{.j}+\underline{\delta}_{1j}$
$\hat{Y}_{ij} = -0.01+0.52X_{ij}$	$\hat{Y}_{ij} = -0.06+0.78X_{ij}-$ $0.15\,\overline{X}_{.j}-$ $0.09(X_{ij}\,\overline{X}_{.j})$	$\hat{Y}_{ij} = -0.06+0.78X_{ij}-$ $0.04\text{NORM}-$ $0.09(X_{ij}\,\overline{X}_{.j})$

	Model 14 parameter estimate	z-test	Model 17 parameter estimate	z-test	Model 20 parameter estimate	z-test
γ_{00}	-0.01			0.06		0.06
γ_{10}	0.52	16.18	0.78	17.20	0.78	17.37
γ_{01}	-0.04	2.24	-0.15	0.59	-0.04	2.20
γ_{11}	a	a	-0.09	8.55	-0.09	8.53
Variances and Co-Variances						
ω_{00}	0.002	2.92	0.01	4.23	0.003	3.16
ω_{11}	0.01	4.05	0.01	4.33	0.01	4.14
ω_{10}	0.01	3.52	0.01	4.11	0.01	3.80

NOTE: a = absent

result indicates that, in high-mean schools, more students change for the better, after correcting for NORM, than in low-mean schools. It is worth investigating in further analyses whether this is the result of a ceiling effect or a new fact. For instance, an interaction effect would exist between NORM and mean alcohol level of a school, with the implication that prevention programs are more effective in schools with a high mean alcohol consumption than in schools with a low mean alcohol consumption.

A technical summary of the RC model in relation to the macro-error components

Note that the error term stayed the same over all three random models (14), (17), and (20) used in this chapter. The two coefficients, the intercept, and the slope are defined as random throughout the analyses. Within the available software for RC models (see last section for a list of available software), choices can be made about which first-level coefficient is fitted as random and which is fitted as fixed. The choice is anything between the two extremes: all coefficients random, or only a random intercept. For this analysis, with only two coefficients and one individual-level regressor, the choice was between defining the coefficient for the prealcohol slope as random or as fixed. The error term in all models of table 6 is $(\underline{\delta}_{0j}+\underline{\delta}_{1j}X_{ij}+\underline{e}_{ij})$. *A model with a fixed instead of a random slope would have a less complicated error term* $(\underline{\delta}_{0j}+\underline{e}_{ij})$ *because the macro-error for the slope* $(\underline{\delta}_{1j}X_{ij})$ *is not present in the model estimation. The last error term is comparable to the definition of the error structure in variance component models (e.g., Searle et al. 1992).*

SUMMARY

As is shown in the literature, researchers have struggled for some time with concepts such as hierarchically nested observations, intraclass correlation, the unit of analysis, and random instead of fixed factors. The problems for experimental researchers are summarized in Anderson and Ager (1978), Crits-Christoph and Mintz (1991), and Murray and

Hannan (1990). Traditional analysis models are limited in the way they solve the technical problems of nested designs. They also are limited in the questions they can address. RC models provide more reliable solutions for nested designs with unbalanced data and take the intraclass correlation into account. By estimating random instead of fixed effects, these models acknowledge the fact that the design has random factors instead of a fixed number of treatments. The treatments can be real treatments but more often are defined as groups within treatments. The random processes taking place within groups are modeled as random effects.

The RC model is a useful extension of the traditional variance component models as discussed in Searle and colleagues (1992) and Winer (1971). For drug prevention researchers, the model offers the possibility to make use of within-school differences in parameter estimates by turning it from a within-group error (or nuisance) into a meaningful source of variation.

Questions do remain, however. For example: Are RC models more powerful than traditional methods? This question never really is addressed, and should be. If RC models prove to be less powerful, why would a researcher go through the trouble of learning another technique if loss of power is the tradeoff of a statistically more "correct" model? After all, researchers evaluating real-life experiments are more interested in the promises and the usefulness of RC models than in the statistical correctness of such models.

Aside from the unresolved issue of power, what do RC models offer that others do not? The promise of the RC model is that it can help build theories that predict the effect of drug prevention programs for special groups of students, in the sense that some programs may work well for some students but not for others. The attractiveness of the RC model is that it estimates effects over all schools together, it is parsimonious, and it can test macro-effects in combination with micro-effects and their cross-level interactions.

As illustrated in this chapter, traditional models have their specific problems for drug prevention research, which can be solved by using RC models. Problems are found in the aggregate model, which measures schools instead of students, while students are the object of interest in drug prevention programs. ANCOVA problems are intraclass correlation, pooled within-slope estimate, and no opportunity to introduce macro-level characteristics to explain school differences. The separate models for separate schools or slopes as outcomes approach (Burstein et al. 1978) is cumbersome, too general, and not parsimonious.

The decision to use RC models and how to decide if a set of effects is fixed or random depends on several things. These include the context of the data, the manner in which the data are collected, the environment from which they come and, most importantly, the inferences that are made based on the analysis to groups, to students, or to types of students. Last but not least, theories are needed that state meaningful relationships between individual characteristics and contexts. Theories are needed that can help find aptitude/treatment interactions as advocated by Cronbach (1957, p. 679):

> The job of applied psychology is to improve decisions about people. The greatest social benefit will come from applied psychology if we can find for each individual the treatment to which he can most easily adapt. This calls for the joint application of experimental and correlational methods.

The problems mentioned above are familiar problems to new techniques being developed, but it is an active area of research in the statistical and educational communities, and there are hopeful signs that some, perhaps many, of these problems will be solved in the next few years.

ANALYSIS PACKAGES FOR THE ANALYSIS OF MULTILEVEL DATA USING HIERARCHICAL LINEAR MODELS

Multilevel modeling software now has become readily available, although under different names. One package clearly used the already existing random effects model from the experimental research tradition (Dempster et al. 1981) by naming the software package VARiance Component anaLysis (VARCL). Others (Bryk et al. 1988) had a class of substantive problems out of the observational research tradition in mind and named their package Hierarchical Linear Models (HLM). Rasbash and colleagues (1989) highlighted the way the data are collected at three levels of the hierarchy by naming their package ML3, where the name "Multilevel" is combined with the number three, the number of hierarchies the package is able to handle. ML3 and VARCL allow for three levels of nesting, while GENMOD and HLM allow for two levels of nesting. These programs are described in more detail below:

GENMOD was written by Hermalin and Anderson at the Population Studies Center, University of Michigan, from instructions provided by Mason and colleagues (1983).

HLM, Version 2.20, was written by Bryk and colleagues. They also have written a manual for its use (Bryk et al. 1988).

ML3, Version 2.2, is software for two- or three-level analysis written by Rasbash. The manual is by Rasbash and colleagues (1989). The program is based on theoretical work by Goldstein. Prosser and colleagues (1991) have written a booklet on data analysis with ML3.

VARCL was initiated by Aitkin and Longford (1986) and was written and is maintained by Longford. Longford (1991) has written a manual to accompany the program.

NOTES

1. Technical summary sections may be skipped without loss of continuity.

REFERENCES

Aitkin, M.A., and Longford, N. Statistical modelling in school effectiveness studies. *J R Stat Soc* 149A:1-43, 1986.

Anderson, L.R., and Ager, J.W. Analysis of variance in small group research. *Pers Soc Psychol Bull* 4(2):341-345, 1978.

Barcikowski, R.S. Statistical power with group mean as the unit of analysis. *J Educ Stat* 6(3):267-285, 1981.

Brook, J.S.; Brook, D.W.; Gordon, A.S.; Whiteman, M.; and Cohen, P. The psychosocial etiology of adolescent drug use: A family interactional approach. *Genet Soc Gen Psychol Monogr* 116(2):111-267, 1990.

Bryk, A.S.; Raudenbush, S.W.; Seltzer M.; and Congdon, R.T. *Manual for HLM.* Chicago: Scientific Software, 1988.

Burstein, L. The analysis of multilevel data in educational research in evaluation. *Rev Res Educ* 8:158-233, 1980.

Burstein, L.; Linn, R.L.; and Capell, F.J. Analyzing multilevel data in the presence of heterogeneous within-class regressions. *J Educ Stat* 3:347-383, 1978.

Cochran, W.G. *Sampling Techniques.* 3d ed. Toronto: Wiley, 1977.

Crits-Christoph, P., and Mintz, J. Implications of therapist effects for the design and analysis of comparative studies of psychotherapies. *J Consult Clin Psychol* 59(1):20-26, 1991.

Cronbach, L.J. The two disciplines of scientific psychology. *Am Psychol* 12(11):671-684, 1957.

Cronbach, L.J., and Webb, N. Between class and within class effects in a reported aptitude x treatment interaction: A reanalysis of a study by G.L. Anderson. *J Educ Psychol* 67:717-724, 1975.

Dempster, A.P.; Rubin, D.B.; and Tsukawa, R.K. Estimation in covariance components models. *J Am Stat Assoc* 76:341-353, 1981.

Hansen, W.B., and Graham, J.W. Preventing alcohol, marijuana, and cigarette use among adolescents: Peer pressure resistance training versus establishing conservative norms. *Prev Med* 20:414-430, 1991.

Hawkins, D.J.; Catalano, R.F.; and Miller, J.Y. Risk and protective factors for alcohol and other drug problems in adolescence and early adulthood. *Psychol Bull* 112(1):64-105, 1992.

Kreft, I.G.G., and De Leeuw, E.D. The seesaw effect: A multilevel problem? *Qual Quant* 22:127-137, 1988.

Longford, N.T. *Manual for VARCL.* Princeton, NJ: Educational Testing Service, 1991.

Mason, W.M.; Wong, G.Y.; and Entwisle, B. Contextual analysis through the multilevel linear model. In: Leinhart, S., ed. *Sociological Methodology.* San Francisco: Jossey-Bass, Inc., Publishers, 1983. pp. 72-103.

Murray, D.M., and Hannan, P.J. Planning for the appropriate analysis in school-based drug-use prevention studies. *J Consult Clin Psychol* 58(4):458-468, 1990.

Prosser, R.; Rasbash, J.; and Goldstein, H. *Data Analysis With ML3.* London: Institute of Education, University of London, 1991.

Rasbash, J.; Prosser, R.; and Goldstein, H. *ML3, Software for Three-Level Analysis.* Manual. London: University of London, Institute of Education, 1989.

Robinson, W.S. Ecological correlations and the behavior of individuals. *Am Sociol Rev* 15:351-357, 1950.

Rutter, M. Resilience in the face of adversity: Protective factors and resistance to psychiatric disorder. *Br J Psychiatry* 147:598-611, 1985.

Searle, S.R.; Casella, G.; and McCulloch, C.E. *Variance Components.* New York: Wiley, 1992.

Winer, B.J. *Statistical Principles in Experimental Design.* 2d ed. New York: McGraw-Hill, 1971.

Sources Providing an Introductory Overview of Multilevel Modeling

Aitkin, M.A., and Longford, N. Statistical modelling in school effectiveness studies. *J R Stat Soc* 149A:1-43, 1986.

Bryk, A.S., and Raudenbush, S.W. Applying the hierarchical linear model to measurements of change problems. *Psychol Bull* 101:147-158, 1987.

De Leeuw, J., and Kreft, I.G.G. Random coefficient models for multilevel analysis. *J Educ Stat* 11:57-85, 1986.

Goldstein, H. Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika* 73:43-56, 1986.

Hox, J.J., and Kreft, I.G.G. Multilevel analysis methods. *Sociol Meth Res* 22(3):283-300, 1994.

Kreft, I.G.G., and De Leeuw, J. The gender gap in earnings: A two-way nested multiple regression analysis with random effects. *Sociol Meth Res* 22(3):319-341, 1994.

Kreft, I.G.G.; De Leeuw, J.; and Kim, K. *Comparing Four Different Statistical Packages for Hierarchical Linear Regression: GENMOD, HLM, ML3, and VARCL.* CSE Technical Report 311. Los Angeles: UCLA, Center for Research on Evaluation, Standards, and Student Testing, 1990.

Mason, W.M.; Wong, G.Y.; and Entwisle, B. Contextual analysis through the multilevel linear model. In: Leinhart, S., ed. *Sociological Methodology.* San Francisco: Jossey-Bass, Inc., Publishers, 1983. pp. 72-103.

Searle, S.R.; Casella, G.; and McCulloch, C.E. *Variance Components.* New York: Wiley, 1992.

Sources Providing Empirical Applications of Multilevel Modeling

Bock, R.D., ed. *Multilevel Analysis of Educational Data.* San Diego: Academic Press, 1988.

Hox, J.J.; Kreft, I.G.G.; and Hermkes, P.L.J. Multilevel models for factorial surveys. *Sociol Meth Res* 19(4):493-511, 1991.

Kreft, I.G.G., and De Leeuw, J. Model based ranking of schools. *Int J Educ Res* 15(1):45-61, 1991.

Kreft, I.G.G., and De Leeuw, E.D. The seesaw effect: A multilevel problem? *Qual Quant* 22:127-137, 1988.

Kreft, I.G.G.; De Leeuw, J.; and Kim, K. *Comparing Four Different Statistical Packages for Hierarchical Linear Regression GENMOD, HLM, ML3, and VARCL.* CSE Technical Report 311. Los Angeles: UCLA, Center for Research on Evaluation, Standards, and Student Testing, 1990.

Raudenbush, S.W., and Wilms, J.D., eds. *Review of Schools, Classrooms and Pupils: International Studies of Schooling From a Multilevel Perspective.* San Diego: Academic Press, 1991.

AUTHOR

Ita G.G. Kreft, Ph.D.
Associate Professor
School of Education
Division of Educational Foundations
California State University, Los Angeles
5151 University Drive
Los Angeles, CA 90032-8143

Seven Ways To Increase Power Without Increasing N

William B. Hansen and Linda M. Collins

ABSTRACT

Many readers of this monograph may wonder why a chapter on statistical power was included. After all, by now the issue of statistical power is in many respects mundane. Everyone knows that statistical power is a central research consideration, and certainly most National Institute on Drug Abuse grantees or prospective grantees understand the importance of including a power analysis in research proposals.

However, there is ample evidence that, in practice, prevention researchers are not paying sufficient attention to statistical power. If they were, the findings observed by Hansen (1992) in a recent review of the prevention literature would not have emerged. Hansen (1992) examined statistical power based on 46 cohorts followed longitudinally, using nonparametric assumptions given the subjects' age at posttest and the numbers of subjects. Results of this analysis indicated that, in order for a study to attain 80-percent power for detecting differences between treatment and control groups, the difference between groups at posttest would need to be at least 8 percent (in the best studies) and as much as 16 percent (in the weakest studies). In order for a study to attain 80-percent power for detecting group differences in pre-post change, 22 of the 46 cohorts would have needed relative pre-post reductions of greater than 100 percent. Thirty-three of the 46 cohorts had less than 50-percent power to detect a 50-percent relative reduction in substance use. These results are consistent with other review findings (e.g., Lipsey 1990) that have shown a similar lack of power in a broad range of research topics. Thus, it seems that, although researchers are aware of the importance of statistical power (particularly of the necessity for calculating it when proposing

research), they somehow are failing to end up with adequate power in their completed studies.

This chapter argues that the failure of many prevention studies to maintain adequate statistical power is due to an overemphasis on sample size (N) as the only, or even the best, way to increase statistical power. It is easy to see how this overemphasis has come about. Sample size is easy to manipulate, has the advantage of being related to power in a straightforward way, and usually is under the direct control of the researcher, except for limitations imposed by finances or subject availability. Another option for increasing power is to increase the alpha used for hypothesis-testing but, as very few researchers seriously consider significance levels much larger than the traditional .05, this strategy seldom is used.

Of course, sample size is important, and the authors of this chapter are not recommending that researchers cease choosing sample sizes carefully. Rather, they argue that researchers should not confine themselves to increasing N to enhance power. It is important to take *additional* measures to maintain and improve power over and above making sure the initial sample size is sufficient. The authors recommend two general strategies. One strategy involves attempting to maintain the effective initial sample size so that power is not lost needlessly. The other strategy is to take measures to maximize the third factor that determines statistical power: effect size.

MAINTAINING EFFECTIVE SAMPLE SIZE

Preventing Attrition

One of the best ways to increase power without increasing N is to avoid decreasing N through attrition. Of course, attrition has other consequences besides loss of power, such as internal and external validity problems. However, independent of these problems, a loss of subjects through attrition is accompanied by a loss of statistical power.

Many articles about attrition (Biglan et al. 1987; Ellickson et al. 1988; Hansen et al. 1990; Pirie et al. 1989) have helped to alert the research community to the potential causes of attrition so that measures can be taken to prevent it. Attrition has many and varied causes. Sometimes the causes are as simple as subjects moving out of the school district where the study is taking place. Usually, though, the causes are more complex and not totally unrelated to the study. In some studies where the treatment is aversive in some way, treatment group subjects drop out at a higher rate; in other studies where the treatment is a plum and nothing is done to compensate the control group, the opposite occurs. In substance use prevention studies, high-risk subjects are more likely to drop out (Hansen et al. 1985). Attrition can even reflect a political problem as, for example, when an institution like a school or a school district drops out of a study (Hansen et al. 1990). Researchers should become familiar with the studies that have examined retention of subjects (Ellickson et al. 1988) and political units (Goodman et al. 1991; O'Hara et al. 1991) to gain an understanding of how to manage attrition in practical terms. Every prevention effort should include funds in its budget for tracking and collecting data from subjects who have dropped out of the study.

Missing Data Analysis

Missing data analysis (Graham et al., this volume) is an exciting new data analysis strategy that recovers some (but not all) of the loss of power incurred through attrition. This is not a way of replacing missing data; rather, it is a way of making the most out of the remaining data. This methodology provides a way for the user to model the mechanisms behind attrition, allowing for estimation of what the results would have been if the full sample had been maintained. The chapter by Graham and colleagues (this volume) presents an in-depth look at this important topic.

MAXIMIZING EFFECT SIZE

Take a closer look at effect size:

$$\text{effect size} = \frac{\mu_A - \mu_B}{\sigma} \tag{1}$$

The numerator of equation (1) is the difference between the population mean for the treatment group (μ_A) and the population mean for the control group (μ_B). The denominator is the population variance (assuming homogeneity of variance, that is, the two populations have identical variances). The strategies suggested here are intended to increase effect size either by increasing the size of the numerator of equation (1), that is, increasing the difference between the mean of the treatment group and the mean of the control group, or decreasing the denominator of equation (1), that is, decreasing the population variance.

STRATEGIES TO INCREASE THE MAGNITUDE OF GROUP DIFFERENCES

Targeting (and Affecting) Appropriate Mediators

All prevention programs seek to change behavior by changing some mediating process. The choice of which mediating process to intervene on is the key to a powerful intervention. Only if the researchers developing a program understand the basic underlying processes that account for substance use behavior can they hope to identify the most appropriate mediators. Such an understanding is gained by examining very carefully and thoroughly existing theory and empirical evidence about the modifiable predictors and determinants of substance use behavior.

For example, in a series of studies conducted by Hansen and colleagues (1988, 1991), two mediating processes were targeted: the development of normative beliefs intolerant of alcohol and drug use and the development of skills for resisting overt offers to use substances. Two programs were compared, each designed to address one mediator specifically and, to the extent possible, to not affect the mediator associated with the other program. The results consistently have shown success in achieving differential impacts on behavior.

Some program developers prefer a less systematic emphasis on which a mediator is targeted for change, basing program content and strategy on strongly held personal beliefs rather than on empirical evidence about which components offer potential for change in particular mediators. Such programs developed solely from instinct or good intentions will, over the long run, fail to have as much power as programs developed more scientifically.

Identifying the appropriate mediators is a necessary but not sufficient condition for increasing statistical power—the intervention must be strong enough to have an effect, ideally a large one, on the mediators. It is difficult to give advice on how to achieve this goal. It seems that, even at their bést, researchers have little more than an intuitive understanding of what it takes programmatically to change mediating processes. Although the literature in this area can be of some help, the best methods for reaching school-age children change constantly. The impact of interventions probably could be increased, thereby increasing statistical power, by making better use of input from the people who know best how to teach youth, namely teachers, counselors, and youth workers.

Maintaining Program Integrity

Program integrity, the degree to which the program is adhered to in delivery, has predictable effects on outcome (Botvin et al. 1990; Hansen et al. 1991; Pentz et al. 1990); when program integrity is compromised, the treatment is less effective and differences between treatment and control shrink. Researchers have yet to develop a complete understanding of program integrity. For example, integrity to date has been defined by researcher standards rather than target audience-centered standards. Researchers may need to account for issues that they have not considered when defining integrity, such as the need to tailor a program for specific audiences.

For some programs, there is a tradeoff between N and program integrity. In fact, Tobler (1993) found in a meta-analysis that effect size was reduced in prevention studies involving more than 400 subjects per condition. If the sample size is so large that a large staff must be hired to deliver the program and the researcher, therefore, cannot be highly selective about this staff and cannot supervise them closely, it is unlikely that the program will be delivered uniformly well. It is important for the researcher to be aware of this tradeoff, because there may be times when power is maximized in the long run by choosing a smaller N and a more manageable intervention.

Appropriate Timing of Longitudinal Followup

The magnitude of the difference between treatment and control groups partly is a function of the length of time between program implementation and followup. Hansen (1992) concluded that many prevention studies are conducted for too short a period of time. Prevention researchers sometimes argue that long-term impacts cannot be expected from prevention programs. The authors disagree for two reasons. First, the goal of prevention is to maintain existing nonbehavior. There is reason to be much more sanguine about the possibility of prevention to have long-term effects, especially if the forces that foster experimentation with alcohol and drugs have really been changed. Second, the outcome of interest in prevention studies is based, not only on the treatment group maintaining its level of use or nonuse, but on the control group changing its behavior. Since this change takes time, it makes sense to measure behavioral outcomes repeatedly over a long period of time in order to increase the potential for observing differences between treatment and control groups when they reach their peak. For more about timing of observations and its effects on results, see Cohen (1991) and Collins and Graham (1991).

STRATEGIES FOR REDUCING VARIANCE

Sampling Control

There often is some pressure on prevention researchers to make sure the studies they are planning involve heterogeneous samples. There are two reasons for this. One reason is the need to maximize external validity. The more representative the sample is of the population at large, the better the external validity of the study is. The second reason is political; for example, it is important to make sure that women and minority groups are not excluded from prevention studies.

These two reasons for using heterogeneous samples are very good ones. However, researchers should balance these considerations with the effects of heterogeneity on statistical power. When heterogeneity is enhanced and homogeneity is diminished, power is reduced. The reason for this is straightforward: All else being equal, a heterogeneous population has more variance than a homogeneous population. Consider two populations with identical variances, σ^2, but with different means. If these two populations are combined into one, the new variance, σ_*^2, will be:

$$\sigma_*^2 = \sigma^2 + \left(\frac{\mu_A - \mu_B}{2} \right)^2 \qquad (2)$$

Thus, the larger the diffeıence in means between the two populations is, the larger the variance of the combined population will be. This larger variance results directly in a decreased effect size (see equation [1]) and, therefore, decreased power.

The problem is compounded if analyses then are conducted separately on subgroups in the data because these analyses necessarily will be based on a smaller N and may have dramatically reduced power. Where appropriate, covariates can be used to model subgroup differences. This maintains degrees of freedom and, therefore, can reduce the threat that sampling from heterogeneous groups brings.

Using Reliable and Appropriate Measures

The disciplines of psychology and epidemiology have both greatly influenced the field of substance use prevention research. These fields have different, and at times opposing, methodological traditions, particularly with respect to measurement. Epidemiology has emphasized relatively straightforward measurement and the use of manifest, and often dichotomous, variables. In contrast, psychology has a long tradition of measurement theory, emphasizing scale development, multiple indicator models, latent variables, and continuous variables. Classical test theory, including reliability theory, came from psychology.

An immediate question that is raised by contrasting these two approaches is, "Which is more appropriate, using continuous measures of substance use or using dichotomous measures?" Of course, the answer depends partly upon the research question that is being posed. The ramifications of this question for statistical power are complex. Cohen (1983) showed that dichotomizing a normally distributed continuous variable essentially throws away information and leads to a considerable loss of power. The situation is less clear with the skewed distributions that are more the rule in substance use prevention research. In general, though, unless the distributions are severely nonnormal, a loss of power can be expected if continuous variables are dichotomized.

It also is worth noting the relationship between measurement reliability and statistical power. This relationship is more complex than it may appear at first glance. Recall that according to classical test theory, the total variance in a measure is made up of true score variance and error variance. Measurement reliability is defined as the proportion of total variance that is made up of true score variance. Zimmerman and Williams (1986) showed that the direction of the relationship between reliability and power depends upon which of the three components—total variance, true score variance, or error variance—is held constant while the others are varied. If a constant true score variance is assumed, it follows that the greater the reliability (that is, the less error variance there is in a measure), the greater the statistical power will be. This is true because,

because, under these conditions, when the error variance decreases, the total variance decreases, resulting in a decrease to the denominator in equation (1). However, if a constant error variance is assumed, when reliability is increased, the true score variance is increased and, therefore, the total score variance is increased, leading to a decrease in power.

Zimmerman and Williams (1986) pointed out that the answer to this seeming paradox lies in how reliability is increased in practice. If a measure is improved by, say, discarding a few items that do not belong in the instrument, then generally this improves reliability by decreasing error variance. This strategy can be expected to improve statistical power. On the other hand, if reliability is improved by changing the sample so that it is more heterogeneous and, therefore, there is more true score variance, this is likely to result in an overall increase in variance and, hence, a loss of power.

CONCLUSIONS

This chapter argues that, while obtaining a sufficiently large sample is important, it is not all there is to statistical power. Other strategies are important if statistical power is to be maintained over the course of a substance use prevention study. The authors made seven suggestions for ways to improve power without increasing N in prevention research. Except for missing data analysis, none of these suggestions are new. Most of them are based on common sense, and many of them will be recognized as recommendations often made to colleagues and students. It is ironic that scientists, researchers, and social advocates have largely failed to use these principles systematically to improve the power of research. They persist in thinking of statistical power only in terms of sample size but must adopt a wider view, as suggested here.

The suggestions made here do not translate directly into formulas that can be inserted "as is" into proposals or research designs. Instead, they represent principles that can be used to guide decision-making in practice. In the end, it is not the proposal or the research report that is the essence of science, but increased understanding of the phenomenon of substance abuse and the procedures employed to prevent it. If researchers are ever to develop a thorough understanding of substance abuse and highly effective methods for preventing it, they must be aware of how research decisions affect statistical power.

REFERENCES

Biglan, A.; Severson, H.; Ary, D.; Faller, C.; Gallison, C.; Thompson, R.; Glasgow, R.; and Lichtenstein, E. Do smoking prevention programs really work: Attrition and the internal and external validity of an evaluation of a refusal skills training program. *J Behav Med* 10:613-628, 1987.

Botvin, G.J.; Baker, E.; Dusenbury, L.; Tortu, S.; and Botvin, E.M. Preventing adolescent drug abuse through a multimodal cognitive-behavioral approach: Results of a three-year study. *J Consult Clin Psychol* 58:437-446, 1990.

Cohen, J. The cost of dichotomization. *Appl Psychol Meas* 7:249-253, 1983.

Cohen, P. A source of bias in longitudinal investigations of change. In: Collins, L.M., and Horn, J.L., eds. *Best Methods for the Analysis of Change: Recent Advances, Unanswered Questions, Future Directions.* Washington, DC: American Psychological Association, 1991.

Collins, L.M., and Graham, J.W. Comments on "A source of bias in longitudinal investigations of change." In: Collins, L.M., and Horn, J.L., eds. *Best Methods for the Analysis of Change: Recent Advances, Unanswered Questions, Future Directions.* Washington, DC: American Psychological Association, 1991.

Ellickson, P.L.; Bianca, D.; and Shoeff, D.C. Containing attrition in school-based research: An innovative approach. *Eval Rev* 12:331-351, 1988.

Goodman, R.M.; Smith, D.W.; Dawson, L.; and Steckler, A. Recruiting school districts into a dissemination study. *Health Educ Res* 6:373-385, 1991.

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

Hansen, W.B.; Collins, L.M.; Malotte, C.K.; Johnson, C.A.; and Fielding, J.E. Attrition in prevention research. *J Behav Med* 8:261-275, 1985.

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* 10:559-579, 1987.

Hansen, W.B.; Graham, J.W.; Wolkenstein, B.H.; Lundy, B.Z.; Pearson, J.L.; Flay, B.R.; and Johnson, C.A. Differential impact of three alcohol prevention curricula on hypothesized mediating variables. *J Drug Educ* 18:143-153, 1988.

Hansen, W.B.; Graham, J.W.; Wolkenstein, B.H.; and Rohrbach, L.A. Program integrity as a moderator of prevention program effectiveness: Results for fifth grade students in the adolescent alcohol prevention trial. *J Stud Alcohol* 52:568-579, 1991.

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

Lipsey, M.W. *Design Sensitivity: Statistical Power for Experimental Research.* Newbury Park, CA: Sage Publications, 1990.

O'Hara, N.M.; Brink, S.; Harvey, C.; Harrist, R.; Green, B.; and Parcel, G. Recruitment strategies for health promotion research. *Health Educ Res* 6:363-371, 1991.

Pentz, M.A.; Trebow, E.A.; Hansen, W.B.; MacKinnon, D.P.; Dwyer, J.H.; Johnson, C.A.; Flay, B.R.; Daniels, S.; and Cormack, C. Effects of program implementation on adolescent drug use behavior: The Midwestern Prevention Project (MPP). *Eval Rev* 14:264-289, 1990.

Pirie, P.; Murray, D.M.; Peterson, A.V.; Thomson, S.J.; Mann, S.L.; and Flay, B.R. Tracking and attrition in longitudinal school-based smoking prevention research. *Prev Med* 18:249-256, 1989.

Tobler, N.S. "Meta-Analysis of Adolescent Drug Abuse Prevention Programs." Paper presented at the National Institute on Drug Abuse Technical Review on Meta-Analysis of Drug Abuse Prevention Programs, Bethesda, MD, July 26-27, 1993.

Zimmerman, D.W., and Williams, R.H. Note on the reliability of experimental measures and the power of significance tests. *Psychol Bull* 100:123-124, 1986.

AUTHORS

William B. Hansen, Ph.D.
Associate Professor
Department of Public Health Sciences
Bowman Gray School of Medicine
Medical Boulevard
Winston-Salem, NC 27157-1063

Linda M. Collins, Ph.D.
Professor
Department of Human Development and Family Studies
The Pennsylvania State University
University Park, PA 16802-6504

Designing and Analyzing Studies of Onset, Cessation, and Relapse: Using Survival Analysis in Drug Abuse Prevention Research

Judith D. Singer and John B. Willett[1]

ABSTRACT

Many questions arising in drug abuse prevention and intervention studies focus on *whether* and, if so, *when* events occur. When do adolescents start using drugs? Does participation in a drug prevention program at school decrease the risk that high school students will initiate drug use? Does failure to participate in a relapse prevention program at a community health center increase the risk that newly abstinent ex-abusers will start using drugs again? Research questions about event occurrence present unique design and analytic difficulties. The fundamental problem is how to handle censored observations, observations of those people who do not experience the target event during data collection. The methods of survival analysis overcome these difficulties and allow prevention researchers to describe patterns of occurrence, compare these patterns among groups, and build statistical models of the risk of occurrence over time.

In this chapter, the authors present a nonmathematical introduction to survival analysis for drug abuse prevention researchers. After developing the basic concepts, they focus on two topics—study design and data analysis—and identify for each the key issues researchers face and provide guidelines for making informed decisions about them. In the process, the authors review how prevention researchers have used the methods to date and point towards new directions for the application of these methods.

INTRODUCTION

Many questions arising in drug abuse prevention and intervention studies focus on *whether* and, if so, *when* events occur. Researchers investigating pathways into alcohol abuse, for example, have examined the age at first use (Adler and Kandel 1983), age at first abuse (Johnston et al. 1989), how long people continue to use alcohol over extended periods of time (Hawkins et al. 1991), how long successfully treated individuals remain abstinent before relapse (Hunt and General 1973), and whether participation in a treatment program affects the risk of relapse (Cooney et al. 1991). Similar questions about event occurrence arise in studies of the onset, cessation, and relapse of other addictions (e.g., illicit drugs, smoking, gambling, and criminal activities), as well as studies of the efficacy of interventions in the prevention of drug use and addiction and the effects of drug use on other event outcomes, such as unemployment, premarital pregnancy, suicide, and withdrawal from school.

Research questions about event occurrence present unique design and analysis difficulties. The core problem is that, no matter when data collection begins and no matter how long any subsequent followup lasts, some people may not experience the target event before data collection ends—some current nonusers may not initiate drug use, some current users may not quit, and some former users may not relapse. Should the researcher assume that none of these people will ever experience the event? All the researcher knows is that, by the end of data collection, usually an arbitrary point in time, the event has not yet occurred. Statisticians say that such observations are *censored.*

The prospect of censoring complicates research design; the presence of censoring complicates statistical analysis. Many researchers have responded to these complications with ad hoc strategies, none entirely satisfactory: categorizing the outcome and placing the censored observations in a single group (Condiotte and Lichtenstein 1981), restricting attention to noncensored cases (Lelliott et al. 1989), deleting censored cases (Litman et al. 1979), or using the censored outcome as a categorical predictor of another outcome that varies over time (Coelho 1984). Others sidestep the "when" question entirely and ask only the "whether" question: "Does the event occur by a particular point in time (Grey et al. 1986) or by each of several successive points in time?" (Glasgow et al. 1988).

Although researchers in the drug abuse field were among the first to recognize the severe limitations of these strategies—most notably the sensitivity to the length of data collection (Hunt et al. 1971; Nathan and Lansky 1978; Sutton 1979)—until recently, relatively few analytic alternatives were available. However, new developments in statistical theory, accompanied by new developments in statistical computing, have changed how researchers can study time. The new methods—known as survival analysis, event history analysis, or hazard-modeling—were developed by biostatisticians modeling human lifetimes (Cox 1972; Kaplan and Meier 1958) and have been extended by economists and sociologists studying social transitions (Heckman and Singer 1985; Lancaster 1990; Tuma and Hannan 1984). Differences in labels aside, these techniques use similar mathematical roots to reach similar goals: to help researchers simultaneously explore *whether* events occur (do people start using illicit drugs, stop smoking, begin drinking again?) and, if so, *when*. Using specific techniques within the broad class of methods, researchers can describe patterns of occurrence, compare these patterns among groups, and build statistical models of the risk of occurrence over time.

Owing to its genesis in modeling human lifetimes, where the target event is death, survival analysis is shrouded in dark, foreboding terms. However, beyond the terminology lies a powerful methodology that appropriately uses data from all observations, noncensored and censored cases alike. Data collection can be prospective or retrospective, experimental or observational. Time can be measured continuously or discretely. The only requirements are: (1) that, at every timepoint of interest, each individual be classified into one of two or more mutually exclusive and exhaustive states, and (2) that the researchers know, for at least some of the individuals, when the transition from one state to the next occurs.

In this chapter, a nonmathematical introduction to survival analysis for drug abuse prevention researchers is presented; readers seeking a more technical presentation should consult one of the references cited at the end of the chapter (Singer and Willett 1993; Willett and Singer 1993). After developing the basic concepts, the authors focus on two topics—study design and data analysis—and, for each, identify the key issues researchers face and provide guidelines for making informed decisions about them. In the process, the authors review how prevention researchers have used the methods to date and point towards new directions for their application. The presentation is based on the authors' experience with the methods (Singer and Willett 1991, 1993; Willett and Singer 1991, 1993) and examples drawn from the recent literature.

THE CONCEPTS UNDERLYING SURVIVAL ANALYSIS

The concepts underlying survival analysis differ markedly from the familiar means, standard deviations, and correlations of traditional parametric statistics. These concepts are developed here using data reported by Stevens and Hollis (1989), who evaluated the efficacy of supplementing a smoking cessation program with followup support sessions designed to help ex-smokers cope with abstinence. The researchers randomly assigned 587 adults who successfully completed a 4-day program to one of three conditions: (1) 3 weeks of coping skills training; (2) 3 weeks of support sessions without skills training; or (3) no supplemental sessions. For 1 year after quitting, participants returned a monthly postcard noting their smoking status. Defining abstinence as smoking no more than five cigarettes per month, Stevens and Hollis asked *whether* the followup support helped people remain abstinent and, if it did not, *when* people were most likely to relapse.

Survivor Function

Survival analysis begins with the survivor function. When studying abstinence after smoking cessation, as in this example, the population survivor function indicates the probability that a randomly selected ex-smoker will remain abstinent over time. Given a representative sample from a target population, the sample survivor function estimates the population probability that a randomly selected person will remain abstinent longer than each time assessed—in this example, 1 month, 2 months, and so on—until everyone relapses or data collection ends (whichever comes first).

Panel A of figure 1 presents the sample survivor function for the 198 people in Stevens' and Hollis' control group.[2] At the beginning of the study (i.e., the beginning of "time"), the estimated survival probability was 1.0. As time passed and people relapsed, the sample survivor function dropped toward 0. In this study, 82 percent successfully abstained from smoking (i.e., "survived") more than 1 month following cessation, 66 percent abstained more than 2 months, 60 percent abstained more than 3 months, and so forth. By 12 months, when data collection ended, 38 percent remained abstinent. These individuals had censored relapse times, either because they never relapsed or because, if they did, it was after data collection ended. Because of censoring, sample survivor functions rarely reached 0.

The sample survivor function helps researchers answer the descriptive question, "On average, how many months pass before the abstinent smoker relapses?" When the sample survivor function reaches 0.5, half of the ex-smokers have relapsed, half have not. The estimated *median lifetime* identifies this midpoint, which indicates how much time passes before half of the sample experiences the target event. As shown in figure 1, among ex-smokers without followup support, the answer is 4 months. The median lifetime statistic incorporates data from both the 123 uncensored individuals who relapsed within 12 months of data collection and the 75 censored individuals who did not.

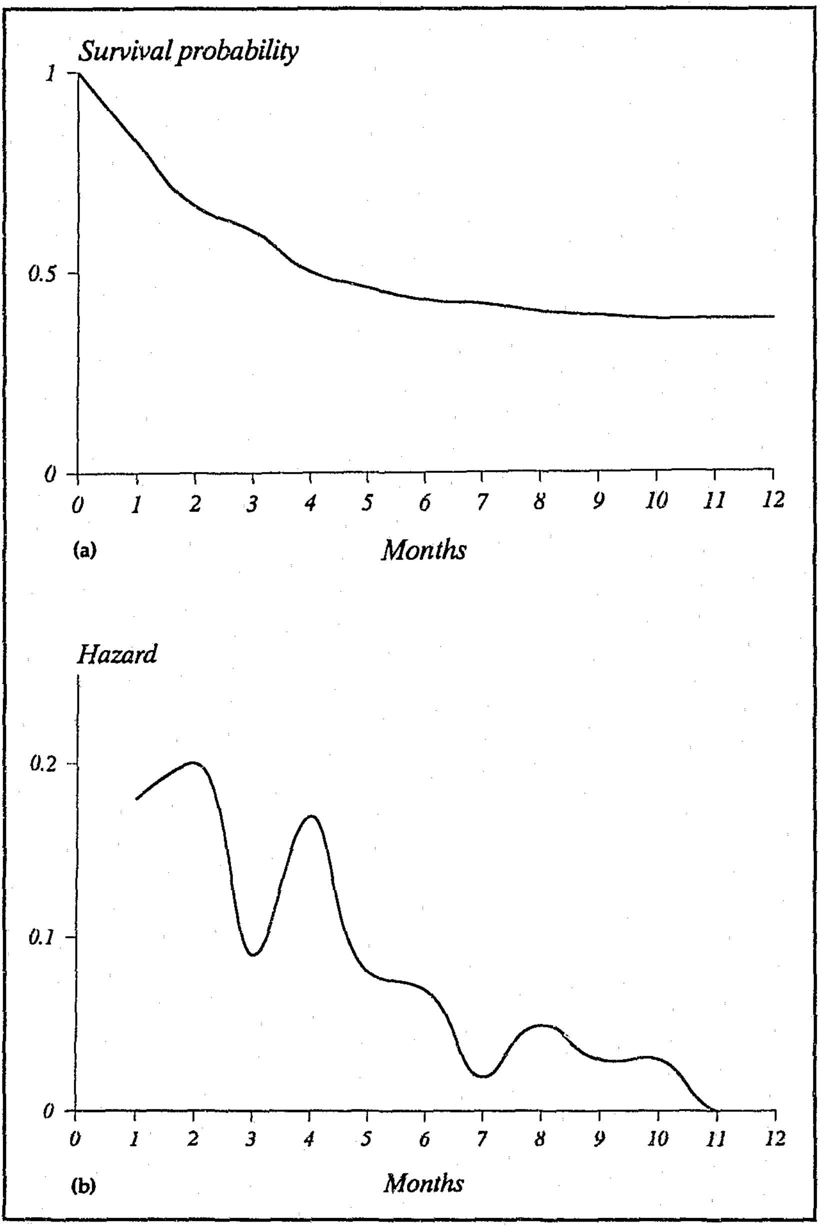


FIGURE 1. *Sample survivor (panel A) and hazard (panel B) functions for 198 ex-smokers based on data reported by Stevens and Hollis (1989)*

All survivor functions have a shape similar to that displayed in figure 1—a negatively accelerating extinction curve, a monotonically nonincreasing function of time. Well before the advent of modern survival methods, Hunt and Bespalec (1974*a*, 1974*b*), Hunt and General (1973), Hunt and Matarazzo (1970), and Hunt and colleagues (1971) noted this generalization. After finding similarly shaped survivor functions in nearly 100 studies of smoking, heroin, and alcohol cessation, Hunt and colleagues (1971) presaged the utility of another plot (to which the authors now turn) when they wrote that they "hoped to use the differences in slope between individual curves as a differential criterion to evaluate various treatment techniques" (p. 455).

Hazard Function

If a large proportion of successful abstainers suddenly relapses in a given month, the survivor function drops sharply, as happens in figure 1, during each of the first few months after smoking cessation. When this happens, ex-smokers are at greater risk of relapse. Examining the changing slope of the survivor function is one way to identify such "risky" time periods. A more sensitive way to assess the risk of event occurrence is to examine the hazard function, a mathematical function related to the survivor function that registers these changing slopes of the (negative log) survivor function.

Mathematical definitions of hazard differ depending upon whether time is measured discretely or continuously. If time is measured discretely, hazard is defined as the conditional probability that an ex-smoker will relapse in a particular time interval, given that the person has not relapsed prior to the interval. As the interval length decreases, the probability that an event will occur during any given interval decreases as well. In the limit, when time is measured continuously, the definition of hazard must be modified because the probability that an event occurs at any "infinitely thin" instant of time will approach 0 (by definition). So, continuous-time hazard is defined as the instantaneous rate of relapse, given uninterrupted abstinence until that time. While hazard always is nonnegative, when time is measured discretely, it can never exceed 1; when time is measured continuously, hazard can assume any value greater than or equal to 0.

Like the survivor function, the hazard function can be plotted versus time, yielding a profile of the risk of relapsing each month, given uninterrupted abstinence until that month. The magnitude of each month's hazard indicates the risk of relapsing in that month—the higher the hazard, the greater the risk. Each month's hazard is calculated using data on only those individuals still eligible to experience the event during the month (i.e., the *risk set*); individuals who already have relapsed are not included.

Panel B of figure 1 presents the sample hazard function corresponding to the sample survivor function in panel A. The risk of relapse is high in each of the first few months of the study and then declines over time. Ex-smokers are at greatest risk of relapse immediately after they quit; those who successfully abstain for several months are likely to abstain for at least a year.

Use of the hazard function in prevention research was proposed well before the use of modern survival methods but, because the associated statistical models were not available yet, much information in the function remained unexploited. Litman and colleagues (1979), McFall (1978), and Sutton (1979), all suggested that researchers examine relapse on a period-by-period basis—as the hazard function does—and identify who relapses, and when. These authors appropriately dismissed the survivor function as too crude a summary because of its consistent shape regardless of the distribution of risk.

The strength of the hazard function is that it effectively portrays the distribution of risk across time. To illustrate its utility, consider the three hazard functions in figure 2, which depict the risk of first use of alcohol, cigarettes, and marijuana by grade in school. These plots were constructed using data presented by Johnston (1991) on the age at first use of these three drugs among members of a high school graduating class of 1988.

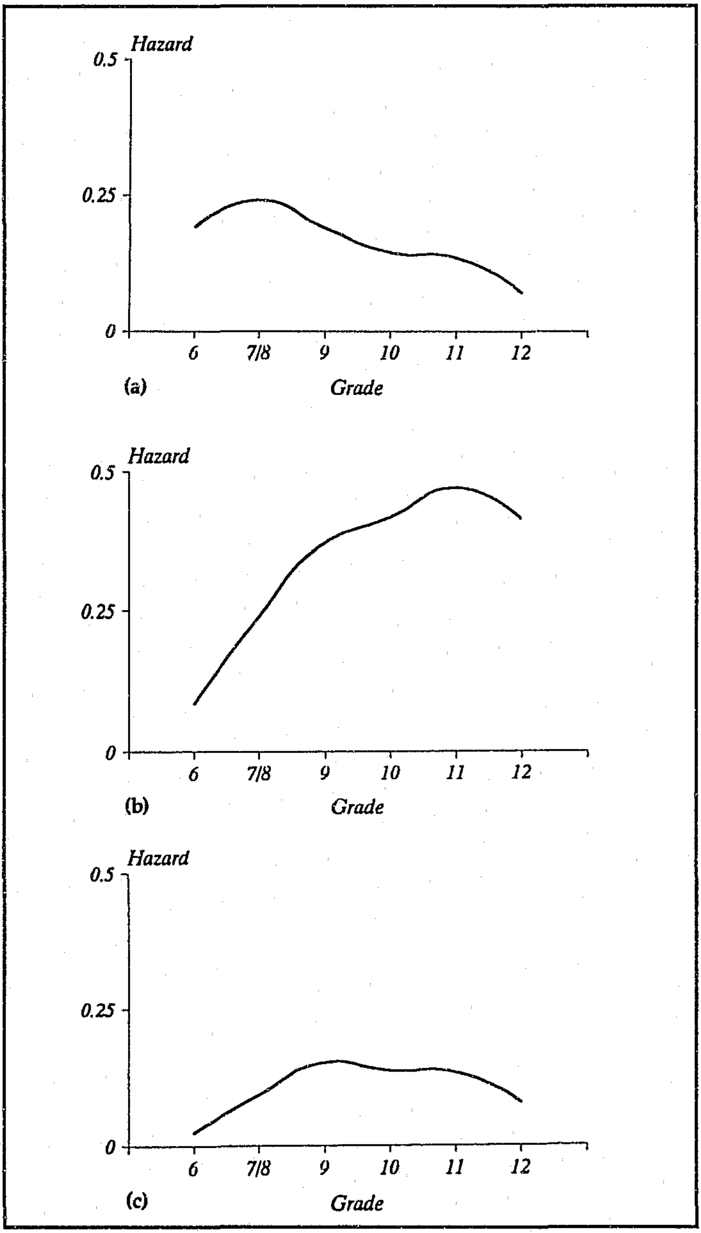


FIGURE 2. *Three hazard functions depicting the grade-by-grade risk of first use of selected drugs: (panel A) cigarettes; (panel B) alcohol; and (panel C) marijuana. This figure is based on data reported in Johnston (1991) from a high school graduating class of 1966.*

Because peaks in the hazard function indicate periods of elevated risk, they pinpoint *when* the target event (here, initiation of drug use) is most likely to occur. Begin by examining the hazard function for cigarettes (panel A). Its high elevation in the sixth and seventh/eighth grades indicates that the risk of first trying cigarettes is greatest in these middle school years. After this initial risky period, when many preadolescents experiment with tobacco, the risk of trying cigarettes, *among those who have not already done so*, declines steadily over time. Indeed, by 12th grade, the risk of initial use of cigarettes is less than 0.1.

The risk of initial use of alcohol, in contrast, increases steadily over time (panel B). Relatively few students take their first drink in sixth grade, for example, as indicated by the low level of hazard (< 0.10) in this period. Over time, however, the risk of trying alcohol increases steadily so that, by 11th grade (the period of greatest experimentation), hazard nears .5.

Now consider the hazard function for initial use of marijuana (panel C), which differs from that of these other two substances in two important ways. First, it consistently is lower, indicating that in every grade, the risk of first use of marijuana is lower than the risk of first use of cigarettes or the risk of first use of alcohol. Second, the risk of first use of marijuana peaks in the middle of the time axis--in ninth grade—not in the beginning (as for cigarettes) or the end (as for alcohol). This indicates that the time period when adolescents are most likely to experiment with drugs differs by drug type. By examining the hazard function, which illuminates such differential profiles of risk, researchers can learn *when* to target specific types of prevention interventions for different types of drugs.

Incidence and Prevalence: An Analogy for Hazard and Survival

Because hazard and survival functions may be unfamiliar concepts, the authors offer an epidemiological analogy to concepts that some readers may find more familiar—incidence and prevalence. *Incidence* measures the number of new events occurring during a time period (expressed as a proportion of the number of individuals at risk), while *prevalence* cumulates these risks to the total number of events that have occurred by a given time (also expressed as a proportion) (e.g., Kleinbaum et al. 1982; Lilienfeld and Lilienfeld 1980). Incidence and prevalence correspond directly to hazard and survival: hazard represents incidence, and survival represents cumulative prevalence.

This analogy reinforces the importance of examining both the survivor and hazard functions. Epidemiologists have long recognized that, while prevalence assesses the extent of a problem at a particular point in time, incidence is the key to disease etiology (Mausner and Bahn 1974). Why? Because prevalence confounds incidence with duration. Conditions with longer durations may be more prevalent, even if they have equal or lower incidence rates. To determine *when* people are at risk, epidemiologists study incidence. When they study incidence, they are actually studying hazard.

DESIGN: COLLECTING SURVIVAL DATA

The conduct of survival analysis requires data summarizing the behavior of a sample of individuals over time. Data can be collected prospectively (as in Stevens' and Hollis' smoking cessation study) or retrospectively (as in Johnston's grade at initial drug use study). The best studies tailor the timeframe to the target event. When studying the side effects of a nicotine patch, a 10-day or 10-week segment might suffice but, when studying the link between drug use and coronary heart disease, even a 10-year window might not.

The following sections discuss eight questions that arise when designing a study of event occurrence: Who will be studied? What is the target event? When does "time" begin? How often should data be collected? How can event histories be reconstructed from retrospective data? How can attrition be minimized? What should be done with repeated events? How long should data be collected? How many people should be studied?

Who Will Be Studied?

As with any statistical method, getting the full advantages of survival analysis requires a representative sample of individuals selected from an appropriate target population. Although data collected from convenience samples can be used, probabilistic statements, population generalizations of sample summary statistics, or statistical inferences may be rendered incorrect. Because many prevention researchers work with epidemiologists accustomed to using probabilistic sampling schemes, there are many excellent examples of survival analyses using data collected from representative samples (e.g., Kandel and Yamaguchi 1987; Rosenbaum and Kandel 1990). The authors hope this standard will persist as survival methods find their way into smaller-scale studies and in clinical settings.

A more problematic issue concerns the need to define carefully the target population from which the sample will be selected. Subtle variations in population definitions inadvertently can distort the distribution of time—the very quantity of interest. Consider the tempting strategy of eliminating censoring altogether by restricting the target population to only those individuals with known event times. A simple example from the research literature on the duration of foster-care arrangements illustrates the problems that can arise. When studying discharge times for children in foster care, Milner (1987) defined his target population as the 222 children in a State agency who were released from care between 1984 and 1985 (thus disregarding those who were not discharged). In a random sample of 75 of these children, he found that 37 percent had entered care within 5 months of discharge, 29 percent had entered care within 6-11 months of discharge, 14 percent had entered care within 12-24 months of discharge, and the remaining 20 percent had entered care over 25 months before discharge.

The estimated median time to discharge in this sample was 6-11 months. Should it be concluded that the "average" child stayed in foster care for under a year? Although this study used a probability sample from a well-defined target population, the answer to this question is not known, for the target population is unsuitable for answering it. Milner knew about discharge times only among children *already discharged*; he ignored those who remained in care. Children in foster care for long periods of time were most likely to be excluded from his study. Determining how long the *average* child stayed in care requires a random sample of *all* children in care. It is likely that Milner's sampling strategy led to an underestimate of the average duration of foster care in the full population.

Some definitions of the target population create more subtle biases. Hidden biases are common especially in retrospective studies because a population defined at a particular point in time excludes people who already experienced an event that made it impossible for them to enter the target population. If a researcher conducted a retrospective study of age at first cocaine use based on a random sample of high school seniors, for example, he or she necessarily would exclude students who had died already because of cocaine use or students who already had dropped out of school.

When a sample excludes individuals who already have experienced the event of interest before data collection begins, statisticians say that the sample is *left truncated*. Left truncation has received very little attention in the methodological literature, perhaps because the nature of the problem—the *omission* of any information—makes it difficult to evaluate the extent or impact of the truncation. As Hutchison (1988*a*, 1988*b*) notes, many methodologists ignore left truncation entirely or incorrectly fail to distinguish it from another methodological difficulty discussed below—*left censoring*. To avoid the complications arising from left truncation, the authors offer some design advice: Whenever possible, define the target population using delimiters unrelated to time and, if this is impossible, fully explore the potential biases created by whatever definition is used.

What Is the Target Event?

At every timepoint of interest, each individual under study must occupy one, and only one, of two or more states. The states must be mutually exclusive (i.e., nonoverlapping) and exhaustive (of all possible states).

Each individual is either using drugs or not, smoking or abstinent, in treatment or not. The target event occurs when an individual moves from one state to the next.

States must be defined precisely, with clear guidelines indicating the specific behaviors, responses, or scores constituting each state. The definition of states is always difficult, even when clinical definitions of event occurrence exist. When reviewing the literature on the onset, recovery, relapse, and recurrence of depression, for example, members of the MacArthur Foundation Research Network on the Psychobiology of Depression concluded that "one investigator's relapse is another's recurrence" (Frank et al. 1991, p. 851).

Fortunately for prevention intervention researchers, the specification of criteria for defining states precisely has received much attention in recent years (Brownell et al. 1986; Velicer et al. 1992). This can be seen in the recent trend toward multiple classification systems that employ biochemical assays, clinical judgment, and self-reports together. Many researchers who once relied solely on a clinical criterion, such as total abstinence, for example, now augment this definition with a less rigid one that permits temporary lapses (Baer and Lichtenstein 1988). Similarly, many researchers who once relied solely on self-report now augment their definition with biochemical data.

Regardless of the source of data, researchers must strike a balance between restrictive definitions, which lead to underestimates of the time to relapse, and less rigorous definitions, which bias estimates towards late relapse. Brownell and colleagues (1986), for example, argue that prevention researchers routinely consider at least two definitions when studying recurrence—lapse (a temporary slip that may or may not lead to relapse) and relapse. Velicer and colleagues (1992) provide a helpful review of the issues arising in the definition of outcome in smoking cessation studies.

Why do methodologists dwell on these definitional issues? They do so because of their serious methodological ramifications. It is clear, for example, that some of the observed variation in relapse rates reported in the literature is attributable not to the differential effectiveness of various interventions but to variation in the definition of event states. Consider, for example, the different conclusions that a research reviewer could cull from just the first month of data on unaided smoking cessation collected by Marlatt and colleagues (1988). By the end of the month, 23 percent of the sample never actually had quit (they smoked again within 24 hours), 36 percent had quit for at least 24 hours but subsequently relapsed within the month, 16 percent had been primarily abstinent but smoked one or two cigarettes, and only 25 percent had been successfully abstinent. In no time at all, a research reviewer could reasonably calculate at least three different relapse rates: by setting aside individuals who never really quit, by pooling the primarily abstinent individuals with the relapsers, or by pooling them with the successfully abstinent individuals.

Given the important role of substantive issues in the definition of event states, all measurement considerations necessary for deriving reliable and valid definitions of event states cannot be reviewed here. Instead, the authors offer more modest general advice: Collect data with as much precision as possible so that transitions can be coded appropriately from one state to the next. With refined data, individuals always can collapse together to derive broader definitions; with coarse categorized data, it is difficult (and often impossible) to recoup more differentiated definitions. When describing results, operationalize definitions as precisely as possible (specifying the criteria for onset, recovery, relapse, and recurrence as clearly as possible in terms of the number, intensity, and duration of symptoms) so that others can compare their findings.

When Does "Time" Begin?

The problem of "starting the clock" is more complex than it may appear. When studying the onset of addictive behaviors, birth certainly is the logical choice. In their community survey of substance abuse among adolescents and young adults, for example, Kandel and Logan (1984) used chronological age (i.e., time since birth) to examine when respondents reported first using marijuana, alcohol, cocaine, and psychedelic drugs.

However, chronological age is not the optimal metric for all research questions arising in drug abuse prevention. Many are better addressed by starting the clock after a precipitating event occurs. Coryell and colleagues (1990), for example, started the clock when patients first presented to a therapeutic setting, Cooney and colleagues (1991) used the date of discharge from an inpatient setting and others (e.g., Brownell et al. 1986; Havassy et al. 1991) have used the date when individuals stopped using a particular drug. Such alternative starting times should be considered whenever an individual is at risk of the target event (e.g., remission, relapse, or recurrence) only after experiencing the prior event.

Consideration of the process under study usually leads to a defensible decision. When it does not, an arbitrary time can be used. Researchers conducting randomized clinical trials, for example, typically use the date of randomization (Greenhouse et al. 1991; Peto et al. 1976) or the date of intervention (Greenhouse et al. 1989). Beware of the measurement imprecision created when the chosen precipitating event only approximates the conceptual beginning of time. When modeling illnesses, for example, the conceptual beginning of time is the onset of the illness episode, yet medical researchers often use the date of evaluation or diagnosis. Since the time between onset and entry into treatment can vary greatly across individuals (Monroe et al. 1991) and the magnitude of this lag time may be an important predictor of a treatment's efficacy, use of these more easily measured dates actually may add even more error into the definition of event occurrence.

What happens if the start date is unknown for some individuals under study? Statisticians say that such observations are *left censored* (to distinguish them from *right-censored* observations in which the *event* times are unknown). Statistical methods for including left-censored data in analyses that also have right-censored data remain in their infancy. Although Turnbull (1974, 1976) offered some basic descriptive approaches and Cox and Oakes (1984) and Flinn and Heckman (1982) offered some guidelines for developing statistical models (under a very restrictive set of assumptions), most methodologists dismiss the topic soon after introducing the terminology (e.g., Blossfeld et al. 1989, p. 29; Tuma and Hannan, 1984, p. 135). The most common advice is that researchers should define the beginning of time so that left censoring never arises, or they should set the left-censored spells aside from analysis (Allison 1984; Tuma and Hannan 1984).

How Often Should Data Be Collected?

Few researchers have the luxury of monitoring subjects continuously. Financial and logistical constraints usually demand that researchers contact subjects at a finite number of preselected intervals. Using these "chunky" data, researchers then try to retrospectively reconstruct pseudo-continuous event histories. Reconstruction can be made more effective if researchers judiciously select the preselected intervals at which study subjects will be contacted.

The collection of data in discrete time can add measurement imprecision. If transitions occur in continuous time but data are collected in discrete time, for example, a researcher will never know an individual's mental state at the moment of transition. Such imprecision has serious consequences if information about the transition moment is critical for predicting the timing of events, as when the coping skills of the ex-smoker, ex-gambler, ex-drinker, ex-overeater, or ex-drug abuser may determine whether the person succumbs to temptation. Shiffman (1982) used an innovative design to overcome this restriction; he interviewed 183 ex-smokers who called a smoking cessation hotline *because* they were in crisis. His design may be useful in other studies requiring data collected at the precise moment of transition.

Carefully constructed interview questions can improve the quality of the event history data. Bradburn and colleagues (1987) provide strategies for helping respondents construct temporal autobiographies. They recom-' mend letting respondents create their own timelines based on personally salient anchors (e.g., birthdays, anniversaries, or holidays) and then sequentially placing other events (and symptoms) on this timeline (see Young et al. 1991 for an application). In multiwave studies, bounded-recall probes can enrich the quality of data describing behavior between interviews. At the beginning of the second and subsequent interviews, for example, Neter and Waksberg (1964) suggest that interviewers first remind respondents of their responses during the previous interview.

Where should limited data collection resources be targeted? Although collection at equally spaced time intervals is systematic, this strategy may omit information about the periods of greatest interest. A simple but effective strategy, which maximizes information on the occurrence of the target event, is to collect data more frequently when events are the most likely to occur.

Information on the anticipated shape of the hazard function is helpful in selecting times for data collection. The idea is to collect data more frequently when hazard is high and less frequently when hazard is low. This allocation strategy was used effectively, for example, by Hall and colleagues (1984) who, in their 1-year prospective study of smoking abstinence following behavioral skills-training, placed their four data collection periods at 3, 6, 26, and 52 weeks after treatment. If they had spaced data-collection episodes equally, waiting until week 13 to first collect followup data, they would have been unable to determine that the risk of relapse was highest in the few weeks immediately following cessation.

How Can Event Histories Be Reconstructed From Retrospective Data Collection?

In 1837, William Farr wrote, "Is your study to be retrospective or prospective? If the former, the replies will be general, vague, and I fear of little value" (cited in Lilienfeld and Lilienfeld 1980). His words remain true today. Whenever possible, researchers should collect data prospectively. However, when studying infrequent events—initiation into opiate drug use, for example—prospective data collection may be unfeasible.

Many researchers, therefore, opt for a different approach: interviewing people and asking, "Has the event *ever* occurred?" and, if so, "*When* did it first occur?" Retrospective data collection has been used successfully by researchers studying the age at first use of many different addictive substances and remains a fruitful strategy for drug abuse research (e.g., Adler and Kandel 1983).

Researchers contemplating a retrospective data-collection effort should be forewarned, however, that their data will be imperfect. Although rare events—suicide attempts or hospitalization—may be remembered indefinitely and highly salient events—initial use of drugs or first symptoms of an illness—may be remembered for 2 or 3 years, habitual events like ongoing symptoms and substance use are too embedded in an individual's life to be remembered precisely (Bradburn 1983; Sudman and Bradburn 1982). The longer the time period, the greater the error. (As noted earlier, if the target event can lead to death, the collection of retrospective data from a cohort ensures that sampling will be biased by the omission of those who already have succumbed.)

Three errors are common in retrospective data collection: (a) *memory failures*—respondents forget events entirely; (b) *telescoping*—events are remembered as having occurred more recently than they actually did; and (c) *rounding*—respondents drop fractions and report even numbers or numbers ending in 0 or 5. These errors create different biases: memory failures lead to underreporting, telescoping to overreporting, and rounding to both.

Supplemental aids and records can help reduce errors. Records control overreporting due to telescoping but have no effect on omission; aided recall, where the subject is presented explicitly with the possible options and is asked directly whether any particular event happened, reduces the number of omissions but may increase telescoping (Sudman and Bradburn 1974). Researchers developing items for retrospective recall would do well to consult strategies described in the ongoing series *Cognition and Survey Measurement* published by the National Center for Health

Statistics (e.g., Lessler et al. 1989; Means et al. 1989) and in the recent book *Measurement Errors in Surveys* (Biemer et al. 1991).

If retrospective recall is the only alternative, is it worth the effort? The authors believe it is. In their retrospective study of suicide ideation, Bolger and colleagues (1989) successfully used several approaches to improve recall (see also Wittchen et al. 1989). Although studying a "threatening" event, they couched the study in less threatening terms about the development of the concept of death and suicide. They never asked about respondents' mental health or suicidal behavior—only about thoughts and knowledge about others. Questionnaires were anonymous and self-administered in a group setting. Respondents were college students—close enough in age to the time period of interest (adolescence) but old enough to be removed.

How Can Attrition Be Minimized?

Given the expense and difficulty of prospective data collection, researchers want to keep every case they can. It is well known that, as sample size decreases, statistical power decreases and, if attrition is nonrandom, generalizability may suffer as well. As Hansen and colleagues (1985) clearly show, drug abuse prevention studies have been plagued by attrition problems. Indeed, in their recent review of the attrition problem, Biglan and colleagues (1991) noted several studies with attrition rates in excess of 50 percent!

Researchers most successful at minimizing attrition have used some of the following strategies: explain to respondents why they have to be followed; ask them to contact a study representative if they move; visit their homes and ask neighbors for information about them; pay them for participation in each interview; have them pay an earnest deposit refundable at the end of the last interview; offer lottery prizes for those who successfully complete all required interviews, mail a newsletter at regular intervals, record the names and addresses of several relatives or friends not living with them, record each respondent's social security number, convene reunion meetings, maintain contact at regular intervals even if data are not being recorded as frequently, send birthday and seasonal greeting cards, and consult official records (e.g., jail, hospital, welfare, or driver registration). Crider and colleagues (1971, 1973), Farrington and colleagues (1990), and Murphy (1990) offer many helpful strategies for minimizing attrition.

Despite diligent effort, most researchers lose some individuals to followup. Researchers attempting to improve their study by using a long followup period face a further conundrum: the longer the followup, the greater the attrition. At first sight, attrition seems nonproblematic for survival analysis because it leads to additional right-censored event times—a problem that survival analysis was designed to handle. However, censoring due to attrition may not be the "noninformative" censoring for which survival methods are valid. Individuals lost to followup can differ substantially from individuals who continue to participate. In their longitudinal study of drug abuse, for example, Biglan and colleagues (1991) present clear evidence that those who remain in the sample differed from those who did not.

What should a researcher do with the data on individuals lost to followup? While multiple imputation methods offer much promise (Little and Rubin 1987), three simple strategies sometimes can suffice. One is to assign each case a censored event time equal to the length of time the person was observed (without the event occurring). If an individual participated for the first 6 months of a 12-month study before attriting, censor the event time at 6 months. A second approach is to use a "worst-case" scenario—assume that the event actually occurred when the case was lost to followup. Under this strategy, the event time is not censored. The findings from analyses carried out under both types of recoding then can be contrasted with each other in a sensitivity analysis. Persistence of findings obtained under multiple strategies or explainable differences between the findings reinforces the strength of the analytic results. The third approach is to conduct a "competing-risks" survival analysis, in which study attrition is treated as another event that "competes" to end an individual's lifetime (Singer and Willett 1991).

The appropriateness of these alternative strategies depends, in part, upon the target behavior under study. Be especially careful when assuming that the event o‹ curred at the time when the observation is censored, for this converts a nonevent into an event. Of course, when studying relapse, this conclusion may be sound because former drug abusers notoriously are unfaithful subjects, and those who are "clean" are more likely to stay in touch. The key idea is to let reason be the guide. Within 12 weeks after beginning a study of 221 treated alcoholics, opiate users, and cigarette smokers, for example, Hall and colleagues (1990) lost 73 people (one-third of their sample) to followup despite valiant attempts to minimize attrition. To ascertain the impact of attrition on their findings, the researchers conducted extensive sensitivity analyses, including: (1) coding of relapse as occurring the week after the last interview completed, and (2) setting aside these cases from analysis. All the analytic findings were similar in sign and magnitude, although the standard errors of parameter estimates were higher under the second strategy because of a loss of statistical power.

What Should Be Done With Repeated Events?

Many events marking the "careers" of drug abusers are repeatable. Indeed, with the exception of initiation into drug use, most other events—ongoing use, abuse, hospitalization, treatment, and relapse—can occur over and over again. When studying the timing of potentially repeatable events, make every attempt to note the "spell number" under study, for the natural course of a first spell may differ from the natural course of second and subsequent spells. So, too, the efficacy of treatment may vary depending upon how many prior spells the individual has experienced.

Drug abuse prevention researchers can learn much about this issue by examining the literature on depression. For example, Kupfer and colleagues (1989) designed a study to investigate differential recovery patterns across multiple spells when studying patients with recurrent depression. Separately analyzing the time to stabilization in two consecutive episodes, they found virtually identical median lifetimes (between 11 and 12 weeks). They also found, however, that the efficacy of treatment varied across spells—early intervention in the second episode, as opposed to the first episode, was what worked particularly well.

The authors believe that the unidentified presence of multiple spells in a single data set may help explain some of the major puzzles in prevention research. This belief stems from a parallel finding in the literature on depression, where Klerman (1978) demonstrated that some of the observed variation in relapse rates was attributable to researchers' failure to note how many prior episodes of depression each subject had had. Given the tendency toward renewed abstinence on the part of formerly abstinent people who relapsed early after quitting, it seems reasonable to hypothesize that previous treatment, even if unsuccessful, may increase the probability of success under subsequent treatments.

How Long Should Data Be Collected?

Once the clock starts, it must stop eventually. Clocks in retrospective studies stop on the date of interview; clocks in prospective studies can, in theory at least, continue indefinitely. As a practical matter, though, most prospective studies follow a sample for a finite, preselected period of time. The length of data collection determines the amount of right censoring (hereafter referred to as "censoring"). Because longer data collection periods yield fewer censored observations, the simple maxim is "the longer, the better." But beware—longer studies are more expensive, have more missing data, and may lead to out-of-date results.

When deciding on the length of followup, remember that, to determine *when* the event is likely to occur, it actually must occur for enough people under study. If the target event never occurs during data collection, all observations are censored. The researcher has little information, knowing only that it generally takes longer than this period for the event to occur.

There is no universally appropriate length of followup. The answer depends on many factors. To decide on a reasonable followup period, the shape of the anticipated hazard function, the probable median lifetime, the sample size, and proposed statistical analyses must be considered. As shown in the section on determining sample size, a good rule of thumb is to follow participants long enough for at least half of them to experience the target event during data collection. This ensures sufficient information for estimating a median lifetime and provides reasonable statistical power.

What have researchers done in practice? Noting that ex-smokers often start smoking again soon after quitting, McFall (1978) suggested that smoking-relapse studies use a 6- to 12-month followup. In a review of smoking-relapse studies published during the 1980s, Singer and Willett (1991) found that this guideline is accepted widely; the modal followup period was 1 year, and this period yielded an average censoring rate below 50 percent. However, Nathan and Skinstad (1987) note that "3- or 6-month posttreatment follow-ups are likely to be insufficient. . . . 2 years or more are probably necessary to determine the long-term effects of a treatment program" (p. 333). Furthermore, when studying infrequent events, even 5 years of data collection may be insufficient. In their review of the link between alcoholism and suicide, for example, Murphy and Wetzel (1990) lament the fact that many of the available studies "are relatively short: less than 10 years" (p. 387).

Before deciding on the length of data collection, be sure to consider the substantive ramifications of this choice. It is clear that variation across studies in the length of followup explains some of the seemingly discrepant conclusions about treatment efficacy that arise in the literature. Length of followup has been identified as a major explanatory factor in several literature reviews, including Murphy and Wetzel's (1990) review of suicidality among alcoholics. Even when it has not been identified as a key explanatory factor, its impact seems certain. In their review of 26 longitudinal studies of teenage alcohol and other drug use, for example, Flay and Petraitis (1991) found that the length of followup varied from a low of 5 months to a high of 19 years. Although the link between length of followup and study findings was not investigated, this design feature may explain why some studies successfully predicted subsequent outcomes while others did not.

Because of the effect of design on conclusions, a researcher always must note the length of followup. Any relapse rate cited must be linked to a specific time period. What can be concluded, for example, from the statement of Seltzer and colleagues (1982) that 65 percent of the mentally retarded adults in their study were not reinstitutionalized, given that the timeframe being referenced is not known? How can researchers know whether this percentage is low or high? How can this rate be compared to those found in other studies? Even well-documented longitudinal studies using sophisticated analytic techniques occasionally omit this important piece of information (Zatz 1985). The length of data collection is key to understanding the ultimate course of survival.

How Many People Should Be Studied?

Having specified in broad outline the design of a study, the final step is to determine how many people to include in the sample. Statisticians determine the minimum number of people a researcher should study by conducting a statistical power analysis (Cohen 1990; Kraemer and Thiemann 1988). This requires specification of the particular hypothesis to be tested, the desired Type I and Type II error rates, and the minimum effect size considered important; for survival analysis, it also requires presaging the anticipated distribution of the hazard function and the proposed length of followup.

Biostatisticians have derived many methods for determining sample size for survival analysis, each applicable under different circumstances. Donner (1984) and Lachin (1981) review the literature; Freedman (1982) provides tables for two-group comparisons; Makuch and Simon (1982) provide formulae for multiple-group comparisons; Schoenfeld and Richter (1982) provide monograms for the same purpose; Bernstein and Lagakos (1978) and Dupont and Plummer (1990) describe computer programs that perform these and other calculations for several designs; and Lachin and Foulkes (1986), Moussa (1988), and Rubinstein and colleagues (1981) provide formulae for complex designs with stratification, covariate information, or allowances for loss of individuals to followup. In the presentation that follows, the authors have computed minimum sample sizes using the computer program developed by Dupont and Plummer (1990).

No single table or formula can cover all possible design configurations. Here, ballpark estimates of sample size are provided that are similar to those provided elsewhere for more familiar statistical analyses (Light et al. 1990). This discussion does not replace consultation with a statistician *before* data collection or, in Kraemer and Pruyn's (1990) words:

> Answers to questions as to what the optimal approach depend on the specific research question to be addressed and can and do not have simple answers. How to demonstrate adequate power and how to assess power when there are multiple outcomes are questions that must be addressed, perhaps differently, in each research study, and these questions require the participation of experts at addressing such issues (p. 1169).

Rather, this discussion should provide researchers with a better sense of the factors affecting the power of survival analyses, a general sense of how many people they must study to ensure a reasonable chance of detecting an effect that really exists, and a language for talking with a statistical consultant. The need for improved design is clear. As Kazdin and Bass (1989) note, too many studies of differences between alternative treatments lack sufficient statistical power to detect the small-to-medium effect sizes likely to occur in practice.

Table 1 presents the minimum total sample sizes necessary to achieve a power of .80 for a simple two-group comparison at the .05 level (two-tailed). The rows of the table indicate minimum detectable effect sizes (R); the columns indicate the length of followup (F); the cell entries indicate the minimum total sample size used in the analysis (N).

TABLE 1. *Minimum total number of individuals needed to detect differences in survival between two groups*

Effect size	Followup period				
	0.5	1.0	1.5	2.0	2.5
1.25	> 2,162	1,260	976	840	766
1.5	654	382	296	254	232
1.75	344	200	156	134	122
2.00	224	130	102	88	80

NOTE: Assuming a two-tailed test at the .05 level, at a power of .80, and exponentially distributed survival times, all individuals followed for the same period of time.

Researchers should inflate these sample size estimates appropriately to adjust for cases lost to followup. The calculations were made assuming a flat hazard function—a restrictive assumption indeed, but the simplest, and the one researchers generally assume in the absence of more detailed information.

To use the table, first specify the smallest effect size deemed important for detection. Although biostatisticians have developed several measures of effect size, perhaps the simplest is the ratio of median lifetimes in the two groups, denoted by R. Letting m_1 be the median lifetime in one group and m_2 the median lifetime in the other, $R = m_1/m_2$. When R = 1.25, the median lifetime of one group is 25 percent longer than the median lifetime of the other; when R = 1.50, the median lifetime of one group is 50 percent longer; and when R = 2.00, the median lifetime of one group is twice as long (100 percent) as the other group.

How can the minimum detectable effect size be specified in advance of data collection? One way is to use prior research. Consider a two-group experiment that might follow from the smoking-relapse study conducted by Stevens and Hollis (1989). The median survival time in the control group of this experiment was 4 months ($m_2 = 4$). If the median survival time in a new experimental group is expected to be as high as 8 months ($m_1 = 8$), the new study can be designed to detect an R of 2.00; if the median survival time in the new experimental group is expected to be only 6 months ($m_1 = 6$), the study should be designed to detect an R of 1.50. In the absence of such prior information, Schoenfeld and Richter (1982) suggest that R = 1.50 be used because a 50-percent increase in survival is "clinically important and biologically feasible" (p. 163).

After specifying the minimum detectable effect size, the length of followup must be specified. Because the length of followup can vary a lot across studies, a standardized measure is needed that is applicable to a variety of settings and metrics. This goal is achieved by dividing the length of followup by the average anticipated median lifetime in the two groups. More precisely, letting $A = (m_1+m_2)/2$ be the average median lifetime in the two groups, and T the total length of followup, the standardized measure of followup, F, is T/A. If a study follows individuals to only half the average median lifetime, F = 0.5; if a study follows individuals to the average median lifetime, F = 1.0; and if a study follows individuals for twice as long as the average median lifetime, F = 2.0.

By using a standardized measure of the length of followup, the table can be used with studies of widely varying length. It is equally applicable if the average median lifetime is 6 minutes, 6 days, 6 months, or 6 years. If the average median lifetime (A) is 6 (in any of these units), a followup (T) of 3 yields an F of 0.5, a followup of 6 yields an F of 1.0, a followup of 9 yields an F of 1.5, and a followup of 12 yields an F of 2.0. The particular time units cancel each other out in the standardization.

Now examine the minimum sample sizes presented in table 1, focusing first on differences in effect size displayed across the rows. Small effects (R = 1.25) are difficult to detect. Regardless of the length of followup, a study must include many hundreds or well over 1,000 individuals to have a reasonable chance of detecting such effects. Medium-sized effects (R = 1.50 to R = 1.75) can be detected with moderate-sized samples; approximately 200-400 individuals generally will suffice, depending upon the length of followup. Large effects (R = 2.00) are relatively easy to detect, even using small samples. If the median lifetime in one group is twice as long as the median lifetime in the other, there is an 80-percent chance of detecting this difference using only 100-200 individuals.

Table 1 also can be used for another purpose: to decide on the length of data collection. Reexamine the table, focusing now on the variation in sample sizes across the columns corresponding to followups of widely differing lengths. The great variation in minimum sample sizes for a given effect size emphasizes the importance of following individuals under study for as long as possible.

Consider, for example, how the minimum sample size needed to detect an R of 1.50 depends upon the length of followup. If a sample is followed only halfway to the average median lifetime, F = 0.5, 654 people are required to detect the 50-percent difference in median lifetimes. However, if people are followed for longer periods of time, fewer people are needed. If the followup can be extended to the average median lifetime (F = 1.0), the same power of .80 can be achieved with almost half as many individuals (N = 382). If the followup is extended further to twice the average median lifetime (F = 2.00), the same power can be achieved with only one-third as many individuals (N = 254).

The message for research design is clear. Much statistical power can be gained by following people for longer periods of time. Researchers would do well to follow people for at least as long as the average median lifetime (F = 1.00). By doubling the length of followup, the same statistical power can be achieved with approximately one-third fewer individuals. If the length of followup is less than the average median lifetime, only studies of many hundreds of individuals will have adequate statistical power.

ANALYSIS: EXAMINING SURVIVAL DATA

Most researchers begin their data analyses with exploratory and descriptive approaches; they move on to fitting statistical models and testing hypotheses only after a full exploration of the data (Ehrenberg 1982; Mosteller and Tukey 1977). In the following sections, the authors present an array of strategies for analyzing survival data, beginning with simple descriptive approaches and moving on to statistical model building.

How Can Survival Data Be Described?

There is much to be learned by straightforward "eyeball" analysis. Inspection of sample survivor and hazard profiles and comparison of these profiles computed separately for substantively interesting subsamples can be very informative. Figure 3 illustrates this using hypothetical data on time to exit from a residential treatment facility. The figure presents the sample survivor and hazard functions describing time to exit for two groups of patients—those who had "severe" addiction problems and those who had "mild" addiction problems.

These sample survivor and hazard profiles contain a great deal of information. Examining the sample survivor profiles by severity shows that those with mild problems have better long-term cumulative prospects for release than do those with severe problems. About half of those with mild problems left the facility 2-3 months after admission; those with severe problems wait a month longer on average.

The subsample hazard profiles disentangle these exit patterns month by month and provide a more sensitive magnifying glass for identifying when patients are likely to be released. Immediately after entry into the facility, the risk of leaving rises as patients improve. After a few months, however, the risk of leaving declines. In every month, the hazard for those with mild problems is higher than the hazard for those with severe problems, indicating that the former group is more likely to be discharged at all times.

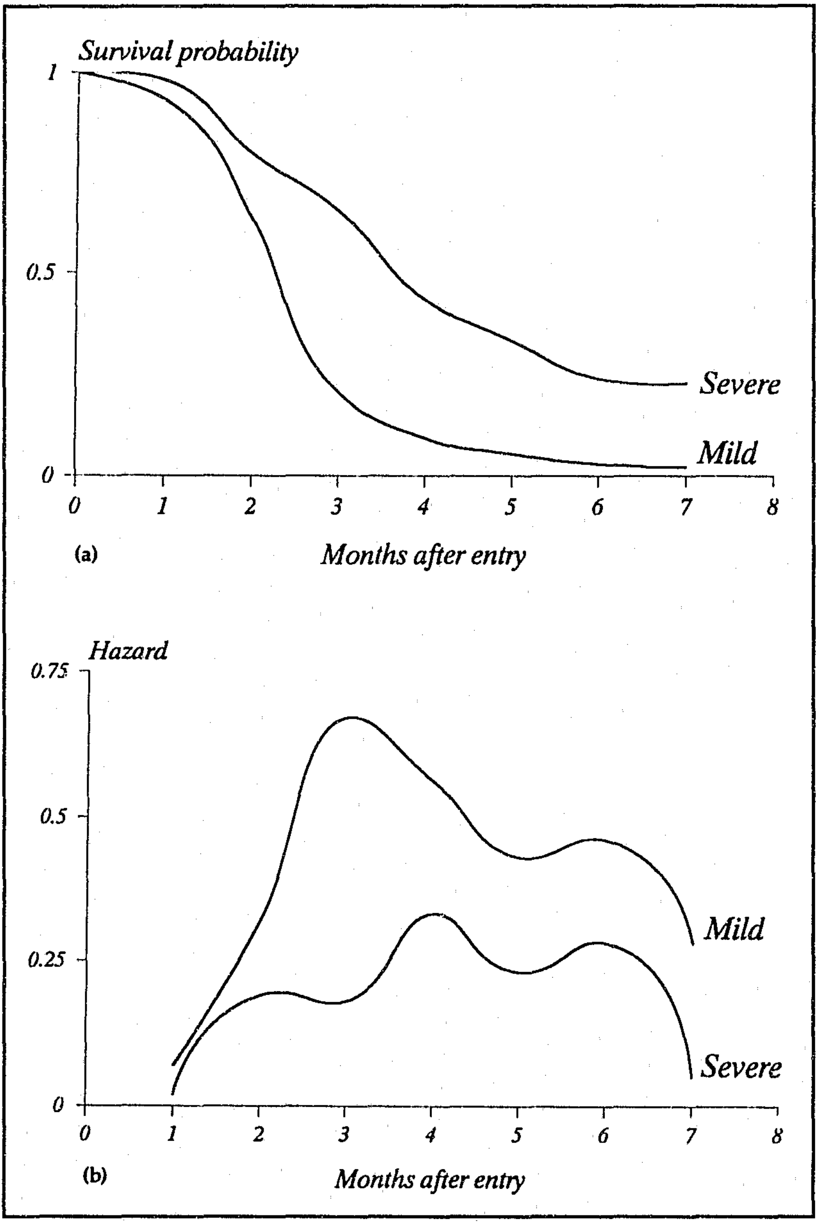


FIGURE 3. *Sample survivor (panel A) and hazard (panel B) functions for time to exit from a residential treatment facility, for those with "mild" and "severe" problems*

When hazard profiles for the two groups of people are compared, level of severity implicitly is treated as a predictor of the entire hazard profile. The comparison of profiles illustrates how the risk of leaving is related to severity. The sample could be divided in other ways, and these divisions could be treated as predictors of hazard as well.

Exploratory comparisons of sample survivor and hazard profiles provide simple persuasive descriptions of when events occur and how the timing of event occurrence varies across groups. Descriptive statements then can be buttressed by simple statistical tests of between-group differences. Lawless (1982) and Lee (1980) provide a compendium of tests for comparing survivor and hazard profiles among groups, tests that are the survival-analytic equivalent of the *t*-test and one-way analysis of variance (ANOVA). The most popular are the Wilcoxon and Log-rank tests of homogeneity of survivor function across populations—the former test placing more weight on early survival times, the latter on later survival times when the test statistic is computed.

Graphical displays and multigroup comparisons are limited, however, because they do not help researchers address the complex questions arising in prevention research. The examination of the effects of continuous predictors on hazard would yield a cumbersome collection of profiles, one per predictor value. Simple bivariate methods are ill suited for exploring the effects of several predictors simultaneously or for evaluating the influence of interactions among predictors. In their study of the relationship between adolescents' length of stay in a psychiatric hospital and two categorical predictors—diagnostic category (i.e., affective, organic, or conduct) and number of prescribed medications (i.e., none, one, two, or more)—Borchardt and Garfinkel (1991) encountered these problems. While these authors elegantly display survival profiles for each of these two predictors separately, they do not examine the joint effect of both variables simultaneously or the effects of each after controlling statistically for the other. They do not investigate the possibility of a two-way interaction between the predictors. Nor do they extend their survival analyses to explore the effects of other predictors, such as funding sources, even though their preliminary exploration suggested that such additional variables were associated with length of stay. To conduct further analysis, researchers require a comprehensive approach to the modeling of event occurrence, a topic discussed next.

How Can Statistical Models of Hazard Be Built?

Statistical models of hazard express hypothesized population relationships between entire hazard profiles and one or more predictors. To clarify the author's representation of these models, examine the two sample hazard profiles in panel B of figure 3 and think of the level of severity as a dummy variable, MILD, which can take on two values (0 for severe, 1 for mild). From this perspective, the entire hazard function is the conceptual outcome, and MILD is a potential predictor of that outcome.

Ignoring minor differences in shape, now consider how the predictor seems to affect the outcome. When MILD = 1, the sample hazard function is higher relative to its location when MILD = 0. So conceptually, the predictor MILD somehow displaces or shifts one sample hazard profile vertically, relative to the other. A population hazard model formalizes this conceptualization by associating this vertical displacement with variation in predictors in much the same way as an ordinary linear regression model associates differences in mean levels of a continuous (noncensored) outcome with variation in predictors.

The difference between a hazard model and a linear regression model, of course, is that the entire hazard profile is no ordinary outcome. The continuous-time hazard profile is a profile of risks bounded by 0. Methodologists postulating a statistical model to represent a bounded outcome as a function of a linear combination of predictors generally transform the outcome so that it becomes unbounded. Transformation prevents derivation of fitted values that fall outside the range of theoretical possibilities—in this case, fitted values of hazard less than 0. When time is measured continuously, researchers build statistical models of the *natural logarithm* of hazard; when time is measured discretely and hazard is a conditional probability, a *logit* transformation is used for the same reason.

The effect of the logarithmic transformation on hazard is illustrated in figure 4, which presents sample log-hazard functions corresponding to the plots in panel B of figure 3. The log transformation has its largest effect on rates near 0, expanding the distance between values at this extreme. Nevertheless, in the transformed world of log-hazard, the predictor MILD works as it did before. When MILD = 1, the log-hazard function consistently is higher, relative to its location when MILD = 0, indicating that, at every possible time among individuals still in residence, those who have mild problems are more likely to leave. Still ignoring the minor differences in the shapes of the profiles, then, the predictor MILD essentially displaces the log-hazard profiles vertically, relative to each other.

Inspection of the sample relationship between the predictor MILD and the entire log-hazard profile in figure 4 leads to a reasonable specification for a population model of the hazard profile as a function of predictors. Letting *h(t)* represent the entire population hazard profile, a statistical model that captures this vertical displacement relates the log transformation of *h(t)* to the predictor MILD as follows:

$$\mathit{log}\ h(t) = \beta_0(\mathrm{t})+\beta_1\mathbf{MILD} \qquad (1)$$

The model parameter, β_0(t), is known as the *baseline log-hazard profile.* It represents the value of the outcome (the entire log-hazard function) in the population when the predictor (MILD) is 0 (i.e., because of the way the predictor MILD has been coded, it specifies the profile for individuals with severe problems). The baseline .s written as β_0(t), a function of time, and not as β_0, a single term unrelated to time (as in regression analysis), because the outcome (*log h(t)*) is a temporal profile. The model specifies that differences in the value of MILD "shift" the baseline log-hazard profile up or down. The "slope" parameter, β_1, captures the magnitude of this shift; it represents the vertical shift in log-hazard

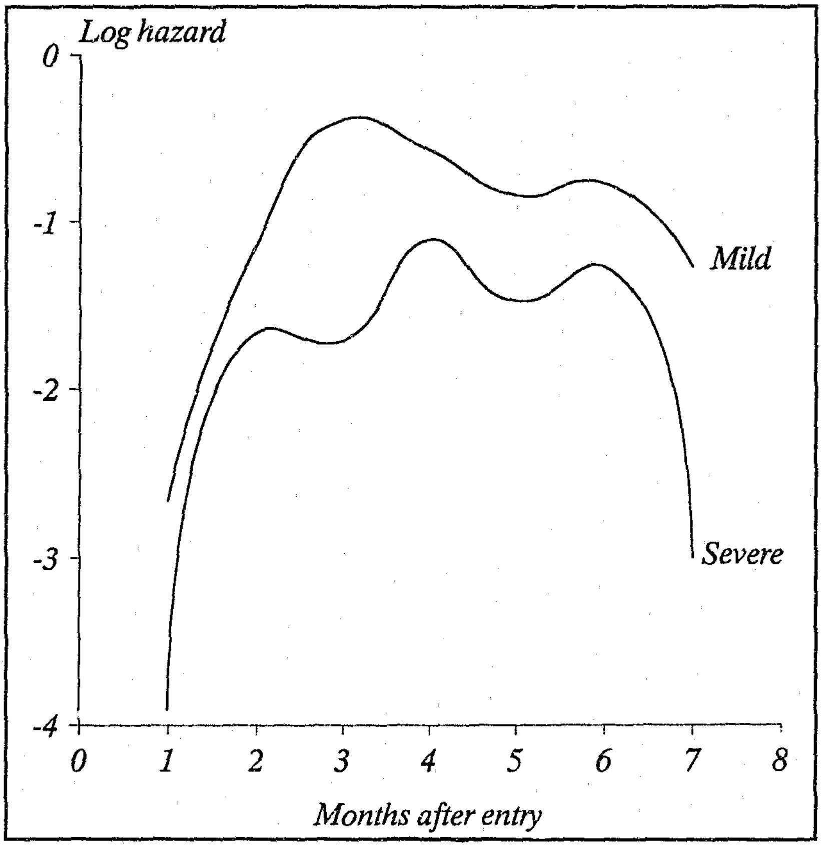


FIGURE 4. *Sample log-hazard functions for residents with "mild" and "severe" problems*

attributable to a one-unit difference in the predictor. Because the predictor in this example (MILD) is a dichotomy, β_1 captures the differential risk of leaving between individuals with mild and severe problems. If the model were fitted to these data, the obtained estimate of β_1 would be positive because those with mild problems are at greater risk of leaving in every month.

Hazard models closely resemble familiar regression models. Several predictors can be incorporated by including additional variables expressed as linear (or nonlinear) functions of additional unknown "slope" parameters on the right-hand side of the equation. This model expansion allows examination of one predictor's effect while controlling statistically for others' effects. Inclusion of cross-product terms enables examination of statistical interactions between predictors. It does not seem excessive to argue that hazard models provide the powerful, flexible, and sensitive approach to analyzing event occurrence that many drug abuse prevention researchers should be using. The goodness of fit of a hypothesized population model can be evaluated with data, allowing inferences about population relationships between hazard and predictors. As shown later, reconstructed survivor and hazard functions and estimated median lifetimes can depict the effects of predictors, providing answers to research questions in the original metric of interest—time.

Are the Hazard Profiles Proportional or Nonproportional?

Simple hazard models like equation (1) implicitly assume that all the log-hazard profiles corresponding to successive values of a predictor differ only by their relative elevation (described here by β_1). Under such models, but in the antilogged world of raw hazard, all the hazard profiles simply are magnifications or diminutions of each other—they are *proportional*. Under this *proportionality assumption*, which in continuous-time survival analysis is called the *proportional hazards assumption*, the entire family of log-hazard profiles represented by all possible values of the predictors share a common shape and are mutually parallel. Singer and Willett (1991, 1993) draw an analogy between this assumption and the assumption of homogeneity of regression slopes in the analysis of covariance.

Proportional hazards models are among the most popular survival analysis approaches used today, in part because most major statistical packages now provide programs for estimating their parameters using a method developed by Cox (1972). (Computer software for fitting hazard models is discussed in the **Where To Go To Learn More About Survival Analysis** section.) This ingenious strategy allows estimation of parameters like β_1 without the specification or estimation of the shape of the baseline hazard function, $\beta_0(t)$. For this reason, analogous to traditional nonparametric methods (which make no distributional assumptions), Cox regression is called semiparametric.

However, the tremendous boon of the semiparametric method—its ability to evaluate the effects of predictors without estimating the shape of baseline hazard profile—also is its principal disadvantage. The method is so general that it works for an unspecified baseline hazard profile of any shape. Without needing to explore the baseline hazard, investigators can examine effects of predictors without exploring absolute levels of risk. Because the baseline hazard function can be easily ignored, researchers may fail to recognize substantively and statistically important information contained only in the baseline hazard function.

What kinds of information can be found? The baseline hazard function and, under the proportionality assumption, its magnified and diminished cousins, describe the *pattern* and *magnitude* of risk over time—it indicates *when* the target event will occur and *how likely* that occurrence is (as in figure 2). The hazard profiles in figure 3, for example, show that individuals still in residence are most likely to be discharged in the third and fourth months after admission. All the predictor does is magnify or diminish this basic pattern of risk.

The ease with which the hazard function's shape can be ignored under the semiparametric method has a further ill consequence: it promotes the unthinking and dubious acceptance of the proportional hazards assumption. Currently available computer software makes it all too easy to examine the effects of predictors without examining the tenability of the underlying proportional hazards assumption. Notice, for example, that the sample log-hazard profiles in figure 4 are neither identical in shape nor parallel, suggesting that the proportional hazards assumption might not be tenable.

The tenability of the proportional hazards assumption must be viewed with some circumspection because those few researchers who have examined its tenability have found clear evidence of its violation. In their own research on employee turnover, for instance, the authors have found that violations of the assumption are the rule rather than the exception (Murnane et al. 1991; Singer 1993*a*, 1993*b*). A similar conclusion was reached by Bolger and colleagues (1989) in their study of adolescent suicide ideation.

This is an important issue because violation of the proportional hazards assumption is far more than a methodological nuisance. The magnitude and direction of the effects of predictors may be estimated incorrectly if the hypothesized statistical model inappropriately constrains the log-hazard profiles to be parallel with identical shapes. Ignoring such underlying failures can lead to incorrect substantive conclusions. In a very informative paper, Trussel and Hammerslough (1983) document differences in interpretation that arise when the proportional hazards assumption injudiciously is assumed tenable in a study of child mortality (compare their tables 3 and 4, particularly the effects of gender, birth order, and age of mother at birth). So uncertain is the veracity of the proportional hazards assumption that the authors always begin their own data analyses with the entirely opposite view. Along with unicorns and normal distributions (Micceri 1989), the authors regard the proportional hazards assumption as problematic in any set of data until proven otherwise. Before adopting a proportional hazards model, researchers at least should subdivide their sample by substantively important values of critical predictors and inspect the shapes of the sample hazard profiles within these subgroups. Arjas (1988), Harrell and Lee (1986), Kalbfleisch and Prentice (1980), and Willett and Singer (1993) provide methods for exploring the tenability of the proportionality assumption. Finally, as discussed below, researchers easily can adopt a broader analytic approach—one that tests the proportional hazards assumption and fits nonproportional hazard models if they are required.

What Different Types of Predictors Can Be Included in Hazard Models?

One important advantage of the hazard-modeling framework is that it permits the simultaneous study of both *time-invariant* and *time-varying* predictors. As befits their label, time-invariant predictors describe immutable characteristics of individuals; the values of time-varying predictors, in contrast, may fluctuate over time. While investigating the monthly risk of initiating marijuana use in late adolescence, for example, Yamaguchi and Kandel (1984) examined predictors of both types. In the study, 1,325 adolescents were interviewed once in high school and reinterviewed 9 years later at age 24 or 25. In the followup interview, respondents retrospectively reconstructed monthly charts of their drug and life histories. The researchers examined the effects of truly time-invariant predictors, such as race, whose values are immutable over time, but other variables such as friends' use of marijuana, involvement in delinquent activities, and belief that marijuana use is not harmful also were treated as time-invariant predictors of the risk of initiation of marijuana use because they were measured on a single occasion during the initial high school interview.

The researchers also examined the effects of time-varying predictors, such as current alcohol use and current cigarette use, whose monthly values were obtained during life-history reconstruction at followup. Using hazard models, the researchers were able to present convincing evidence that the "current use of alcohol and cigarettes have strong effects on the initiation of marijuana use among men and women" and "controlling for selected antecedent behavioral, attitudinal, and environmental factors measured in adolescence, . . . friends' use of marijuana has the strongest positive influence on initiation of marijuana" (Yamaguchi and Kandel 1984, p. 675). Interestingly, when the initiation of *prescribed* psychoactive drug use was examined later in the paper, Yamaguchi and Kandel found that "multiple factors are involved in the progression to prescribed drugs, with adolescent depressive symptomatology and use of other illicit drugs important for both sexes, and maternal use of psychoactive drugs, dropping out of school, and prior use of marijuana of additional importance for women" (p. 673). These same authors also have used hazard-modeling to study links between time-varying drug consumption and the risk of premarital pregnancy (Yamaguchi and Kandel 1987) and the risk of job turnover (Kandel and Yamaguchi 1987).

The hazard model in equation (1) includes a single time-invariant predictor, MILD. The information contained in this predictor—whether the patient suffers mild or severe problems—remains constant over time. β_1 quantifies the time-invariant effect of this time-invariant predictor on the risk of discharge. Hazard models like equation (1) can be extended easily to include time-varying predictors. Such extensions can be helpful particularly in prevention research, where the values of important predictors often vary naturally over time.

Hazard models with time-varying predictors closely resemble the model in equation (1). In Yamaguchi and Kandel's study (1984) of the risk of marijuana initiation, for example, one possible population hazard model might include: (1) HSDEPRESS, a predictor treated as time invariant because it describes whether the individual ever suffered clinical depression during high school (authors' coding: 0 = never clinically depressed, 1 = suffered clinical depression during high school); and (2) ALCOHOL, a time-varying predictor whose monthly values are known throughout adolescence (0 = not currently using, 1 = currently using). Such a model might be:

$$\boldsymbol{\mathit{log\ h(t)} = \beta_0(t)+\beta_1\mathrm{HSDEPRESS}+\beta_2\mathrm{ALCOHOL}\mathit{(t)}} \tag{2}$$

The parenthetical "t" in the predictor ALCOHOL*(t)* indicates that the values of this predictor may vary over time. Unit differences in ALCOHOL correspond to shifts in the log-hazard profile of β_2. Although the values of the predictor ALCOHOL may differ over time, each one-unit difference anywhere produces the same shift of β_2 in the appropriate part of the log-hazard profile. So, while the model includes a time-varying predictor, the per-unit effect of that predictor on log-hazard is constant over time.

Another way to understand the effects of time-varying predictors is to conceptually regard the outcome in equation (2)—the log-hazard profile—as a temporally sequenced list (a vector) of marijuana-initiation risks. The predictors also can be viewed as an ordered list of values that, for each person, describes the values of HSDEPRESS and ALCOHOL over time. Each element in the hazard list corresponds to an element in each predictor's list. For a predictor treated as time invariant, such as HSDEPRESS, all elements in each person's predictor list are identical—1 for every person who was ever clinically depressed in high school, and 0 for every person who was not. For a time-varying predictor like ALCOHOL, in contrast, the values in the predictor list may differ from month to month. If an individual does not use alcohol initially, the early elements in the ALCOHOL vector are 0; when alcohol use begins, the values change to 1. If alcohol use persists, the values stay as 1; if it ends, the values revert to 0. Each person has his or her own alcohol use pattern; the number of patterns across individuals is limited only by the number of possible states and occasions of measurement. The hazard model simply relates the values in one list (the hazard vector) to the values in the other (the predictor vector), regardless of whether the elements in the latter list are identical to each other.

Time itself is the fundamental time-varying predictor. So, conceptually at least, one might argue that it, too, should be included as a time-varying predictor in equation (2), mapping intrinsic changes in the risk of marijuana initiation over time. Although intuitively appealing, this approach produces complete redundancy in the model because this time-varying effect already is captured by the baseline log-hazard function, $\beta_0(t)$. $\beta_0(t)$ describes the chronological pattern of baseline risk—the differences in log-hazard attributable solely to time. Estimation of the baseline hazard function is tantamount to estimation of "the main effect of time." This analogy reinforces the need to examine the shape of the baseline hazard, for it provides information about the effects of the fundamental time-varying predictor—time itself.

Can Predictors in Hazard Models Interact With Time?

Not only can predictors themselves be time invariant or time varying, their *effect* on hazard also can be constant or vary over time. By including a main effect of the predictor HSDEPRESS in equation (2), the vertical displacement associated with clinical depression in high school is assumed to be the same at age 16 and age 24 (and equal to β_1). However, the assumption of temporally immutable effects may not hold in reality—the effects of some predictors will vary over time. The impact of depression in high school on the risk of marijuana initiation might decline as time passes and the individuals mature. If so, the distance between the hazard profiles associated with different values of the predictor HSDEPRESS would narrow over time.

When the effect of one predictor on an outcome differs by levels of another predictor, statisticians say that the two predictors *interact.* If the effect of a predictor like HSDEPRESS on an outcome like the risk of marijuana initiation differs across time, the predictor HSDEPRESS is said to *interact with time.* Predictors that interact with time have important substantive interpretations, allowing researchers to build complex models of the relationship between predictors and risk. If a predictor primarily affects early risks, the hazard profiles will be separated widely in the beginning of time and converge as time passes. If a predictor primarily affects late hazards, it will have little effect at the beginning of time but will widen the distance between hazard functions on each subsequent occasion.

One's understanding of event occurrence can be improved vastly by exploring whether the effects of predictors remain constant or vary over time. As Verhulst and Koot (1991) note, "what may be a risk factor at one developmental phase may not be at another" (p. 363). Some recent studies that look for such interactions indeed are finding their presence. In their study of the age at first suicide ideation, for instance, Bolger and colleagues (1989) detected interactions between two key predictors and time. Dividing time into two broad periods—adolescence and preadolescence—they found that the effects of *respondent race* and *parental absence in childhood* both differed across these periods. With regard to race, during preadolescence, Bolger and colleagues (1989) found that white children were less likely to consider suicide than nonwhite children but, during adolescence, they were more likely to do so. With regard to parental absence, they found that, during preadolescence, children who experienced a parental absence were more likely to consider suicide than those who did not experience such absence, but, during adolescence, parental absence had little impact on the risk of suicidal thought. In a reanalysis of the National Institute of Mental Health Collaborative Study of Maintenance Treatment of Recurrent Affective Disorders, Greenhouse and colleagues (1991) found that the efficacy of selected antidepressants in preventing recurrence was pronounced only during the first few weeks after treatment initiation. By including interactions between predictors and time, researchers can better identify the predictors of risk over time.

If a predictor interacts with time, the proportionality assumption is violated, and models such as the proportional hazards model introduced in equations (1) and (2) do not represent reality. The proportionality assumption is tested easily by adding an interaction with time to the hazard model and assessing the effect of this new predictor. If the assumption holds, the interaction term will have no effect and can be removed. If the interaction term proves to be an important predictor of the hazard profile, then a violation of the proportionality assumption has been detected and the interaction with time must remain in the model to ensure the appropriate estimation of predictor effects. It is recommended that researchers routinely examine the effects of such interactions in their hazard models, just as they would routinely examine interactions among other predictors in traditional linear models.

What Is Discrete-Time Survival Analysis?

The hazard models posited above, which assume that time can take on any nonnegative value, represent the hazard profile as a *continuous* function of time as reflected, for example, in the parenthetical inclusion of the symbol "t" in the expression for the baseline hazard function, $\beta_0(t)$. When data are collected in *discrete* time, however, either because the events occur or are measured only at specific times—perhaps every week, month, academic semester, or year—researchers should consider a different class of survival methods known as *discrete-time survival analysis*. The method is easy to apply, facilitates the estimation of the baseline hazard function, encourages the testing of the proportionality assumption, and enables researchers to fit hazard models using procedures available in most statistical computer packages. For all these reasons, the authors encourage its wider application to studying questions about time.

The discrete-time survival analysis approach is described in detail in two recent papers (Singer and Willett 1993; Willett and Singer 1993); this chapter simply gives an overview. A researcher conducts a discrete-time survival analysis by altering the data structure, transforming the standard one-person, one-record data set (the "person" data set) into a one-person, multiple-period data set (the "person-period" data set). In the new person-period data set, a dichotomous variable is created to summarize the pattern of event occurrence in each discrete time period for every person in the sample. If relapse into cocaine use were being studied, for instance, this variable (RELAPSE) would be coded "0" if no relapse occurred and "1" if it did occur, *in each discrete time interval.* So, for instance, an ex-addict who relapsed in the sixth month after treatment would have six lines of "data" in the new person-period data set and, in each line, RELAPSE would take on a value specific to that interval—the first five being "0," the last being "1." The researcher also creates a set of "time indicators" that index and distinguish the discrete time intervals themselves.

Under the discrete-time approach, the relationship between the dichotomous event summary (RELAPSE) and predictors (including the time indicators) can be fit using a modification of standard logistic regression programs. Interactions among predictors, and between predictors and the time indicators, are included easily by forming cross-products in the person-period data set and using them as predictors. Adding these interactions to main-effects models facilitates easy testing of the proportional hazards assumption, and, if the assumption is violated, retention of the interactions in the fitted model ensures the appropriate estimation of the effects.

The use of a standard logistic regression computer package to fit discrete-time hazard models eliminates the need for dedicated software and, consequently, brings the new methodology within the grasp of all empirical researchers. The logistic regression parameter estimates, standard errors, and goodness-of-fit statistics are exactly those required for testing hypotheses about the effect of predictors on the discrete-time hazard profile (Singer and Willett 1993). Allison (1982, p. 82) comments that these estimates are "consistent, asymptotically efficient, and asymptotically normally distributed" and that, despite the apparent inflation of sample size on creation of the person-period data set, the estimated standard errors are consistent estimators of the true standard errors.

Because of the frequency with which prevention researchers use discrete-time data collection strategies, readers are encouraged to learn more about discrete-time survival methods. In the Yamaguchi and Kandel (1984) study of drug use described earlier, for example, participants reconstructed their life histories on a *month-by-month* basis. Many other researchers follow subjects at discrete points in time. Morgan and colleagues (1988), for example, conducted followups 2, 3, and 8 weeks after cessation. Harackiewicz and colleagues (1987) used 3-month intervals after an initial 6-week followup. Marlatt and colleagues (1988) conducted followups after 1 and 4 months and 1 and 2 years.

How Can Fitted Models Be Interpreted?

Fitting statistical models is of little use unless the researcher can interpret the resultant information clearly and persuasively. Interpretation includes at least three components: identification of "statistically significant" effects, computation of numerical summaries of effect size, and graphical display of the magnitude and direction of the effects. In traditional ANOVA, for example, a researcher first might determine whether the difference in average outcome between two groups is statistically significant, and if it is, he or she then might express one group's advantage in "standard deviation" units and provide data plots comparing the distribution of the outcome across groups.

The interpretation of survival analysis also must include the same three components. However, because hazard models may be difficult to conceptualize (describing, as they do, variation in entire hazard profiles), graphical techniques may provide a better vehicle for reporting findings.

Graphics can help communicate complex and unfamiliar ideas about whether an event occurs, and, if so, when. Yet, even the most effective graphical displays must be supported by documentation of parameter estimates and associated standard errors. So the discussion of interpretation begins with the computer output commonly generated by statistical packages.

Computer output that documents the results of fitting hazard models closely resembles output that documents the results of other statistical techniques. Most programs output estimates of the "slope" parameters, the standard errors of these estimates, the ratio of each parameter estimate to its standard error (a "*t*-statistic"), and a *p*-value based on the *t*-statistic for testing the null hypothesis that the corresponding parameter is 0 in the population (given that the other predictors are in the model). Some programs output a χ^2 statistic in lieu of a *t*-statistic; the accompanying *p*-value assesses the improvement in fit resulting from adding the predictor to a reduced model containing all the other predictors.

Researchers frequently provide tables of some, or all, of these summary statistics in the accounts of their analyses (e.g., Yamaguchi and Kandel 1984, tables 1, 2, and 3). When doing so, however, researchers should not ignore the sign and magnitude of the "slope" estimate by focusing on the associated *p*-values. Although *p*-values can help identify critical predictors, they indicate nothing about the direction and relative magnitude of effects.

Because hazard models represent relationships between the entire hazard profile and predictors, specifying an understandable effect size is not easy. One useful approach is to interpret the parameter estimate associated with each predictor in a way similar to interpreting a regression coefficient. In continuous-time survival analysis, the parameter estimate represents a difference in elevation of the log-hazard profile corresponding to predictor values one unit apart. The parameter estimate's sign indicates the direction of the movement, indicating whether positive differences in the value of the predictor correspond to positive or negative differences in the risk of event occurrence. It may be helpful to imagine the profile on a log-hazard plot "moving" up (or down, if the estimate is negative) for a one-unit difference in the predictor. Predictors with larger parameter estimates produce larger elevation differences per unit difference in the predictor. (In discrete-time survival analysis, the conceptualization is identical but the interpreter of the findings is dealing with differences in the elevation of the *logit*, rather than log, hazard profile.)

Even after considerable experience with hazard models, however, ready visualizations in the transformed world of log-hazard may remain tortured. A mathematically complex but intuitively simple approach involves the transformation of the outcome back into the more familiar metric of "risk" antilogging parameter estimates as necessary. Of course, a researcher must use different transformations and interpretations depending on whether continuous- or discrete-time models have been fitted.

These ideas are illustrated with the continuous-time hazard model in equation (1). Antilogging botl sides:

$$h(t) = e^{\beta_0(t)} e^{\beta_1 MILD} \qquad (3)$$

Because MILD = 1 for individuals with mild problems and MILD = 0 for those with severe problems, the hazard functions corresponding to these two groups are:

$$h(t\colon severe) = e^{\beta_0(t)} \quad and \quad h(t\colon mild) = e^{\beta_0(t)} e^{\beta_1} \qquad (4)$$

The risk profile in the mild group simply is the risk profile in the severe group multiplied by e^{β_1}. This multiplicative rule applies to both categorical and continuous predictors. So in continuous-time hazard models, antilogged parameter estimates yield numerical multipliers of risk-per-unit difference in the predictor. If the antilogged parameter estimate is greater than 1, risk is higher in the reference group; if it is less than 1, risk is lower.

This transformation strategy enabled Hall and colleagues (1991) to document the strong effect of commitment to abstinence on the risk of relapse to cocaine use. After controlling statistically for selected demographic covariates and route of administration, the researchers obtained a parameter estimate of 0.42 for a time-varying covariate indicating whether the former cocaine users had a goal of absolute abstinence (χ^2 (1, N = 103) = 7.14, p = .0076). Hall and colleagues (1991) interpreted the antilog of this estimate ($e^{0.420}$ = 1.5) by writing that "subjects who endorsed abstinence were less than half as likely to lapse subsequently as were subjects who endorsed less stringent goals" (p. 529).

Another way to interpret hazard-model parameter estimates is in terms of *percentages difference in risk.* Doubling the baseline risk (multiplying by a factor of 2) is equal to a 100-percent increase in risk; halving the baseline risk (multiplying by a factor of .5) is equal to a 50-percent decrease. So, in the cocaine relapse study conducted by Hall and colleagues (1991) above, multiplying the baseline hazard by .5 corresponds to a 50-percent decrease in the risk of relapse for those with a commitment to total abstinence. The general rule is simple: The percentage difference in risk-per-unit difference in the predictor is $100(e^{\beta}-1)$. Some researchers automatically add these estimates of e^{β} (or $100(e^{\beta}-1)$) to tables reporting parameter estimates, standard errors, *t*-statistics, and *p*-values.

Similar but modified interpretations can be made after fitting discrete-time hazard models. Since discrete-time hazard is the conditional probability that an event will occur in a particular time interval (given that it has not yet occurred before the interval), the discrete-time hazard model, which uses logit-hazard as the outcome, expresses the relationship between predictors and the *log odds* of event occurrence. Estimates of e^{β} or $100(e^{\beta}-1)$, therefore, are multipliers of, or percentage increases or decreases in, the *odds* of an event occurring (Rosenbaum and Kandel 1990).

As these illustrations document, numeric and algebraic strategies are not the last word in the clear communication of the findings of survival analysis. Apart from being arithmetically convoluted, they have at least two other drawbacks. First, they ignore the shape of the baseline hazard function—they indicate only the extent to which one risk profile is a magnification or diminution of another. As argued earlier, the shape of the hazard profile—the temporal placement of its peaks and valleys—indicates much about the survival process under investigation. Second, algebraic interpretations are useful only if the proportionality assumption is met. If the effect of predictors differs over time, risk profiles no longer will be parallel in log- or lógit-space, and so it makes little sense to talk about one profile being "rescaled" to generate the other. If the shapes of the risk profiles differ dramatically, algebraic interpretations may not only oversimplify findings, they may even misrepresent them completely.

Presenting fitted hazard plots, fitted survival plots, and estimated median lifetimes resolves these problems. Most computer programs provide procedures for recovering fitted profiles from parameter estimates. By appropriately substituting back into the hazard model, a researcher can generate fitted hazard profiles at substantively interesting values of the predictors for the range of time values spanning the data collection period. The use of fitted hazard profiles is clear, comprehensive, and intuitively meaningful. Fitted profiles demonstrate the effect of predictors on risk and pinpoint whether these effects rise, fall, or remain constant with the passage of time. By presenting fitted hazard functions, a researcher need not struggle to describe effects using abstract scaling factors and percentage increases that ignore important interactions with time.

Researchers should consider their original questions and analytic findings when selecting predictor values for constructing fitted plots. Questions to ask include: "Which predictors were emphasized in the research questions?" and "Which predictors were significantly associated with hazard?" Use predictors that are substantively and statistically important when generating the fitted profiles; lesser variables can be included as "controls" by equating their value to their sample averages.

Fitted survivor functions and estimated median lifetimes also can be reconstructed from the fitted hazard profiles in order to illustrate the magnitude and direction of important effects. However, fitted hazard profiles generally are more informative because they identify the specific times when the events of interest are most likely to occur. It usually is more difficult to discern differences between fitted survivor profiles than between fitted hazard profiles because the survivor function is "smoothed" by the cumulation of risk over time.

The advantages of this graphical approach are illustrated in figure 5 using data from Hall and colleagues (1991), who studied the risk of relapse to cocaine use among 104 former users who participated in a treatment program. Among the many predictors Hall and her colleagues studied, there was a strong and statistically significant effect of the route of administration prior to entry into treatment (ROUTE), here divided into two groups: those who used cocaine intranasally and all others. Figure 5 presents fitted hazard and survivor functions based upon a discrete-time hazard model that included this single predictor. Because a discrete-time hazard model has been fitted here, the fitted values of the survivor function and hazard function are joined using line segments rather than a smooth curve.

Comparison of the two fitted hazard functions in figure 5 demonstrates the large differential in risk of relapse associated with route of administration. In every week after treatment, intranasal users are far less likely than other users to relapse. These fitted functions have the same basic shape, and one appears to be a magnification of the other.[3] Were these hazard functions to be replotted on a logit-hazard scale, they would have a constant vertical separation. The functions have been constrained to appear this way by the proportionality assumption, which was tested for and found to be met.

The fitted survivor plots in panel B of figure 5 show the cumulative effects of the large weekly differentials in risk. Unlike the fitted hazard functions that emphasize large and consistent differences in risk, the fitted survivor functions condense the effects of these weekly risk differentials together to reveal a substantial difference between the groups. Focusing on the last fitted survival probability, for example, it is estimated that 12 weeks after treatment ended, 63 percent of the intranasal users remained abstinent, as compared with 28 percent of other users.

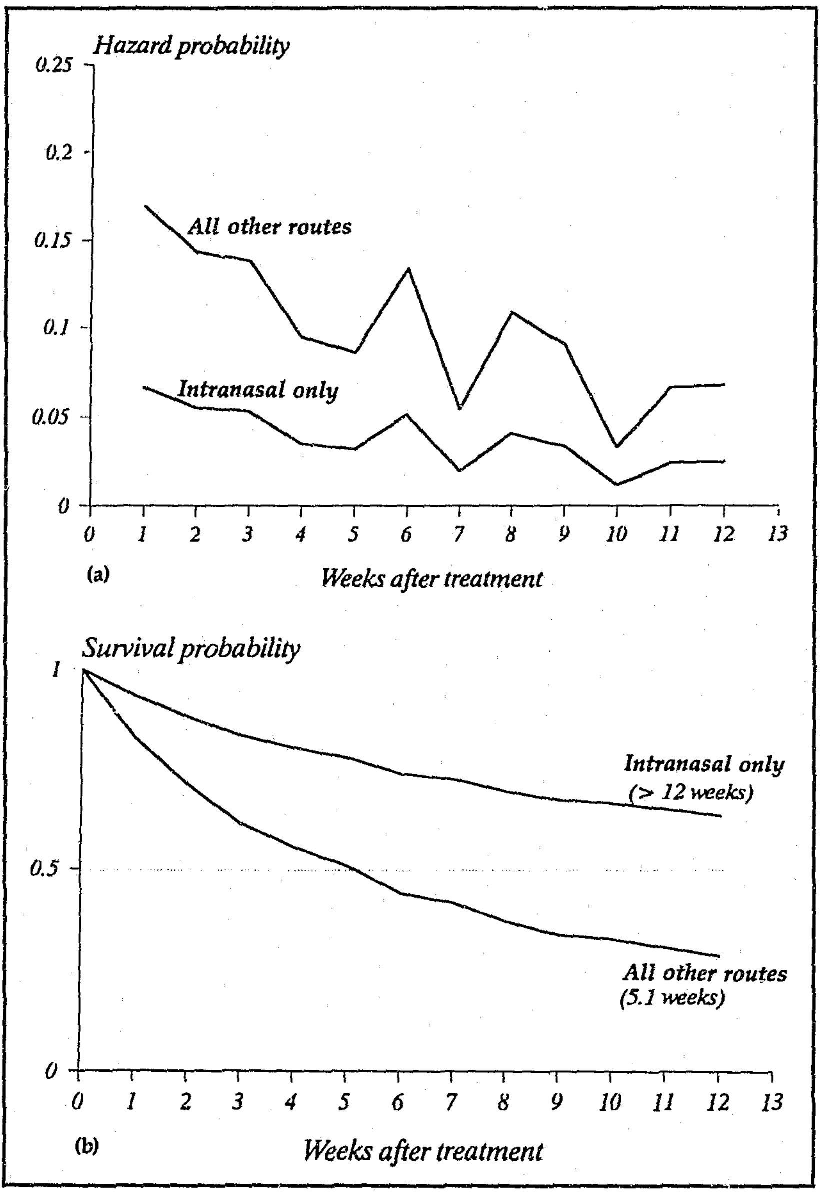


FIGURE 5. *Fitted hazard functions (panel A) and survivor functions (panel B) describing the risks of relapse for 104 former cocaine abusers following treatment, by route of cocaine administration prior to treatment (intranasal versus all others).*

SOURCE: Based on data reported by Hall and colleagues (1991)

A third perspective on the divergent relapse patterns of these two groups comes from comparison of the estimated median lifetimes displayed in panel B of figure 5: more than 12 weeks for intranasal users versus 5.1 weeks for all other users. Even though censoring prevented estimating a median lifetime precisely for intranasal users, the large difference between these "average" relapse times powerfully communicates the analytic results.

IS SURVIVAL ANALYSIS REALLY NECESSARY?

The methods of survival analysis provide a powerful and flexible set of tools for studying many questions arising in drug abuse prevention and intervention. Although increasing numbers of researchers are using the methods, many others studying onset, duration, recovery, recidivism, relapse, and recurrence have yet to exploit this new analytic tool.

One reason survival methods have not yet been used widely when studying questions about event occurrence is that many researchers still wonder whether the methods really are necessary. Although this view rarely is expressed explicitly, reading between the lines suggests that many researchers believe that traditional analytic approaches usually will suffice.

The authors agree that some skepticism is healthy. Why bother with complex methods if simpler methods will do? Unfortunately, the problem when studying event occurrence is that simpler methods will not always suffice. To illustrate this point, this chapter is concluded by describing five ways in which traditional methods can obscure important information about event occurrence—information that sensitively and assuredly is revealed by survival analysis methods.

First, answers obtained by researchers using traditional methods inextricably are linked to the particular timeframe chosen for data collection and analysis; yet, in prevention intervention research, these timeframes rarely are substantively motivated. Researchers comparing 6-month, 1-year, or

5-year relapse rates for individuals participating in different treatment programs, for example, simply are describing *cumulative* differences in behavior until these times. All other variation over time in the risk of relapse is lost. The literature is filled with examples of disparate risk profiles that lead to comparable relapse rates at specific points in time (e.g., Cooney et al. 1991, figure 1; Harackiewicz et al. 1987, table 2). Just because two groups of subjects have identical relapse rates at one point in time does not mean that they followed similar trajectories to get there—most of those in one group might have relapsed in the first month while those in the other might have been equally likely to relapse at all points in time. The 6-month, 1-year, and 2-year cutpoints used in the past are convenient but not purposeful. By documenting variation in risk over time and by discovering what predicts variation in risk, researchers can better understand why people relapse. Traditional methods disregard this information; with survival methods, variation in risk becomes the primary analytic focus.

Disregard for variation in risk over time leads to a second problem with traditional methods: seemingly contradictory conclusions can result from nothing more than variations in the particular timeframes studied. Had Stevens and Hollis (1989) computed only 1-month and 12-month relapse rates when evaluating the efficacy of their individually tailored skills-training technique for preventing relapse to smoking, for example, they would have reached opposite conclusions: the 1-month rates would have shown that subjects in the skills group were *more* likely to relapse (in comparison to those in a discussion-oriented group) while the 1-year rates would have shown that they were *less* likely. By thoughtfully presenting sample survivor functions, Stevens and Hollis showed that the effectiveness of the skills-training approach revealed itself only after several months. Researchers using traditional methods constantly must remind themselves that conclusions can change as the timeframe changes. While such caveats usually appear in the "Methods" section of an article, they often disappear in the "Discussion" section. In survival analysis, the timeframe itself is integral to the answer; it highlights rather than obscures variation over time.

Third, traditional analytic methods offer no systematic mechanism for incorporating censored observations in the analyses. If all the censored observations occur at the same point in time, traditional data analysis can collapse the sampled individuals into two groups: those who experience the event before the censoring point, and those who do not. In their longitudinal study of unaided smoking cessation, for example, Marlatt and colleagues (1988) compared ex-smoker subjects who relapsed and those who did not at each of four points in time: 1 month, 4 months, 1 year, and 2 years after quitting. If the first days and weeks following cessation are the hardest, individuals who relapse soon after cessation may differ systematically from those who relapse subsequently. Dichotomization conceals such differences; survival methods, which focus on the risk of event occurrence over time, bring such differences to light.

If censoring does not occur at the same timepoint for every individual under study (as when researchers follow cohorts of patients admitted over time until a single fixed point in time), traditional methods create a fourth problem: If censoring times vary across people, the risk periods vary as well. People followed for longer periods of time have more opportunities to experience the target event than do those followed for shorter periods of time. This means that observed differences in rates of event occurrence might be attributable to nothing more than research design. In thւ study by Goldstein and colleagues (1991) of suicidality among 1,906 Iowans with affective disorders, the followup period ranged from 2 to 13 years. As they note, "The highly variable period of follow-up is also a potential limitation, because those patients followed up for the shortest periods may not have been given the opportunity for their suicidal outcome to emerge" (p. 421). Had the researchers used survival methods instead of logistic regression, they would have been better able to address this concern because each person who did not commit suicide simply would have been censored at followup.

Fifth, traditional analytic methods offer few mechanisms for including predictors whose values vary over time or for permitting the effects of predictors to fluctuate over time. To overcome this limitation, researchers studying the effects of time-varying variables tend to use predictor values corresponding to a single point in time, the average of predictor values over time, or the rate of change in predictor values over time. This is not necessary in survival analysis. The analytic effort is identical whether the study is including predictors that are static over time or predictors that change over time; so, too, it is easy to determine whether the *effects* of predictors are constant over time or whether they differ over time. There is no need to create a single-number summary of the temporal behavior of a changing predictor. Traditional methods force researchers into building static models of dynamic processes; survival methods allow researchers to model dynamic processes dynamically.

Researchers in prevention research are encouraged to investigate the design and analytic possibilities offered by survival methods. When these methods were in their infancy and statistical software was either not available or not user friendly, researchers reasonably adopted other approaches. However, experience elsewhere in medicine and in the social sciences shows that these methods, originally developed to model human lifetimes, lend themselves naturally to the study of other phenomena as well. While software lags behind, this is an area of active research with rapidly improving options (Harrell and Goldstein, in press).

Researchers rarely ask questions that they do not have the analytic methods to answer. Many researchers who have been interested in the timing of events have modified their questions because they did not know how to build appropriate statistical models. The authors hope that this presentation of survival analysis will help researchers reframe these modified questions and provide them with strategies for answering those questions as simply and as directly as possible.

Where To Go To Learn More About Survival Analysis

In the body of this chapter, the discussion of technical statistical issues that arise in survival analysis has been purposefully avoided; indeed, the authors have gone to great pains to ensure that the text is relatively free of technicality. The goal of this chapter has been to make a strong case for the use of survival methods in prevention research. For readers actually considering the use of survival methods, this section provides references to written materials to consult before embarking on a study.

Readers interested in acquiring a more sophisticated background in these methods can choose from among a wide range of published material, both in books and in scholarly journals. An introductory monograph (Allison 1984) provides an excellent starting point for readers familiar with regression. It is a well-documented, accessible, and largely nontechnical introduction to a broad range of survival methods. In less than 100 pages, Allison touches on most of the important issues facing the user of survival analysis, including discrete- versus continuous-time methods, the proportional hazards model and partial likelihood estimation (Cox regression), the analysis of competing risks, and repeated events.

Scattered through the scholarly literature are a variety of accessible articles that can be used to supplement Allison's overview. Many of these provide nontechnical reviews of the application of survival methods in particular substantive areas. Anderson and colleagues (1980) use a medical setting to present a readable introduction to many aspects of survival analysis, ranging from displays and single-number summaries through life-table testing and hazards-modeling. And in a recent pair of papers, Singer and Willett (1991) and Willett and Singer (1991), expand on the nontechnical overview offered here by reviewing applications of survival analysis in psychological and educational research.

Readers wishing to supplement these introductions with greater technical detail should consult one of the several "standard" texts. Although mathematically complex, Kalbfleisch and Prentice (1980) is a thorough and well-written source. Other texts of similar stature are Cox and Oakes (1984) and Miller (1981). In addition, there has been important methodological work on survival methods (known in sociology as event history methods) pioneered by Mayer and Tuma (1990), Petersen (1991), and Tuma and Hannan (1984).

Researchers collecting data in discrete rather than continuous time should learn more about discrete-time survival analysis. In addition, because discrete-time hazard models are easy to apply, facilitate the recapturing of the baseline hazard and survivor functions, can be estimated with standard logistic regression software, and allow the testing and, if necessary, the relaxation of the proportionality assumption, even researchers with continuous-time data also might want to explore this approach more fully. In a pair of articles, Singer and Willett (1993) and Willett and Singer (1993) provide an overview of discrete-time methods written for empirical rese?-chers. The article by Willett and Singer (1993) is the place to start for those seeking a data analytic perspective; the article by Singer and Willett (1993) offers a more mathematical presentation. Readers seeking further technical details on discrete-time methods can consult Allison (1982), Efron (1988), or Laird and Olivier (1981).

NOTES

1. The order of the authors was determined by randomization. This chapter was completed while the authors were American Statistical Association/National Science Foundation Fellows at the National Center for Education Statistics. Some of the material presented in this chapter is taken from two earlier papers (Singer and Willett 1991; Willett and Singer 1991). Address correspondence to either author at Harvard University, Graduate School of Education, Appian Way, Cambridge, MA 02138.

2. The authors estimated the sample survivor function in figure 1 using summary data kindly supplied by Dr. Victor J. Stevens (Stevens and Hollis 1989, figure 1, p. 422) using the Kaplan-Meier product limit method (Kalbfleisch and Prentice 1980). The authors then smoothed the obtained discrete estimates using a spline function (after the recommendation of Miller [1981]). The same method was used to create figures 2, 3, and 4. Their intentions were strictly pedagogic. They wished to use continuous-time survivor and hazard functions

to introduce the concepts of survival analysis before discussing the differences between continuous-time and discrete-time methods.

3. Strictly speaking, this apparent magnification of one hazard profile to give the other is only approximate in the discrete-time hazard model and only holds when h_j is small. For further discussion, see Wil!ett and Singer (1993).

REFERENCES

Adler, I., and Kandel, D.B. Adolescence in France and Israel: Application of survival analysis to cross-sectional data. *Soc Forces* 62:375-397, 1983.

Allison, P.D. Discrete-time methods for the analysis of event histories. In: Leinhardt, S., ed. *Sociological Methodology.* San Francisco: Jossey-Bass, 1982. pp. 61-98.

Allison, P.D. Event history analysis: Regression for longitudinal event data. In: *Sage University Paper Series on Quantitative Applications in the Social Sciences* (#07-046). Beverly Hills, CA: Sage Publications, 1984.

Anderson, S.; Auquier, A.; Hauck, W.W.; Oakes, D.; Vandaele, W.; and Weisberg, H.I. *Statistical Methods for Comparative Studies: Techniques for Bias Reduction.* New York: Wiley, 1980.

Arjas, E. A graphical method for assessing goodness of fit in Cox's proportional hazards model. *J Am Stat Assoc* 83:204-212, 1988.

Baer, J.S., and Lichtenstein, E. Classification and prediction of smoking relapse episodes: An exploration of individual differences. *J Consult Clin Psychol* 56:104-110, 1988.

Bernstein, D., and Lagakos, S.W. Sample size and power determination for stratified clinical trials. *J Stat Comput Simul* 8:65-73, 1978.

Biemer, P.P.; Groves, R.M.; Lyberg, L.E.; Mathiowetz, N.A.; and Sudman, S. *Measurement Errors in Surveys.* New York: John Wiley, 1991.

Biglan, A.; Hood, D.; Brozovsky, P.; Ochs, L.; Ary, D.; and Black, C. Subject attrition in prevention research. 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. pp. 213-234.

Blossfeld, H.P.; Hamerle, A.; and Mayer, K.U. *Event History Analysis: Statistical Theory and Application in the Social Sciences*. Hillsdale, NJ: Lawrence Erlbaum, 1989.

Bolger, N.; Downey, G.; Walker, E.; and Steininger, P. The onset of suicide ideation in childhood and adolescence. *Youth Adolesc* 18:175-189, 1989.

Borchardt, C.M., and Garfinkel, B.D. Predictors of length of stay of psychiatric adolescent inpatients. *J Am Acad Child Adolesc Psychiatry* 30(6):994-998, 1991.

Bradburn, N.M. Response effects. In: Rossi, P.H.; Wright, J.D.; and Anderson, A.A., eds. *Handbook of Survey Research*. New York: Academic Piess, 1983. pp. 289-328.

Bradburn, N.M.; Rips, L.J.; and Shevell, S.K. Answering autobiographical questions: The impact of memory and inference on surveys. *Science* 236:157-161, 1987.

Brownell, K.D.; Marlatt, G.A.; Lichtenstein, E.; and Wilson, G.T. Understanding and preventing relapse. *Am Psychol* 41:765-782, 1986.

Coelho, R.J. Self-efficacy and cessation of smoking. *Psychol Rep* 54:309-310, 1984.

Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*. 2d ed. Hillsdale, NJ: Erlbaum, 1990.

Cooney, N.L.; Kadden, R.M.; Litt, M.D.; and Getter, H. Matching alcoholics to coping skills or interactional therapies: Two-year follow-up results. *J Consult Clin Psychol* 59:598-601, 1991.

Condiotte, M.M., and Lichtenstein, E. Self-efficacy and relapse in smoking cessation programs. *J Consult Clin Psychol* 49:648-658, 1981.

Coryell, W.; Keller, M.; Lavori, P.; and Endicott, J. Affective syndromes, psychotic features, and prognosis. *Arch Gen Psychiatry* 47:651-662, 1990.

Cox, D.R. Regression models and life tables. *J R Stat Soc [B]* 34:187-202, 1972.

Cox, D.R., and Oakes, D. *Analysis of Survival Data* London: Chapman and Hall, 1984.

Crider, D.M.; Willits, F.K.; and Bealer, R.C. Tracking respondents in longitudinal surveys. *Public Opin Q* 35:613-620, 1971.

Crider, D.M.; Willits, F.K.; and Bealer, R.C. Panel studies: Some practical problems. *Sociol Meth Res* 2:3-19, 1973.

Donner, A. Approaches to sample size estimation in the design of clinical trials: A review. *Stat Med* 3:199-214, 1984.

Dupont, W.D., and Plummer, W.D., Jr. Power and sample size calculations: A review and computer program. *Controlled Clin Trials* 11:116-128, 1990.

Efron, B. Logistic regression, survival analysis, and the Kaplan-Meier curve. *J Am Stat Assoc* 83:414-425, 1988.

Ehrenberg, A.S.C. A *Primer in Data Reduction.* New York: Wiley, 1982.

Farrington, D.P.; Gallagher, B.; Morley, L.; St. Ledger, R.J.; and West, D.J. Minimizing attrition in longitudinal research: Methods of tracing and securing cooperation in a 24-year follow-up study. In: Bergman, L.R., and Magnusson, D., eds. *Data Quality in Longitudinal Research.* New York: Cambridge University Press, 1990. pp. 122-147.

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. pp. 81-109.

Flinn, C.J., and Heckman, J.J. New methods for analyzing individual event histories. In: Leinhardt, S., ed. *Sociological Methodology.* San Francisco: Jossey-Bass, 1982. pp. 99-140.

Frank, E.; Prien, R.F.; Jarrett, R.B.; Keller, M.B.; Kupfer; D.J.; Lavori, P.W.; Rush, A.J.; and Weissman, M.M. Conceptualization and rationale for consensus definition of terms in major depressive disorder. *Arch Gen Psychiatry* 48:851-855, 1991.

Freedman, L.S. Tables of the number of patients required in clinical trials using the logrank test. *Stat Med* 1:121-129, 1982.

Glasgow, R.E.; Klesges, R.C.; Klesges, L.M.; and Somes, G.R. Variables associated with participation and outcome in a worksite smoking control program. *J Consult Clin Psychol* 56:617-620, 1988.

Goldstein, R.B.; Black, D.W.; Nasrallah, A.; and Winokur, G. The prediction of suicide: Sensitivity, specificity, and predictive value of a multivariate model applied to suicide among 1906 patients with affective disorders. *Arch Gen Psychiatry* 48:418-422, 1991.

Greenhouse, J.B.; Stangl, D.; and Bromberg, J. An introduction to survival analysis methods for analysis of clinical trial data. *J Consult Clin Psychol* 57:536-544, 1989.

Greenhouse, J.B.; Stangl, D.; Kupfer, D.J.; and Prien, R.F. Methodologic issues in maintenance therapy clinical trials. *Arch Gen Psychiatry* 48:313-318, 1991.

Grey, C.; Osborn, E.; and Reznikoff, M. Psychosocial factors in two opiate addiction treatments, *J Clin Psychol* 42:185-189, 1986.

Hall, S.M.; Havassy, B.E.; and Wasserman, D.A. Commitment to abstinence and acute stress in relapse to alcohol, opiates and nicotine. *J Consult Clin Psychol* 58:175-181, 1990.

Hall, S.M.; Havassy, B.E.; and Wasserman, D.A. Effects of commitment to abstinence, positive moods, stress, and coping on relapse to cocaine use. *J Consult Clin Psychol* 59:526-532, 1991.

Hall, S.M.; Rugg, D.; Tunstall, C.; and Jones, R.T. Preventing relapse to cigarette smoking by behavioral skill training. *J Consult Clin Pychol* 52:372-382, 1984.

Hansen, W.B.; Collins, L.M.; Malotte, C.K.; Johnson, C.A.; and Fielding, J.E. Attrition in prevention research. *J Behav Med* 8:261-275, 1985.

Harackiewicz, J.M.; Sansone, C.; Blair, L.W.; Epstein, J.A.; and Manderlink, G. Attributional processes in behavior change and maintenance: Smoking cessation and continued abstinence. *J Consult Clin Psychol* 55:372-378, 1987.

Harrell, F.E., and Goldstein, R. A survey of microcomputer survival analysis software: The need for an integrated framework. *Am Stat*, in press.

Harrell, F.E., and Lee, K.L. Verifying assumptions of the Cox proportional hazards model. In: *SAS Institute, eds. SUGI 11: Proceedings of the 11th Annual SAS Users Group International.* Cary, NC: SAS Institute, 1986.

Havassy, B.E.; Hall, S.M.; and Wasserman, D.A. Social support and relapse: Commonalities among alcoholics, opiate users, and cigarette smokers. *Addict Behav* 16(5):235-246, 1991.

Hawkins, J.D.; Abbott, R.; Catalano, R.F.; and Gillmore, M.R. Assessing effectiveness of drug abuse prevention: Implementation issues relevant to long-term effects and replication. 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. pp. 213-234.

Heckman, J., and Singer, B., eds. *Longitudinal Analysis of Labor Market Data.* New York: Cambridge University Press, 1985.

Hunt, W.A.; Barnett, W.; and Branch, L.G. Relapse rates in addiction programs. *J Clin Psychol* 27:455-456, 1971.

Hunt, W.A., and Bespalec, D.A. An evaluation of current methods of modifying smoking behavior. *J Clin Psychol* 30:431-438, 1974*a*.

Hunt, W.A., and Bespalec, D.A. Relapse rates after treatment for heroin addiction. *J Community Psychol* 2:85-87, 1974*b*.

Hunt, W.A., and General, W.R. Relapse rates after treatment for alcoholism. *J Community Psychol* 1:66-68, 1973.

Hunt, W.A., and Matarazzo, J.D. Habit mechanisms in smoking. In: Hunt, W.A., ed. *Learning Mechanisms in Smoking*. Chicago: Aldine, 1970.

Hutchison, D. Event history and survival analysis in the social sciences. I. Background and introduction. *Qual Quant* 22:203-219, 1988*a*.

Hutchison, D. Event history and survival analysis in the social sciences. II. Advanced applications and recent developments. *Qual Quant* 22:255-278, 1988*b*.

Johnston, L.D. Contributions of drug epidemiology to the field of drug abuse prevention. 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. pp. 213-234.

Johnston, L.D.; O'Malley, P.M.; and Bachman, J.G. *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.

Kalbfleisch, J.D., and Prentice, R.L. *The Statistical Analysis of Failure Time Data.* New York: Wiley, 1980.

Kandel, D.B., and Logan, J.A. Patterns of drug use from adolescence to young adulthood: I. Periods of risk for initiation, continued use, and discontinuation. *Am J Public Health* 74(7):660-666, 1984.

Kandel, D.B., and Yamaguchi, K. Job mobility and drug use: An event history analysis. *Am J Sociol* 92:836-878, 1987.

Kaplan, E.L., and Meier, P. Non-parametric estimation from incomplete observations. *J Am Stat Assoc* 53:457-481, 1958.

Kazdin, A.E., and Bass, D. Power to detect differences between alternative treatments in comparative psychotherapy outcome research. *J Consult Clin Psychol* 57:138-147, 1989.

Kleinbaum, D.G.; Kupper, L.L.; and Morgenstern, H. *Epidemiologic Research: Principles and Quantitative Methods.* Belmont, CA: Lifetime Learning Publications, 1982.

Klerman, G.L. Long-term maintenance of affective disorders. In: Lipton, M.A.; DiMascio, A.: and Killam, K., eds. *Psychopharmacology: A Generation of Progress.* New York: Raven Press, 1978. pp. 1303-1311.

Kraemer, H.C., and Pruyn, J.P. The evaluation of different approaches to randomized clinical trials. *Arch Gen Psychiatry* 47:1163-1169, 1990.

Kraemer, H.C., and Theimann, S. *How Many Subjects?* Beverly Hills, CA: Sage Publications, 1988.

Kupfer, D.J.; Frank, E.; and Perel, J.M. The advantage of early treatment intervention in recurrent depression. *Arch Gen Psychiatry* 46:771-775, 1989.

Lachin, J.M. Introduction to sample size determination and power analysis for clinical trials. *Controlled Clin Trials* 2:93-113, 1981.

Lachin, J.M., and Foulkes, M.A. Evaluation of sample size and power for analyses of survival with allowance for nonuniform patient entry, losses to follow-up, noncompliance, and stratification. *Biometrics* 42:507-519, 1986.

Laird, N., and Olivier, D. Covariance analysis of censored survival data using log-linear analysis techniques. *J Am Stat Assoc* 76:231-240, 1981.

Lancaster, T. *Econometric Analysis of Transition Data.* New York: Cambridge University Press, 1990.

Lawless, J.F. *Statistical Models and Methods for Lifetime Data.* New York: Wiley, 1982.

Lee, E.T. *Statistical Methods for Survival Data Analysis.* Belmont, CA: Lifetime Learning Publications, 1980.

Lelliott, P.; Marks, I.; McNamee, G.; and Tobena, A. Onset of panic disorder with agoraphobia. *Arch Gen Psychiatry* 46:1000-1004, 1989.

Lessler, J.; Tourangeau, R.; and Salter, W. *Questionnaire Design in the Cognitive Research Laboratory: Results of an Experimental Prototype.* National Center for Health Statistics, Vital and Health Statistics, Series 6, No. 1. Washington, DC: Supt. of Docs., U.S. Govt. Print. Off., 1989.

Light, R.J.; Singer, J.D.; and Willett, J.B. *By Design.* Cambridge, MA: Harvard University Press, 1990.

Lilienfeld, A.M., and Lilienfeld, D.E. *Foundations of Epidemiology.* 2d ed. New York: Oxford University Press, 1980.

Litman, G.K.; Eiser, J.R.; and Taylor, C. Dependence, relapse and extinction: A theoretical critique and a behavioral examination. *J Clin Psychol* 35:192-199, 1979.

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

Makuch, R.W., and Simon, R.M. Sample size requirements for comparing time-to-failure among k treatment groups. *J Chronic Dis* 35:861-867, 1982.

Marlatt, G.A.; Curry, S.; and Gordon, J.R. A longitudinal analysis of unaided smoking cessation. *J Consult Clin Psychol* 55:715-720, 1988.

Mausner, J.S., and Bahn, A.K. *Epidemiology.* Philadelphia: W.B. Saunders, 1974.

Mayer, K.U., and Tuma, N.B., eds. *Event History Analysis in Life Course Research.* Madison, WI: University of Wisconsin Press, 1990.

McFall, R.M. Smoking-cessation research. *J Consult Clin Psychol* 46:703-712, 1978.

Means, B.; Nigam, A.; Zarrow, M.; Loftus, E.F.; and Donaldson, M.S. *Autobiographical Memory for Health-Related Events: Enhanced Memory for Recurring Incidents.* National Center for Health Statistics, Vital and Health Statistics, Series 6, No. 2. Washington, DC: Supt. of Docs., U.S. Govt. Print. Off., 1989.

Micceri, T. The unicorn, the normal curve, and other improbable creatures. *Psychol Bull* 103:156-166, 1989.

Miller, R.G. *Survival Analysis.* New York: Wiley, 1981.

Milner, J.L. An ecological perspective on duration of foster care. *Child Welfare* 66:113-123, 1987.

Monroe, S.M.; Simons, A.D.; and Thase, M.E. Onset of depression and time-to-treatment entry: Roles of life stress. *J Consult Clin Psychol* 59:566-573, 1991.

Morgan, G.D.; Ashenberg, Z.S.; and Fisher, E.B., Jr. Abstinence from smoking and the social environment. *J Consult Clin Psychol* 56:298-301, 1988.

Mosteller, F., and Tukey, J.W. *Data Analysis and Regression.* Reading, MA: Addison-Wesley, 1977.

Moussa, M.A.A. Planning the size of survival time clinical trials with allowance for stratification. *Stat Med* 7:559-569, 1988.

Murnane, R.J.; Singer, J.D.; Willett, J.B.; Kemple, J.J.; and Olsen R.J. *Who Will Teach?: Policies That Matter.* Cambridge, MA: Harvard University Press, 1991.

Murphy, G.E., and Wetzel, R.D. The lifetime risk of suicide in alcoholism. *Arch Gen Psychiatry* 47:383-392, 1990.

Murphy, M. Minimizing attrition in longitudinal studies: Means or end? In: Magnusson, D., and Bergman, L.R., eds. *Data Quality in Longitudinal Research.* New York: Cambridge University Press, 1990. pp. 122-147.

Nathan, P.E., and Lansky, O. Common methodological problems in research on the addictions. *J Consult Clin Psychol* 46:713-726, 1978.

Nathan, P.E., and Skinstad, A. Outcomes of treatment for alcohol problems: Current methods, problems, and results. *J Consult Clin Psychol* 55:332-340, 1987.

Neter, J., and Waksberg, J. A study of response errors in expenditures data from household interviews. *J Am Stat Assoc* 59:18-55, 1964.

Petersen, T. The statistical analysis of event histories. *Sociol Meth Res* 19:270-323, 1991.

Peto, R.; Pike, M.C.; Armitage, P.; Breslow, N.E.; Cox, D.R.; Howard, S.V.; Mantel, N.; McPherson, K.; Peto, J.; and Smith, P.G. Design and analysis of randomized clinical trials requiring prolonged observation of each patient: Introduction and Design. *Br J Cancer* 34:585-612, 1976.

Rosenbaum, E., and Kandel, D.B. Early onset of adolescent sexual behavior and drug involvement. *J Marriage Fam* 52:783-798, 1990.

Rubinstein, L.V.; Gail, M.H.; and Santner, T.J. Planning the duration of a comparative clinical trial with loss to follow-up and a period of continued observation. *J Chronic Dis* 34:469-479, 1981.

Schoenfeld, D.A., and Richter, J.R. Monograms for calculating the number of patients needed for a clinical trial with survival as an endpoint. *Biometrics* 38:163-170, 1982.

Seltzer, M.M.; Seltzer, G.B.; and Sherwood, C.C. Comparison of community adjustment of older vs. younger mentally retarded adults. *Am J Ment Defic* 87:9-13, 1982.

Shiffman, S. Relapse following smoking cessation: A situational analysis. *J Consult Clin Psychol* 50:71-86, 1982.

Singer, J.D. Are special educators' career paths special?: Results from a 13-year longitudinal study. *Except Child* 59:262-279, 1993*a*.

Singer, J.D. Once is not enough: Former special educators who return to teaching. *Except Child* 60(1):1993*b*.

Singer, J.D., and Willett, J.B. Modeling the days of our lives: Using survival analysis when designing and analyzing studies of duration and the timing of events. *Psychol Bull* 110(2):268-290, 1991.

Singer, J.D., and Willett, J.B. It's about time: Using discrete-time survival analysis to study duration and the timing of events. *J Educ Stat* 18:155-195, 1993.

Stevens, V.J., and Hollis, J.F. Preventing smoking relapse using an individually tailored skills training technique. *J Consult Clin Psychol* 57:420-424, 1989.

Sudman, S., and Bradburn, N. *Response Effects in Surveys: A Review and Synthesis*. Chicago: Aldine, 1974.

Sudman, S., and Bradburn, N. *Asking Questions: A Practical Guide to Questionnaire Design*. San Francisco: Jossey-Bass, 1982.

Sutton, S.R. Interpreting relapse curves. *J Consult Clin Psychol* 47:96-98, 1979.

Trussel, J., and Hammerslough, C. A hazards-model analysis of the covariates of infant and child mortality in Sri Lanka. *Demography* 20:1-26, 1983.

Tuma, N.B., and Hannan, M.T. *Social Dynamics: Models and Methods*. New York: Academic Press, 1984.

Turnbull, B.W. Non-parametric estimation of a survivorship function with doubly censored data. *J Am Stat Assoc* 69:169-173, 1974.

Turnbull, B.W. The empirical distribution function with arbitrarily grouped, censored and truncated data. *J R Stat Soc [B]* 38:290-295, 1976.

Velicer, W.F.; Proschaska, J.O.; Rossi, J.S.; and Snow, M.G. Assessing outcome in smoking cessation studies. *Psychol Bull* 111:23-41, 1992.

Verhulst, F.C., and Koot, H.M. Longitudinal research in child and adolescent psychiatry. *J Am Acad Child Adolesc Psychiatry* 30(3):361-368, 1991.

Willett, J.B., and Singer, J.D. From whether to when: New methods for studying student dropout and teacher attrition. *Rev Educ Res* 61(4):407-450, 1991.

Willett, J.B., and Singer, J.D. Investigating onset, cessation, relapse and recovery: Why you should, and how you can, use discrete-time survival analysis to examine event occurrence. *J Consult Clin Psychol* 61:952-965, 1993.

Wittchen, H.; Burke, J.D.; Semler, G.; Pfister, H.; Von Cranach, M.; and Zaudig, M. Recall and dating of psychiatric symptoms: Test-retest reliability of time-related symptom questions in a standardized psychiatric interview. *Arch Gen Psychiatry* 46:437-443, 1989.

Yamaguchi, K., and Kandel, D.B. Patterns of drug use from adolescence to young adulthood: Predictors of progression. *Am J Public Health* 74(7):673-681, 1984.

Yamaguchi, K., and Kandel, D.B. Drug use and other determinants of premarital p›egnancy and its outcome: A dynamic analysis of competing life events. *J Marriage Fam* 49:257-270, 1987.

Young, M.A.; Watel, L.G.; Lahmeyer, H.W.; and Eastman, C.I. The temporal onset of individual symptoms in winter depression: Differentiating underlying mechanisms. *J Affect Disord* 22:191-197, 1991.

Zatz, M.S. Los Cholos: Legal processing of Chicano gang members. *Soc Probl* 33:13-30, 1985.

AUTHORS

Judith D. Singer, Ph.D.
Professor

John B. Willett, Ph.D.
Professor

Graduate School of Education
and
National Center for Educational Statistics
Harvard University
Appian Way
Cambridge, MA 02138

Time Series Models of Individual Substance Abusers

Wayne F. Velicer

ABSTRACT

Time series analysis is a statistical procedure appropriate for repeated observations on a single subject or unit. The goal of the analysis may be to determine the nature of the process that describes an observed behavior or to evaluate the effects of a treatment or intervention. *Model identification* involves specifying which of several alternative Autoregressive Integrated Moving Average (ARIMA) models best describes the series and may be used to investigate basic processes. This is illustrated by an example involving selecting the model of nicotine regulation that best represents smokers. *Intervention analysis* involves determining if there are any changes in level or direction for the series as a result of the intervention. Two types of applications have potential for the substance abuse area: (1) evaluation of the effects of an intervention on a single individual, and (2) evaluation of organizational-level changes (i.e., program evaluation). This is illustrated by an example that examines the effect of relaxation therapy on blood pressure. *Pooled time series procedures* are employed to combine the data from several different individuals or units, either by cross-sectional analysis or meta-analysis. In addition, several other issues are discussed that are critical to performing a time series analysis: selection of an appropriate computer program, alternative procedures for handling missing data, procedures for multiple observations at each occasion, and corrections for seasonal data.

INTRODUCTION

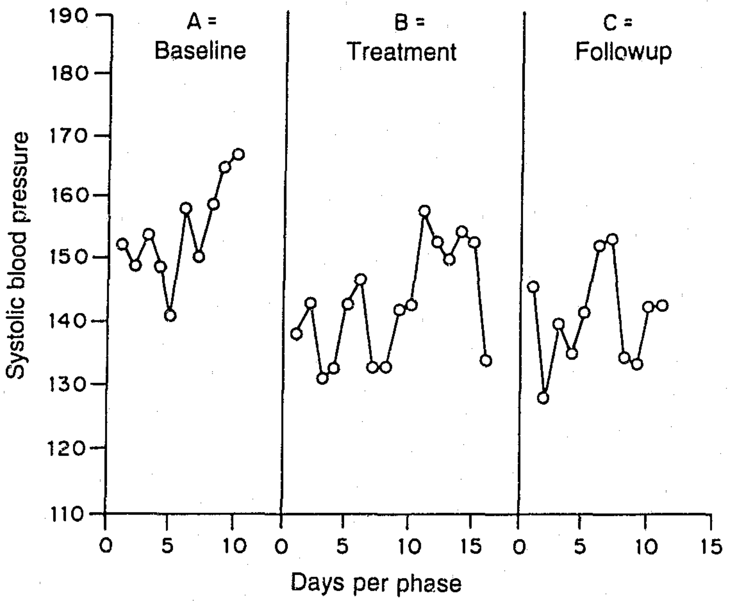
Time series analysis involves repeated observations on a single unit (often a single subject) over time. In the area of prevention and treatment of substance abuse, the analysis of interest usually is an *interrupted time series analysis.* The interruption corresponds to the occurrence of an intervention, and the goal is to evaluate its effect. Traditional between-groups statistical procedures cannot be employed because repeated observations on the same unit cannot be assumed to be independent. The presence of dependency may substantially bias a statistical test that does not take it into account. The direction of the bias will depend on the direction of the dependency. The most widely employed methods of analysis for time series designs are based on the Autoregressive Integrated Moving Average (ARIMA) models (Box and Jenkins 1976; Box and Tiao 1965). These procedures permit the effects of dependency to be statistically removed from the data (Glass et al. 1975; Gottman 1973; Gottman and Glass 1978).

Time series analysis has generated widespread interest for a number of reasons. First, time series are applicable particularly to the study of problems in applied settings where more traditional between-subject designs are impossible or very difficult to implement and may not accurately reflect the situations involved. Many prevention and treatment programs for substance abuse occur in school or clinical settings. Second, time series designs are appropriate particularly for dealing with questions of causality because of the temporal occurrence of both the intervention and effect of the intervention. Third, time series designs possess the additional advantage of permitting study of the pattern of intervention effects (i.e., temporary effects versus permanent effects, changes in slope as well as change in level) over and above the usual question of the existence of a mean treatment effect. The study of substance abuse and the prevention and treatment of substance abuse provides many situations where time series designs are the optimal choice.

The employment of time series methods also suffers from several drawbacks. First, generalizability cannot be inferred from a single study, only through systematic replication. Second, traditional measures may be inappropriate for time series designs; measures are required that can be repeated a large number of times on a single subject, usually at short intervals. Third, a large number of observations is required for accurate model identification. Model identification is a necessary step in order to remove the dependency present in the data. Advances in methods of analysis in the last decade have provided partial solutions to the generalizability issues and the sample size issues.

To illustrate the use of time series analysis, consider two examples. In the first example, the effects of assertion training and muscle relaxation therapy on blood pressure (hypertension) were studied (Printz 1978). Figure 1 presents the results for a single subject. The baseline phase (A) involved a series of regular (3 days/week) observations of the subject's blood pressure. After the 10th observation, the treatment phase (B) started, which involved training in assertiveness and relaxation therapy, and 16 more observations were taken. The followup phase (C) refers to the end of active assertiveness training and relaxation therapy training; only 11 regular measurements occurred. However, the subject was expected to continue to employ assertiveness and relaxation techniques on his or her own.

The analysis estimates two parameters for each phase: level and slope. Conceptually, a straight line is fitted to the data, with the level referring to the intercept of the line and the slope referring the rate of increase or decrease of the line. The slope refers to the rate of increase or decrease of the series over time. A slope near 0.0 is common and would be presented graphically as a nonincreasing line parallel to the time axis. In the case of a near-zero slope, the level also can be interpreted as the mean. During the A phase, the level of the series is 145.02, and the slope is increasing. The introduction of the relaxation therapy results in a decrease of 27.09 in the level of the series (in the figure, Δ Level = change in level) and a decrease in the slope of the series (in the figure, Δ Slope = change in slope). Both changes are significant and represent a positive outcome for relaxation therapy. During the C phase, there is a further (nonsignificant) decline in the level of the series and an additional decrease in the slope of the series, which was significant. This indicates that the positive effects of the relaxation therapy were maintained after the end of the intervention.



Phase	Level	Δ Level	Slope	Δ Slope
A-B	145.02	-27.09*	1.17*	-0.79
B-C	134.92	-10.93*	0.95*	-0.62

FIGURE 1. *Example of an interrupted time series analysis*

This study illustrates several of the strengths and weaknesses of time series analysis. First, the study involved eight different subjects (seven in addition to the one illustrated here), each treated as part of a therapist's regular practice over a period of approximately 1 year, each during a different timeframe. This design illustrated how time series can be incorporated into an applied setting. Second, the abrupt change in the level of the series that occurred at the same time the intervention started permits a strong causal inference about the relation between intervention and the outcome. Third, the change in slope provides information about the nature of the intervention. The drawbacks of time series analysis also are illustrated by this study. The issue of generalizability was addressed by employing multiple subjects to replicate the effect. In this case, the treatment was effective in three of the cases. A potential explanation was that the treatment was effective only when the onset of hypertension was rather recent in origin and less effective when the problem was of long duration. The measure in this case was appropriate for repeated observations. The length of the series was too short to permit model identification. This study followed the Simonton (1977) approach (see below) and assumes that a single model was the appropriate model for all subjects.

The most widely used interrupted time series procedure is described by Glass and colleagues (1975), Gottman (1973), and Gottman and Glass (1978), following the approach of Box and Jenkins (1976) and Box and Tiao (1965). It involves a two-step process: First, the researcher identifies which of a family of ARIMA (p, d, q) models is appropriate for the data; then the researcher employs a specific transformation appropriate to the identified model to transform the dependent observed variable (Z_t) into a serially independent variable (Y_t). Intervention effects then can be evaluated by a generalized least squares estimate of the model parameters. This procedure suffers from a number of drawbacks, including: (1) the requirement of a large number of data points for accurate model identification; (2) excessive mathematical complexity; and (3) problems with accurately and reliably performing the model identification task, even when the recommended minimum number of observations are obtained (Velicer and Harrop 1983). Alternative procedures that avoid model identification have been proposed (Algina and Swaminathan 1977, 1979; Simonton 1977; Swaminathan and Algina 1977; Velicer and McDonald 1984, 1991).

A key concept for time series analysis is dependence. This is assessed by calculating the *autocorrelations* of various *lags*. A typical correlation coefficient estimates the relation between two variables measured at the same time. An autocorrelation estimates the relation between the same variable measured on two occasions. For example, if researchers have a series of observations and pair the second observation with the first, the third observation with the second, and so on until the last observation is paired with the second from the last observation, and they then calculate the correlation between the paired observations, the lag 1 autocorrelation has been calculated. If the third is paired with the first and each subsequent observation with the observation two occasions behind, the lag 2 autocorrelation is calculated. The lag of an autocorrelation refers to how far in the past. Typically, autocorrelations are between 1.00 and -1.00. In the behavioral sciences, the size of the autocorrelation typically will decrease as the lag increases. The exception is seasonal data. The pattern of the autocorrelation and the related *partial auto-correlations* (which will not be defined here) are employed as the basis for identifying the specific ARIMA model. A *white noise model* is one where there is no dependency in the data; i.e., the autocorrelations and partial autocorrelations for all lags are 0.

In this chapter, model identification and intervention analysis are treated separately. Model identification can be the goal of a study. The first section will discuss the problems of model identification and some of the recent solutions to those problems and will present an example of the use of time series model identification to the problem of theory-testing in the addictive behavior area. The second section will review the analysis of interrupted time series data, which is appropriate when an intervention is present. The Box-Jenkins approach (Box and Jenkins 1976) will be described in detail. Several alternative approaches also will be reviewed with an emphasis on procedures that bypass the model identification step. In contrast to the first section, this section will assume that model identification is not a primary goal of the study. The third section will describe procedures for generalization, including testing effects across multiple units (subjects) and meta-analysis procedures. The last section will review some specific problem areas for time series analysis: cyclic data, missing data, software available for the analysis, and multivariate procedures.

TIME SERIES MODEL IDENTIFICATION: GENERAL ISSUES

Model identification can be the goal of a time series analysis. Determining the specific model can identify a basic process. However, model identification is a difficult and problematic procedure. In interrupted time series analysis, model identification often represents a first step, preliminary to the goal of the analysis, which is the estimating and testing of the preintervention and postintervention parameters (Box and Jenkins 1976; Box and Tiao 1965, 1975; Glass et al. 1975; McCleary and Hay 1980; Velicer and McDonald 1984, 1991). A variety of procedures have been developed to identify the model (Akaike 1974; Beguin et al. 1980; Bhansali and Downham 1977; Glass et al. 1975; Grey et al. 1978; Hannan and Rissanen 1982; Kashyap 1977; McCleary and Hay 1980; Parzen 1974; Pukkila 1982; Rissanen 1978, 1986*a*, 1986*b*; Schwartz 1978; Tsay 1984; Tsay and Tiao 1984). However, model identification has been problematic because of the large number of data points required for accurate identification, the complexity of the procedures, and problems with accuracy and reliability, even under ideal circumstances (Velicer and Harrop 1983). This section will illustrate the use of model identification to answer a substantive question and illustrate the procedures and inherent problems in model identification.

Definition of Model Identification

The ARIMA (p, d, q) model represents a family of models with the parameters designating which specific model is involved. The first parameter (p) is the order of the autoregressive parameter, and the last parameter (q) is the order of the moving average parameter. The middle parameter (d) represents the presence of instability or stochastic drift in the series. Each of the parameters of the model may be of order 0, 1, 2, 3, or more, although higher-order models are unusual in the behavioral sciences (Glass et al. 1975). A parameter equal to 0 indicates the absence of that term from the model.

Model identification involves a number of aspects that can be determined with varying degrees of accuracy. *Selection of the model* involves determining which specific model from the ARIMA (p, d, q) family of models most parsimoniously describes the data. This is a difficult task to accomplish accurately because the different models, under certain conditions, can appear very similar. For example, a first-order moving average model is identical to an autoregressive model of high order.

Order refers to how many preceding observations must be considered in order to account for the dependency in the series. Accuracy is difficult because higher-order autocorrelation terms typically are closer to 0 than first-order terms and, therefore, are more likely to be included within the bounds for any error estimate. Order reflects how far into the past one must go to predict the present observation.

Degree of dependency refers to how large the autocorrelations are on a scale from 0.0 to 1.0. As with other dependency indicators, this can be interpreted as the strength of relationship between consecutive measurements. The accuracy of estimation is largely a function of the number of observations with numbers of observations over 100 providing reasonably accurate estimates (Box and Pierce 1970; Glass et al. 1975; Ljung and Box 1978). The degree of dependency indicates the extent to which an observation at any point in time is predictable from one or more preceding observations. For example, if data were collected every 12 hours, then finding an order 1 model would suggest that the previous observation (t-1 = 12 hours ago) was more important than the second previous observation (t-2 = 24 hours ago) in predicting the level of the series at time t.

Direction of dependency refers to whether the autocorrelation is positive or negative. This can be determined with a high degree of accuracy when the dependency clearly is nonzero. The direction is of less interest as the degree of dependency approaches 0. The direction of dependency has clear implications. If the sign of the autocorrelation is negative, a high level for the series on one occasion will predict a low level for the series on the next occasion. If the sign is positive, an above-average level of the series on one occasion will predict a higher-than-average level on the next occasion.

Illustrations of Alternative Time Series

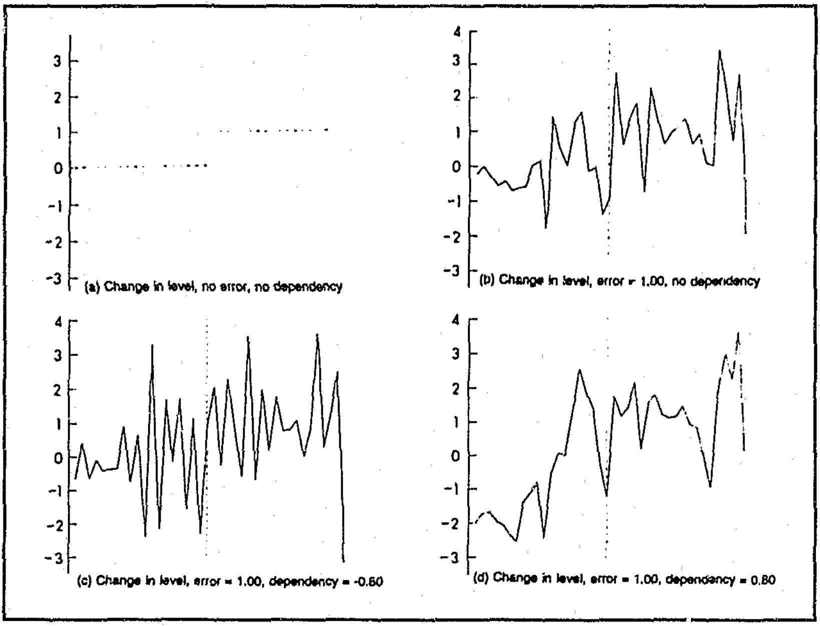
Figure 2 illustrates four different types of models with computer-generated data ($N_1 = N_2 = 20$) for an ARIMA (1, 0, 0) model. Graph (a) represents an ideal interrupted time series example with no error and an immediate change in the level of one unit at the time of intervention. Graph (b) is the same model with the same change in level but with a random-error component added. The variance or the random error is 1.00. There is no autocorrelation in this model. Graph (c) is a model with the same change in level and error variance but with a large negative autocorrelation (-.80). Graph (d) is a model with the same change in level and error variance as (b) but with a large positive autocorrelation (+.80). The impact of dependency can be observed easily. The negative dependency results in an exaggerated "sawtooth" graph with increased apparent variability. The positive dependency results in a smoother graph with decreased apparent variability. The inclusion of an intervention effect (the change in level) illustrates how difficult it is to determine by visual inspection alone if an intervention had an effect.

Example

To illustrate the use of model identification in theory-testing, the author will present briefly the results of a recent study (Velicer et al. 1992*a*) designed to determine which of three models of nicotine regulation best represented most smokers. These models seek to explain the mechanism that determines how smokers increase or decrease their level of smoking in order to maintain a certain level of nicotine in their systems. Three measures were employed in the study but only one, number of cigarettes, is described here.

Nicotine Regulation Models. Three alternative models have been employed to account for nicotine's effectiveness in maintaining smoking: (1) the nicotine fixed effect model, (2) the nicotine regulation model, and (3) the multiple regulation model. Leventhal and Cleary (1980) provide a review of the literature and a description of each of the three models. Each of the three models is identified with one of three broad classes of time series models: (1) a positive dependency model, (2) a white noise model (no dependency), and (3) a negative dependency model.

The *nicotine fixed effect model* assumes that smoking is reinforced because nicotine stimulates specific reward-inducing centers of the



(a) Change in level, no error, no dependency

(b) Change in level, error = 1.00, no dependency

(c) Change in level, error = 1.00, dependency = -0.60

(d) Change in level, error = 1.00, dependency = 0.80

FIGURE 2. *Illustrations of four time series using computer-generated data for ARIMA (1, 0, 0) models*

SOURCE: Reprinted from *Addictive Behaviors*, 17; W.F. Velicer, C.A. Redding, R.L. Richmond, J. Greeley, and W. Swift; A time series investigation of three nicotine regulation models, 325-345; Copyright (1992), with kind permission from Elsevier Science, Ltd., The Boulevard, Langford Lane, Kidlington OX5 1GB, UK

nervous system. These have been identified as either autonomic arousal or feeling of mental alertness and relaxation or both. There is not a large body of evidence or a good formal statement available for this model. Following this model, an increase on one occasion should be followed by an increase on the next occasion, or a decrease on one occasion should be followed by decreased consumption on a subsequent occasion if the same level of arousal is to be maintained. In time series model terms, this would result in a positive autocorrelation.

The *nicotine regulation model* assumes that smoking serves to regulate or titrate the smoker's level of nicotine. Departures from the optimal level, or *set point,* will stimulate an increase or decrease in smoking to return to this optimal nicotine level. Jarvik (1973) presents a review of a large body of evidence which supports this model (also see Schachter [1977] and Russell [1977]). The model suggests that any increase or decrease in smoking caused by events in a person's environment should be temporary. The person should return immediately to their personal set point when the environment permits. This would result in a white noise model with an autocorrelation of 0.

The *multiple regulation model* represents a more complex model designed to overcome some of the problems of the nicotine regulation model—specifically, how the nicotine set point develops and how deviations from the set point generate a craving for cigarettes. Leventhal and Cleary (1980) summarize some of the evidence the nicotine regulation model cannot adequately account for, and they suggest the multiple regulation model as an alternative. This model is an elaboration of similar models by Solomon and Corbit (1973, 1974) and Tomkins (1966, 1968); also see Solomon (1980). This model assumes that the smoker is regulating emotional states. Drops in nicotine level stimulate craving. One way to link craving to nicotine level is the opponent-process theory (Solomon 1980; Solomon and Corbit 1973, 1974), which posits that nicotine gives rise to an initial positive-affect reaction that is followed automatically by a slave opponent negative-affect reaction. The opponent state becomes stronger with repeated activation and can be eliminated by reinstating the initial positive state. External stimulus provides an alternative source for craving. The theory would predict that an increase (or decrease) in smoking rate caused by events in a person's environment should be followed by an opposite decrease (or increase) in smoking rate. This would result in a negative autocorrelation at lag 1 and alternating positive and negative autocorrelations at subsequent lags.

As an analogy, view each model as positing a predetermined level for each smoker. The environment (both internal and external) produces a "shock" to the system, causing nicotine intake to exceed or fall below the predetermined level. The three models differ on the strength of the forces that return the smoker to his or her level. Researchers can think of this as a physiological or psychological "rubberband." The nicotine fixed effect model proposes a weak rubberband so that some of the shock remains in the system at the next observation. This would result in a positive dependency. The nicotine regulation model assumes that the rubberband is perfectly accurate, returning the system to its original level at the next observation. This would result in a white noise (or zero dependency) model. The multiple regulation model proposes a very strong rubberband that carries the system past the level in the opposite direction on the next observation. The system would oscillate around the individual's set point, slowly damping down to that level. This would result in a negative dependency model.

Subjects. In order to achieve stable autocorrelations, time series analysis requires a minimum of 100 data points (Box and Jenkins 1976; Glass et al. 1975). The study (Velicer et al. 1992*a*) employed 10 smokers (4 male and 6 female), from whom measures were collected twice daily for two months (62 days). Deletion (i.e., deletion of the missing observation and closing up the series) and mean value were both used for missing data.

Measure: Number of Cigarettes. Having subjects monitor their own smoking behavior is one of the most commonly employed measures in smoking research (McFall 1978; Velicer et al. 1992*b*). This is an inexpensive and convenient means of gathering data. The accuracy and reliability of data gathered through self-monitoring are not always as high as that of data gathered through other techniques. However, the advantages of using self-monitoring typically outweigh the disadvantages. Heatherton and colleagues (1989) have found the number of cigarettes smoked per day to be a valuable index of heaviness of smoking (also see Velicer et al. 1992*b*).

Model Identification Procedures

Model identification involves determining if autoregressive terms or moving average terms must be included to describe the data fully. The distribution of the autocorrelations and partial autocorrelation provides the basis for making such decisions. For an autoregressive component, the autocorrelations will decay slowly to 0 for increasing lags, and the partial autocorrelations will drop abruptly to 0 when the appropriate lag (p) is reached. For the moving averages component, the autocorrelations will drop abruptly to 0 when the appropriate lag (p) is reached, and the partial autocorrelations will drop slowly to 0. Model identification in this study was restricted to autoregressive models only, a procedure consistent with current practice (Djuric and Kay 1992; Gottman 1981; Velicer and McDonald 1984, 1991). Diagnostic checks on the residuals were performed to test the appropriateness of this procedure. A third component, drift, was set equal to 0 a priori for all identification problems based on a preliminary evaluation of the data. Models that demonstrate no dependence are called white noise models and are described as ARIMA (0, 0, 0) models.

Five different procedures were employed for model identification. First, traditional visual analysis of the autocorrelations and partial autocorrelations was performed. The visual analysis required the consensus of three raters. Then four different automated methods for order identification of autoregressive models were employed: (1) predictive minimum descriptive length (Rissanen 1986*a*); (2) predictive least squares (Rissanen 1986*b*); (3) predictive least absolute value (Djuric and Kay 1992); and (4) predictive density criterion (Djuric and Kay 1992). Two additional methods were considered and rejected: (1) Akaike information criterion (AIC) (Akaike 1974), and (2) minimum descriptive length (MDL) (Rissanen 1978; Schwartz 1978). A recent simulation study evaluating these six criteria (Djuric and Kay 1992) found that AIC and MDL tended to overestimate the order of series. In this study, these two criteria were inconsistent with either visual analysis or the other four criteria, typically finding a much higher order, so they were eliminated from consideration.

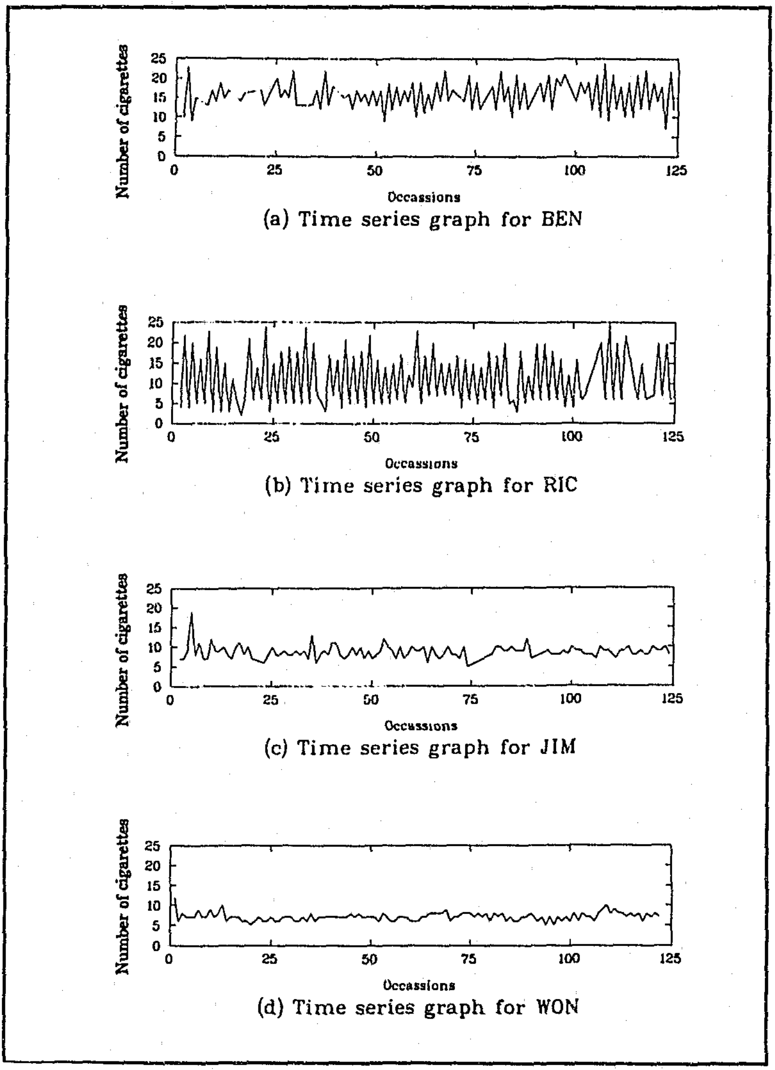
For the majority of model identification, all five procedures converged on the same answer. When disagreement occurred, it typically was a difference of one in order, and all models were reviewed. Disagreements typically involved a low autoregressive coefficient that was approximately equal to the critical value for statistical significance. The more parsimonious fit (lower order) was employed when the evidence for the higher-order model was weak and the inclusion of the additional term would not result in a change in interpretation.

Results

Seven of the subjects were described by a first-order autoregressive model with a high degree of negative dependence (-.30 to -.80). All subjects reported on their smoking behavior in the morning and afternoon. The autocorrelation resulted in a very clear, easily identified model with a high degree of autocorrelation. This pattern is consistent with the multiple regulation model, and the study was interpreted as supporting that model.

Three of the subjects did not show the same pattern. One of the subjects worked some weeks during the day and some weeks at night. This subject also missed a number of sessions and terminated prematurely. One subject was a very controlled smoker, smoking 15 cigarettes at predetermined intervals. All three averaged less than a pack a day. However, two subjects who demonstrated the pattern of high negative dependence also smoked less than a pack a day.

Figure 3 presents the data graphically for four subjects. Two of the subjects (BEN and RIC in panels [a] and [b], respectively) were representative of the seven subjects characterized by a high negative dependence. The exaggerated "sawtooth" shape of this type of time series is clearly observable. Two subjects (JIM and WON in panels [c] and [d], respectively) were representative of the three subjects who demonstrated either a zero or low positive dependence. The time series graphs for these two subjects are much smoother and more regular. The positive dependency subject (WON) produced a smoother pattern than the zero dependency subject (JIM).



(a) Time series graph for BEN

(b) Time series graph for RIC

(c) Time series graph for JIM

(d) Time series graph for WON

FIGURE 3. *Illustrative time series graphs of the number of cigarettes for four subjects*

SOURCE: Reprinted from *Addictive Behaviors*, 17; W.F. Velicer, C.A. Redding, R.L. Richmond, J. Greeley, and W. Swift; A time series investigation of three nicotine regulation models, 325-345; Copyright (1992), with kind permission from Elsevier Science, Ltd., The Boulevard, Langford Lane, Kidlington OX5 1GB, UK

INTERRUPTED TIME SERIES ANALYSIS

The simplest interrupted time series analysis is a design that involves repeated observations on a single unit followed by an intervention that is followed by additional observations of the unit. The purpose of the analysis is to determine if the intervention had an effect. The example presented earlier and figure 1 illustrate this approach. The analysis involves some preprocessing of the data to remove the effects of dependence. Several alternative procedures will be described below. The analysis then involves a general linear model analysis using a generalized least squares or Aitken estimator (Aitken 1934; Morrison 1983). The intervention can be an experimental manipulation, such as a drug or treatment for an addiction, or it can be a naturally occurring event, such as a change in policy or funding for a public program. If the intervention effect is significant, it frequently is of prime interest to evaluate the form of the effect. One of the advantages of time series analysis is the ability to assess the nature of change over time.

The next section will describe the Box-Jenkins procedure (Box and Jenkins 1976). Several variations on this procedure have been proposed to eliminate the problematic model identification step and will be described afterward. Some of the more technical material has been placed in italics so that readers may skip over this material and still follow the presentation.

Box-Jenkins Intervention Analysis

An intervention for the prevention or treatment of substance abuse can be evaluated using a Box-Jenkins analysis. The Box-Jenkins procedure (Box and Jenkins 1976), as adapted by Glass and colleagues (1975), is a two-step process. First, the autocorrelations and partial autocorrelations are calculated for various lags. This information is the basis for identifying the specific ARIMA model (i.e., specifying the value for p, d, and q). Model identification determines the specific transformation matrix to be used. The purpose of this transformation is to remove the dependence from the data so that they can be analyzed by the usual statistical procedures. Second, the data are analyzed with a modified general linear model program, and the parameters are estimated and tested for significance. The general linear model is the general analytic procedure that includes multiple regression, analysis of variance, and analysis of covariance as special cases. After the dependence in the data is accounted for, the analysis follows standard estimation and testing procedures.

A typical problem[1] *would be to determine if the level of the series has changed as a result of the intervention. The analysis will be described without the transformation first. For the simplest analysis, this would involve the estimation of two parameters:* _**L**_*, the level of the series, and* _**D**_*, the change in level after intervention. A test of significance then could be performed on the hypothesis of prime interest, H_o:*_D_ *= 0. This could be expressed in terms of the general linear model as*

$$\mathbf{Z} = X\,\mathbf{b}+\mathbf{a} \qquad (1)$$

where **Z** *is the Nx1 vector of observed variables ($N = n_1+n_2$), where N is the total number of observations with n, occurring before intervention; X is the Nxp design matrix, where p is the number of parameters estimated;* ***b*** *is the px1 vector of parameters; and* ***a*** *is the Nx1 vector of residuals. For this example, the vector of parameters contains two components, namely* _L_ *and* _D_*. The design matrix is presented in panel A in table 1.*

The usual least squares solution is

$$b=(X'X)^{-1}X'Z \qquad (2)$$

and a test of significance for the null hypothesis H_o: $b_i = 0$ (i.e., H_o: _D_ *= 0 is given by*

$$t_{bi}=b_i/s_{bi} \qquad (3)$$

TABLE 1. *Examples of common design matrices (X) for single-unit analysis ($N_1 = N_2 = 5$)*

(a) Immediate and constant changes in level

1	0
1	0
1	0
1	0
1	0
1	1
1	1
1	1
1	1
1	1

(b) Immediate and constant changes in level and slope

1	0	1	0
1	0	2	0
1	0	3	0
1	0	4	0
1	0	5	0
1	1	6	1
1	1	7	2
1	1	8	3
1	1	9	4
1	1	10	5

(c) Delayed change in level

1	0
1	0
1	0
1	0
1	0
1	1.0
1	.5
1	.25
1	.13
1	.07

(d) Delayed change in level

1	0
1	0
1	0
1	0
1	0
1	0
1	0
1	1
1	1
1	1

where

$$s^{2}_{bi}=s^{2}_{a}C^{II} \tag{4}$$

and s^{2}_{u} is the estimate of the error variance and C^{II} is the ith diagonal element of $(X'X)^{-1}$. The test statistic would have a t distribution with degrees of freedom N-p.

Figure 4 illustrates eight different outcomes for a simple one-intervention design. In a typical between-two-groups experimental design, only one assessment occurs after treatment. By inspecting the different patterns of change over time, researchers can see that selecting different points in time for the single assessment would result in very different conclusions for five of the examples (D, E, F, G, and H). The evolutionary effect (D) is a good example of where the intervention results in a temporary negative effect, perhaps while a response pattern is unlearned, followed by a positive effect. An early assessment would conclude that the treatment had a negative effect; a somewhat later assessment would find no treatment effect, while an even later assessment would find a positive treatment effect.

Alternative specifications of the design matrix permit the investigation of different hypotheses concerning the nature of the intervention. Table 1 presents some illustrative examples for an N = 10 ($N_1 = N_2 = 5$) case. Panel (a) is the design matrix for an immediate and constant treatment effect. Panel (b) is the design matrix for testing a change in both level and slope. Panel (c) is the design matrix for a decaying treatment effect. Panel (d) is the design matrix for testing a delayed treatment effect.

Alternative specifications of the design matrix permit the investigation of different hypotheses concerning the nature of the intervention. Table 1 presents some illustrative examples for an N = 10 ($N_1 = N_2 = 5$) case. Panel (a) is the design matrix for an immediate and constant treatment effect. Panel (b) is the design matrix for testing a change in both level and slope. Panel (c) is the design matrix for a decaying treatment effect. Panel (d) is the design matrix for testing a delayed treatment effect.

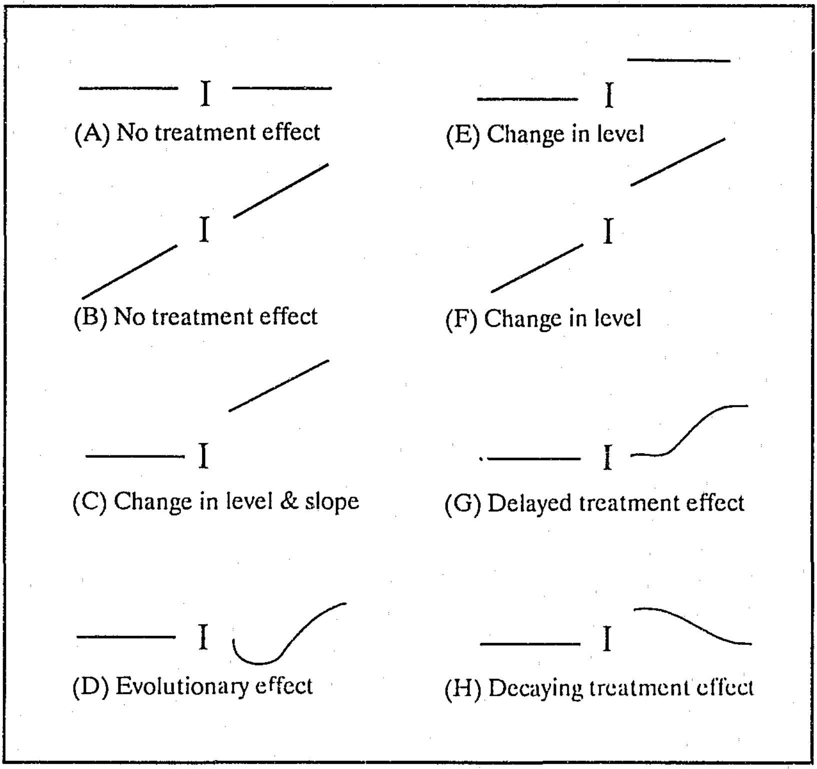


FIGURE 4. *Eight alternative outcomes for a simple intervention design*

The general linear model cannot be applied directly to time series analysis because of the presence of dependency in the residuals. It is necessary to perform a transformation on the observed variable, Z_t, to remove dependency prior to the statistical analysis. A transformation matrix T must be found, yielding

$$Y = TZ \tag{5}$$

and

$$X^* = TX \tag{6}$$

Given T, the estimate of the parameters, b, may be expressed as a generalized least squares problem; that is,

$$b = (X'T'TX)^{-1}X'T'TZ \tag{7}$$

and

$$b = (X^{*\prime}X^{*})^{-1}X^{*\prime}Y \tag{8}$$

The purpose of the model identification step is to determine the appropriate transformation of Z into Y. Table 2 presents six common ARIMA models. After model identification, an estimation procedure is employed to determine the specific numeric values of ϕ and Θ. Appropriate tests of significance are based on asymptotic theory.

The Box-Jenkins approach to intervention analysis suffers from a number of difficulties. First, the number of data points required for model identification often is prohibitive for research in applied settings. Second, even for the required number of points, correct identification is problematic (Velicer and Harrop 1983). Third, the method is complex, making applications by the mathematically unsophisticated researcher difficult. Three alternative approaches are described in the next section, all of which attempt to avoid the problematic model identification step.

Alternative Approaches

Simonton (1977) proposed a procedure that avoids the problem of model identification by using an estimate of the variance-covariance matrix based on a pooling of the observations across all subjects observed. This approach, however, requires a basic assumption. All series are assumed to be (1, 0, 0). While the assumptions seem to be theoretically indefensible, empirical investigations indicate that this procedure works well in a wide variety of cases (Harrop and Velicer 1985).

Algina and Swaminathan (1977, 1979) and Swaminathan and Algina (1977) have proposed an alternative to Simonton's statistical analysis that

TABLE 2. *Common ARIMA models*

Label	(p, d, q)	Descriptive Formula	Comment
White noise	(0, 0, 0)	$Z_t = L+a_t$	No dependency in the data
Autoregressive Order One	(1, 0, 0)	$Z_t-L = \phi(Z_{t-1}-L)+a_t$	Predicted from previous observations
Autoregressive Order Two	(2, 0, 0)	$Z_t-L = \phi(Z_{t-1}-L)+$ $\phi_2(Z_{t-2}-L)+a_t$	Predicted from previous two observations
Moving averages Order One	(0, 0, 1)	$Z_t-L = a_t-\theta_1 a_{t-1}$	Proportion of previous shock affects observation
Moving averages Order Two	(0, 0, 2)	$Z_t-L = a_t-\phi_1 a_{t-1}-\phi_2 a_{t-2}$	Proportion of two previous shocks affecting observations
Integrated average	(0, 1, 1)	$Z_t-Z_{t-1} = a_t-\phi_1 a_{t-1}$	Stochastic drift and proportion of previous shock affect observation

employs a profile analysis. The sample variance-covariance matrix is employed as an estimator for $T'T$ in the modified least squares solution (see equation [7]). This approach, however, requires the assumption that the number of subjects is greater than the number of observations per subject. This is not a condition that is likely to be met in most applied research settings, where time series approaches are most appropriate.

The transformation of the observed variable, Z, is required because the observed data contain dependence and, therefore, do not meet the requirements of the general linear model. All transformation matrices, **T**, have an identical form—a lower triangular matrix with equal subdiagonals. Instead of trying to determine the specific matrix, Velicer and McDonald

(1984) propose a general transformation matrix with the numerical values of the elements of **T** being estimated for each problem. Weight vectors with five nonzero weights are accurate for most cases. A greater number of weights can be employed where indicated by appropriate diagnostics (Velicer and McDonald 1984). The accuracy of this approach has been supported by two simulation studies (Harrop and Velicer 1985, 1990*b*).

TIME SERIES ANALYSIS FOR MULTIPLE UNITS

One of the issues involved in time series analysis is generalizability. How can the results from a single individual be generalized to a larger population? Hersen and Barlow (1976) discuss the problems in terms of systematic replication. The example discussed previously involving the impact of relaxation therapy on blood pressure employed this approach.

However, this procedure involves logical inference rather than formal statistical inference. Two approaches have been developed for statistical inference on multiple units: pooled time series designs and meta-analysis.

The next section will describe an approach to pooled time series analysis that recently was proposed by Velicer and McDonald (1991). This approach is an extension of the general transformation approach described above. However, the same approach can be adapted with only minor alterations to implement the procedures developed by Box and Jenkins (1976), Glass and colleagues (1975), or Simonton (1977).

Pooled Time Series Analysis

This approach to time series analysis for multiple units represents a direct extension of the analysis for single units and requires only the use of a patterned transformation matrix. The specific choice of the design matrix **X** and the number of units will be dictated by the particular questions of interest. The procedure will be illustrated by a two-unit example (K = 2), where the design employed involves only level and change in level (the design matrix in panel A in table 1).

The observations for all the units can be represented by a supervector of observations of length N, which is composed of the vector observations (preintervention and postintervention) for each of the units, or

$$Z = \left[\begin{array}{c} Z_1 \\ \hline Z_2 \end{array}\right] \tag{9}$$

and where there are n_1 observations before intervention and n_2 observations after intervention on both unit 1 and unit 2. Table 3 presents an example of the patterned general transformation matrix that would be employed to transform the serially dependent Z_i variables to the serially independent variables Y_i. The transformation matrix always will take the form

$$\begin{bmatrix} \mathbf{T}^* & \mathbf{0} & \mathbf{0} & . & . & . & \mathbf{0} \\ \mathbf{0} & \mathbf{T}^* & \mathbf{0} & . & . & . & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{T}^* & . & . & . & \mathbf{0} \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & . & . & . & \mathbf{T}^* \end{bmatrix} \tag{10}$$

where $\mathbf{T}^*$ is an NxN lower diagonal transformation matrix ($N = n_1+n_2$) and **0** is an NxN null matrix. The occurrence of the null matrices in all positions except the diagonal reflects the assumption of independence of the different units.

The use of a properly parameterized design matrix will permit comparisons between different units. Table 4 presents an illustrative example. The design matrix in panel (a) includes four parameters that reflect

TABLE 3. *Example of general transformation matrix (T) for cross-sectional analysis ($k = 2$: $n_{11} = n_{12} = n_{21} = n_{22} = 4$)*

1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
W_1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
W_2	W_1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
W_3	W_2	W_1	1	0	0	0	0	0	0	0	0	0	0	0	0
W_4	W_3	W_2	W_1	1	0	0	0	0	0	0	0	0	0	0	0
W_5	W_4	W_3	W_2	W_1	1	0	0	0	0	0	0	0	0	0	0
0	W_5	W_4	W_3	W_2	W_1	1	0	0	0	0	0	0	0	0	0
0	0	W_5	W_4	W_3	W_2	W_1	1	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	W_1	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	W_2	W_1	1	0	0	0	0	0
0	0	0	0	0	0	0	0	W_3	W_2	W_1	1	0	0	0	0
0	0	0	0	0	0	0	0	W_4	W_3	W_2	W_1	1	0	0	0
0	0	0	0	0	0	0	0	W_5	W_4	W_3	W_2	W_1	1	0	0
0	0	0	0	0	0	0	0	0	W_5	W_4	W_3	W_2	W_1	1	0
0	0	0	0	0	0	0	0	0	0	W_5	W_4	W_3	W_2	W_1	1

level and change in level for both units and the difference between the two units on preintervention and postintervention change in level. If the last parameter (i.e., the difference between the units on the postintervention change in level) is not significant, the design matrix in panel (b) would be adopted, reflecting no difference between the two units in intervention effects (change in level). Differences between units would seem likely to be fairly common for most problems. However, if no such differences exist, the design matrix in panel (c) would be appropriate. The design matrix in panel (d) is appropriate if no intervention effects or differences between units exist.

TABLE 4. *Example of design matrix (X) for cross-sectional problem with level and change in level analysis*

(a) Full model	(b) No difference in intervention effects	(c) No difference in individual effects	(d) No intervention effects
1 0 0 0	1 0 0	1 0	1
1 0 0 0	1 0 0	1 0	1
1 0 0 0	1 0 0	1 0	1
1 0 0 0	1 0 0	1 0	1
1 1 0 0	1 1 0	1 1	1
1 1 0 0	1 1 0	1 1	1
1 1 0 0	1 1 0	1 1	1
1 1 0 0	1 1 0	1 1	1
1 0 1 0	1 0 1	1 0	1
1 0 1 0	1 0 1	1 0	1
1 0 1 0	1 0 1	1 0	1
1 0 1 0	1 0 1	1 0	1
1 1 1 1	1 1 1	1 1	1
1 1 1 1	1 1 1	1 1	1
1 1 1 1	1 1 1	1 1	1
1 1 1 1	1 1 1	1 1	1

The procedure can be generalized to any number of units and any choice of design matrix. Implicit is the assumption that a common transformation matrix is appropriate for all units. This assumption seems reasonable if the nature of the series is viewed as determined by an underlying process specific to the construct under investigation. As with any of the analytic approaches, diagnostic indicators like the Ljung and Box test (1978) may be used to test the fit of the model. The basic form of the design matrix should be based on the analyses of the individual units and/or a priori knowledge when available.

The approach described here has a number of advantages. First, it represents a direct extension of the general transformation approach developed by Velicer and McDonald (1984). This approach avoids the problematic model identification step and has received a favorable evaluation in several simulation studies (Harrop and Velicer 1985, 1990*b*).

Second, the approach described here also can be adapted to two of the alternative methods of analysis. For the approach developed by Glass and colleagues (1975), a specific transformation matrix could be specified for a particular ARIMA (p, d, q) model and would replace the general transformation matrix employed here. Following the Simonton (1977) approach, the ARIMA (1, 0, 0) transformation matrix would be used for all cases instead of the general transformation approach.

Third, the approach is a simple, direct extension of existing procedures. It can be implemented easily by a slight modification of existing computer programs like GENTS (Velicer et al. 1986) or TSX (Glass et al. 1974). The problems of adaptation will involve problems of size and speed created by the use of supervectors rather than an increased complexity of the analysis.

Meta-Analysis

An alternative procedure to combining data from several individuals or units is meta-analysis. Procedures for performing a meta-analysis have been well developed for traditional experimental designs (Hedges and Olkin 1985; Hunter and Schmidt 1990; Tobler, this volume). Meta-analysis procedures have not been applied previously to single-subject designs. Two problems exist in applying meta-analysis to this area: (1) primary research reports often have relied on visual analysis rather than time series analysis, resulting in a lack of basic statistical information (O'Rourke and Detsky 1989), and (2) alternative definitions of effect size must be developed. Allison and Gorman (1992) review some alternative effect size calculations appropriate for time series designs.

DISCUSSION

The topics discussed in the previous sections—model identification, intervention assessment, and pooled time series analysis—represent the three critical issues in time series analysis that have received the most attention. There are several other topics that are either of less interest or currently are under development. They will be discussed briefly in this section.

Cyclic Data

A potential confounding variable in time series data is the presence of cyclic or seasonal data. Economic data frequently are affected by the months of the year, or the "season." Daily data gathered on individuals may have a weekly or monthly cycle. Three alternative procedures, discussed below, have been proposed to deal with cyclic data.

Deseasonalization. In some content areas, the cyclic nature of the data is well known. For example, in economics, many of the data are adjusted for seasonal effects before it is reported. These seasonal adjustments, based on a priori information, remove cyclic trends from the data prior to any time series analysis.

Statistical Control. An alternative method of adjusting for seasonal effects is to find some variable that is sensitive to the same seasonal effects as the dependent measure but cannot be affected by the intervention. This variable then could be used as a covariate. The cyclic effects would be statistically controlled. Some of the problems in using a covariate are discussed below.

Combined Models. A third alternative approach involves the use of combined models. McCleary and Hay (1980) discuss this approach in detail. As an example, suppose a time series is represented by a lag 1 moving averages model, as below:

$$Z_t - L = A_t - \Theta_1 A_{t-1} \tag{11}$$

Furthermore, assume that a seasonal component of lag 12 also is present. This could be modeled as

$$Z_t - L = A_t - \Theta_{12} A_{t-12} \tag{12}$$

The time series, therefore, would be described as an ARIMA (0, 0, 1) (0, 0, 1)$_{12}$ model or

$$Z_t - L = (A_t - \Theta_1 A_{t-1})(A_t - \Theta_{12} A_{t-12}) \tag{13}$$

Unlike the first two approaches, the combined models approach presents difficulties for the extension of this procedure to either pooled procedures or multivariate time series approaches and would require longer series.

Missing Data

Missing data are an almost unavoidable problem in time series analysis and present a number of unique challenges. Life events will result in missing data even for the most conscientious researchers. In the model identification study described previously, missing data were a relatively minor problem. Four subjects had no missing data (i.e., all 124 observations were available). For four other subjects, four or fewer observations were missing. Only two subjects showed significant amounts of missing data (115 and 97 observations).

The problem of missing data has received little attention in the behavioral sciences area. Rankin and Marsh (1985) assessed the impact of different amounts of missing data for 32 simulated time series modeled after 16 real-world data examples. They concluded that, with up to 20 percent missing data, there is little impact on model identification, but the impact is pronounced when more than 40 percent is missing. In an extensive simulation study, Colby and Velicer (under review) compared four different techniques of handling missing data: deletion from the analysis, substitution of the mean of the series, substitution of the mean of the two adjacent observations, and a maximum likelihood estimation (Little and Rubin 1987). The mean of the series was judged unacceptable. The mean of the adjacent points and deletion worked well for a large number of cases but not for all cases. The maximum likelihood procedure was the best procedure across all conditions.

Computer Programs

Analysis of time series data requires the use of a computer program. Fortunately, a large number of programs have become available in the last two decades. Unfortunately, the quality of the available programs is quite variable. Harrop and Velicer (1990*a*, 1990*b*) evaluated five programs: BMDP (Dixon 1985), GENTS (Velicer et al. 1986), ITSE (Williams and Gottman 1982), SAS (SAS Institute 1984), and TSX (Bower and Glass 1974). Simulated data from 44 different ARIMA models were employed to assess the accuracy of the programs (Harrop and Velicer 1990*b*). Three programs produced generally satisfactory results (TSX, GENTS, and SAS). One was inaccurate across a wide range of models (ITSE), and one was occasionally inaccurate and occasionally failed to complete the analysis (BMDP). For all five programs, the overall evaluation of the computation features and quality of documentation was not very favorable (Harrop and Velicer 1990*a*). All suffered from at least one flaw, with documentation frequently being either nonexistent or inadequate. In particular, SAS and BMDP did not provide adequate documentation for most social science applications. TSX and GENTS had no documentation aside from published research reports and comments contained in the code.

Multivariate Time Series Analysis

Time series analysis on a single dependent measure involves many of the procedures common to the multivariate statistics because two vectors of unknowns must be estimated simultaneously; these are the vector of parameters and the vector of coefficients, which represent the dependency in the data. The term "multivariate time series" denotes the observation of more than one variable at each point in time. If the additional variables are conceptualized as being unable to be influenced by the intervention, the appropriate analysis has been labeled a concomitant variable analysis (Glass et al. 1975) and is a direct analog of the analysis of covariance. The covariate is employed to statistically remove some variation from the dependent measure, thus increasing sensitivity. Two problems arise: (1) What is the proper lag between the covariate and dependent variable, and (2) how should dependency in the covariate be handled? One application of this procedure is to control the effects of seasonality in the data (see **Cyclic Data** section).

Alternatively, all of the observed variables could be conceptualized as dependent measures. Molenaar (1985, 1987), Molenaar and colleagues (1992), and Peña and Box (1987) have presented two approaches to this problem, but examples of the application of these procedures have not appeared in the literature yet. The problems are direct extensions of the covariate applications (i.e., determining the appropriate lag for relating the dependent measures and dealing with the potential of different dependency models). In addition, alternative approaches could involve dealing with all p dependent measures simultaneously, combining the p measures into a single optimum composite, or defining a set of m new composites (m < p) and interpreting these composites.

Application Issues

A number of critical design issues must be addressed before applying time series analysis to substance abuse problems. First, the unit of analysis must be defined. For several examples discussed here, the unit of analysis was assumed to be a single individual. Treatment outcome studies, even if they involve multiple subjects, can be analyzed profitably as a series of studies at the individual level. The outcome of the studies can be treated as replications and combined using cross-sectional procedures or meta-analysis procedures. If differences exist between subjects, hypotheses can be generated and a systematic replication procedure employed (Hersen and Barlow 1976). Alternatively, the unit can be an aggregate group of people, and the interventions can apply only at the group level, such as policy changes. The same methods of analysis can be applied to the group data.

Studies of this type typically are called evaluation studies (Cook and Campbell 1979). A recent study of this type investigated the impact of two interventions on narcotics use and property crime (Powers et al. 1991); the researchers concluded that methadone treatment has long-term benefits in reducing drug use and property crime but that legal supervision had the contrary effect of increasing both property crimes and narcotics use.

Second, only very simple designs have been described here. More complex designs involving multiple interventions may be appropriate, and the analysis procedures generally differ only with respect to the design matrix employed. A variety of textbooks discuss alternative designs and the relation of the designs to different threats to validity (Campbell and Stanley 1963; Cook and Campbell 1979; Glass et al. 1975).

Time series analysis has a tremendous potential for applications to substance abuse problems. During the last decade, a combination of computational advances and alternative statistical procedures have increased the ease of application and the range of potential applications. Two of the early drawbacks, the large sample size required for model identifications and problems with generalizability, have been largely overcome in the last decade. Time series analysis should be viewed as representing one of a variety of potential methods of analysis available to all researchers rather than a novel and difficult procedure.

NOTES

1. Italicized sections may be skipped without loss of continuity.

2. Requests for reprints should be sent to Wayne F. Velicer, Ph.D., Professor and Co-Director, Cancer Prevention Research Center, University of Rhode Island, Flagg Road, Kingston, RI 02881-0808 (BITNET: KZP101@URIACC).

REFERENCES

Aitken, A.C. On least squares and lineal combination of observations. *Proc R Soc Edinburg H* 55:42-47, 1934.

Akaike, H. A new look at the statistical model identification. *IEEE Trans Autom Control* 19:716-723, 1974.

Algina, J., and Swaminathan, H.A. A procedure for the analysis of time series designs. *J Exp Educ* 45:56-60, 1977.

Algina, J., and Swaminathan, H.A. Alternatives to Simonton's analysis of the interrupted and multiple-group time series designs. *Psychol Bull* 86:919-926, 1979.

Allison, A.B., and Gorman, B.S. "Calculating Effect Sizes for Meta-analysis: The Case of the Single Case." Paper presented at the American Psychological Association Meeting, Washington, DC, August 1992.

Beguin, J.M.; Courieroux, C.; and Monfort, A. Identification of a mixed autoregressive-moving average process: The corner method. In: Anderson, O.D., ed. *Time Series: Proceedings of the International Conference Held at Nottingham University, March 1979.* Amsterdam: North-Holl. , d, 1980. pp. 423-436.

Bhansali, R.J., and Downham, D.Y. Some properties of the order of an autoaggressive model selected by a generalization of Akaike's FPE-criterion. *Biometrika* 64:547-551, 1977.

Bower, C., and Glass, G.V. *TSX.* Computer program. Boulder, CO: University of Colorado, 1974.

Box, G.E.P., and Jenkins, G.M. *Time-Series Analysis: Forecasting and Control.* San Francisco: Holden Day, 1976.

Box, G.E.P., and Pierce, W.A. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *J Am Stat Assoc* 65:1509-1526, 1970.

Box, G.E.P., and Tiao, G.C. A change in level of nonstationary time series. *Biometrika* 52:181-192, 1965.

Box, G.E.P., and Tiao, G.C. Intervention analysis with application to conomic and environmental problems. *J Am Stat Assoc* 70:70-92, 1975.

Campbell, D.T., and Stanley, J.C. *Experimental and Quasi-Experimental Design for Research.* Chicago: Rand McNally, 1963.

Colby, S.M., and Velicer, W.F. "A Comparison of Four Alternative Procedures for Handling Missing Data in a Time Series Analysis." Under review.

Cook, T.D., and Campbell, D.T., eds. *Quasi-Experimentation: Design and Analysis Issues for Field Settings.* Chicago: Rand McNally College Publishing, 1979.

Dixon, W.J. *BMDP Statistical Software.* Computer program manual. Berkeley, CA: University of California Press, 1985.

Djuric, P.M., and Kay, S.M. Order selection of autoregressive models. *IEEE Trans Acoust Speech Signal Process* 40:2829-2893, 1992.

Glass, G.V.; Bower, C.; and Padia, W.L. *TSX.* Computer program. Boulder, CO: University of Colorado, 1974.

Glass, G.V.; Willson, V.L.; and Gottman, J.M. *Design and Analysis of Time Series Experiments.* Boulder, CO: Colorado Associate University Press, 1975.

Gottman, J.M. N-of-one and N-of-two research in psychotherapy. *Psychol Bull* 80:93-105, 1973.

Gottman, J.M. *Time-Series Analysis.* New York: Cambridge University Press, 1981.

Gottman, J.M., and Glass, G.V. Analysis of interrupted time-series experiments. In: Kratochwill, J., ed. *Strategies To Evaluate Change in Single Subject Research.* New York: Academic Press, 1978.

Grey, H.L.; Kelly, G.D.; and McIntire, D.D. A new approach to ARIMA modeling. *Commun Stat* B7:1-77, 1978.

Hannan, E.J., and Rissanen, J. Recursive estimation of mixed autoregressive moving average order. *Biometrika* 69:81-94, 1982.

Harrop, J.W., and Velicer, W.F. A comparison of three alternative methods of time series model identification. *Multivariate Behav Res* 20:27-44, 1985.

Harrop, J.W., and Velicer, W.F. Computer programs for interrupted time series analysis: I. A qualitative evaluation. *Multivariate Behav Res* 25:219-231, 1990*a*.

Harrop, J.W., and Velicer, W.F. Computer programs for interrupted time series analysis: II. A quantitative evaluation. *Multivariate Behav Res* 25:233-249, 1990*b*.

Heatherton, T.F.; Kozlowski, L.T.; Frecker, R.C.; Rickert, W.; and Robinson, J. Measuring heaviness of smoking: Using self-reported time to the first cigarette of the day and number of cigarettes smoked per day. *Br J Addict* 84:791-800, 1989.

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

Hersen, M., and Barlow, D. *Single Case Experimental Designs: Strategies for Studying Behavior Change.* New York: Pergamon Press, 1976.

Hunter, J.E., and Schmidt, F.L. *Methods for Meta-Analysis.* Newbury Park, CA: Sage, 1990.

Jarvik, M.E. Further observations on nicotine as the reinforcing agent in smoking. In: Dunn, W.L., Jr., ed. *Smoking Behavior: Motives and Incentives.* Washington, DC: V.H. Winston, 1973.

Kashyap, R.L. A Bayesian comparison of different classes of dynamic models using empirical data. *IEEE Trans Autom Control* 22:715-727, 1977.

Leventhal, H., and Cleary, P.D. The smoking problem: A review of the research and theory in behavioral risk modification. *Psychol Bull* 88:370-405, 1980.

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

Ljung, G.M., and Box, G.E.P. On a measure of lack of fit in time series models. *Biometrika* 65:297-303, 1978.

McCleary, R., and Hay, R.A., Jr. *Applied Time Series Analysis for the Social Sciences.* Beverly Hills, CA: Sage, 1980.

McFall, R.M. Smoking cessation research. *J Consult Clin Psychol* 76:703-712, 1978.

Molenaar, P.C.M. A dynamic factor model for the analysis of multivariate time series. *Psychometrika* 50:181-202, 1985.

Molenaar, P.C.M. Dynamic factor analysis in the frequency domain: Causal modeling of multivariate psychophysiological time series. *Multivariate Behav Res* 22:329-353, 1987.

Molenaar, P.C.M.; De Gooijer, J.G.; and Schmitz, B. Dynamic factor analysis of nonstationary multivariate time series. *Psychometrika* 57:333-349, 1992.

Morrison, D.F. *Applied Linear Statistical Methods*. Englewood Cliffs, NJ: Prentice Hall, 1983.

O'Rourke, K., and Detsky, A.S. Meta-analysis in medical research: Strong encouragement for higher quality in individual research efforts. *J Clin Epidemiol* 42:1021-1024, 1989.

Parzen, E. Some recent advances in time series modelling. *IEEE Trans Autom Control* 19:723-729, 1974.

Peña, D., and Box, G.E.P. Identifying a simplifying structure in time series. *J Am Stat Assoc* 82:836-843, 1987.

Powers, K.; Hanssens, D.M.; Hser, Y.; and Anglin, M.D. Measuring the long-term effects of public policy: The case of narcotic use and property crime. *Manag Sci* 37:627-644, 1991.

Printz, A.M. "Stress Reduction in the Treatment of Essential Hypertension: A Clinical Trial Utilizing Assertion and Relaxation Coping Skills." Ph.D. diss., University of Rhode Island, Kingston, 1978.

Pukkila, T.M. On the identification of ARIMA (p,q) models. In: Anderson, O.D., ed. *Time Series Analysis Theory and Practice I: Proceedings of the International Conference Held at Valencia, Spain, June 1981*. Amsterdam: North-Holland, 1982. pp. 81-103.

Rankin, E.D., and Marsh, J.C. Effects of missing data on the statistical analysis of clinical time series. *Soc Work Res Abstr* 21:13-16, 1985.

Rissanen, J. Modeling by shortest data description. *Automatica* 14:465-478, 1978.

Rissanen, J. Order estimation by accumulated prediction errors. *J Appl Probab* 12A:55-61, 1986*b*.

Rissanen, J. Stochastic complexity and modeling. *Ann Stat* 14:1080-1100, 1986*a*.

Russell, M.A. Nicotine chewing gum as a substitute for smoking. *Br Med J* 1:1060-1063, 1977.

SAS Institute. *SAS/ETS User's Guide, Version 5 Edition*. Computer program manual. Cary, NC: SAS Institute, 1984.

Schachter, S.L. Nicotine regulation in heavy and light smokers. *J Exp Psychol: [Gen]* 106:5-12, 1977.

Schwartz, G. Estimating the dimension of a model. *Ann Stat* 6:461-469, 1978.

Simonton, D.K. Cross-sectional time-series experiments: Some suggested statistical analyses. *Psychol Bull* 84:489-502, 1977.

Solomon, R.L. The opponent-process theory of acquired motivation. *Am Psychol* 35:691-712, 1980.

Solomon, R.L., and Corbit, J.D. An opponent-process theory of motivation: I. Temporal dynamics of affect. *Psychol Rev* 81:119-145, 1974.

Solomon, R.L., and Corbit, J.D. An opponent-process theory of motivation: II. Cigarette addiction. *J Abnorm Psychol* 81:158-171, 1973.

Swaminathan, H., and Algina, J. Analysis of quasi-experimental time-series designs. *Multivariate Behav Res* 12:111-131, 1977.

Tomkins, S.S. Psychological model for smoking behavior. *Am J Public Health* 68:250-257, 1966.

Tomkins, S.S. A modified model of smoking behavior. In: Borgatta, E.F., and Evans, R.R., eds. *Smoking, Health and Behavior*. Chicago: Aldine, 1968.

Tsay, R.S. Regression models with time series errors. *J Am Stat Assoc* 79:118-124, 1984.

Tsay, R.S., and Tiao, G.C. Consistent estimates of autoregressive parameters and extended sample autocorrelation function for stationary and nonstationary ARIMA models. *J Am Stat Assoc* 79:84-90, 1984.

Velicer, W.F.; Fraser, C.; McDonald, R.P.; and Harrop, J.W. *GENTS.* Computer program. Kingston, RI: University of Rhode Island, 1986.

Velicer, W.F., and Harrop, J.W. The reliability and accuracy of time series model identification. *Eval Rev* 7:551-560, 1983.

Velicer, W.F., and McDonald, R.P. Time series analysis without model identification. *Multivariate Behav Res* 19:33-47, 1984.

Velicer, W.F., and McDonald, R.P. Cross-sectional time series designs: A general transformation approach. *Multivariate Behav Res* 26:247-254, 1991.

Velicer, W.F.; Prochaska, J.O.; Rossi, J.S.; and Snow, M. Assessing outcome in smoking cessation studies. *Psychol Bull* 111:23-41, 1992*b*.

Velicer, W.F.; Redding, C.A.; Richmond, R.; Greeley, J., and Swift, W. A time series investigation of three nicotine regulation models. *Addict Behav* 17:325-345, 1992*a*.

Williams, E.A., and Gottman, J.M. *A User's Guide to the Gottman-Williams Time-Series Analysis Computer Programs for Social Scientists*. Computer program manual. Cambridge, MA: Cambridge University Press, 1982.

ACKNOWLEDGMENT

This work was supported partially by grants CA27821 and CA50087 from the National Cancer Institute.

AUTHOR

Wayne F. Velicer, Ph.D.
Professor and Co-Director
Cancer Prevention Research Consortium
University of Rhode Island
Flagg Road
Kingston, RI 02881-0808

Use and Misuse of Repeated Measures Designs

Robert S. Barcikowski and Randall R. Robey

ABSTRACT

Repeated measures designs should be used more frequently in prevention intervention research. They are the design of choice when one or more measurements have been taken at baseline followed by one or more measurements after prevention intervention. They may be used to ask questions about differences on measurements at different points in time and between measures made on the same scale. In this presentation, prevention intervention researchers are provided with a step-by-step discussion of this design, examples of prevention intervention repeated measures designs, and a discussion of the misuses of this design.

INTRODUCTION

This chapter is written in the form of a dialog between its authors and researchers in drug abuse prevention. Each section begins with a question that would be asked by a prevention intervention researcher, and then the question is answered. This format provides a step-by-step presentation of repeated measures designs, their analyses, and a discussion of why they frequently are misused. The reader interested in pursuing this topic further will find discussions of repeated measures designs in Bock (1975), Crowder and Hand (1990), Davidson (1972), Games (1990*a*, 1990*b*), Keppel (1991), Kirk (1982), Maxwell and Delaney (1990), Morrison (1990), Rich (1983), Stevens (1992), Timm (1975), and Winer (1971).

WHAT IS A REPEATED MEASURES DESIGN?

A repeated measures design is a statistical design wherein units (e.g., subjects) are measured (e.g., tested) more than once with either the same instrument or different but commensurate instruments.

WHAT ARE COMMENSURATE INSTRUMENTS?

Commensurate instruments are instruments measured on the same scale. Examples of commensurate instruments are:

1. The parallel forms used with the Miller's Analogy Test, and

2. The subtests on reading, arithmetic, spelling, and language usage from the Iowa Test of Basic Skills.

WHY DO RESEARCHERS USE REPEATED MEASURES DESIGNS?

There are two main reasons for using repeated measures designs:

1. Data naturally exist in this form (i.e., researchers frequently take more than one measurement on the same subject at different points in time), and

2. Researchers want to use the unit (e.g., subject) as its (e.g., his or her) own control.

WHAT DOES "TO USE THE UNIT (E.G., SUBJECT) AS ITS (E.G., HIS OR HER) OWN CONTROL" MEAN?

One may start to explain the preceding statement by comparing a one-way analysis of variance (ANOVA) with a repeated measures ANOVA.

Later in this chapter, the analysis to be discussed is referred to as part of the "univariate" approach to repeated measures analysis.

The chapter will examine a set of data to be analyzed using a one-way ANOVA. A simple three-treatment design will be used in which 15 subjects have been randomly sampled and 5 subjects have been randomly assigned to each of the three treatments. This arrangement also is referred to as a "completely randomized" one-way design because of the random fashion in which the subjects occur in each treatment.

To give this study meaning, consider an example of a potential drug prevention intervention study. Let the 15 subjects be alcoholics who were about to be involved in different prevention intervention programs. Prior to entering one of these programs, the alcoholics were placed in one of three treatment conditions where they were shown slides of different possible prevention intervention outcomes.

In this part of the study, the researchers wanted to know if the pictures did indeed represent different possible prevention intervention outcomes to the alcoholics. That is, the researchers were attempting to validate their instrument (the slides) for later use as measures in their prevention intervention study. The subjects in each of the treatments viewed 10 slides depicting people in different situations, as illustrated in figure 1.

The slides were placed in random order; past research had indicated no picture order effects. For each slide, the subject was asked: "How much does this picture depict a situation that you feel is possible for you?" The responses to each slide were based on the Likert scale shown below:

Very Possible	Possible	Don't Know	Unlikely	Impossible
(5)	(4)	(3)	(2)	(1)

Treatment	Viewed Ten Slides of People in Various Situations
1	Drinking
2	Working
3	Related to Family

FIGURE 1. *Types of slides viewed in the three treatments*

The research problem was: Are there differences in alcoholics' perceptions of possible treatment outcomes? The resultant data are shown in figure 2.

In figure 2, the score for a subject was based on the average of the subject's Likert item-response scores; however, some items were weighted so that a maximum possible score for each situation was 10. These data were taken from an example provided by Gravetter and Wallnau (1985, p. 470). The authors have modified those data here in order to better illustrate their point.

Goal Situation		
Drinking	Work	Family
4	6	5
3	4	3
2	3	2
3	4	2
3	6	4
M1 = 3	M2 = 4.6	M3 = 3.2

FIGURE 2. *Data for the one-way ANOVA*

A one-way ANOVA of these data yielded the results shown in table 1. Given a .05 level of significance, the researcher would fail to find a significant difference ($p < .0971$) among the treatments.

TABLE 1. *ANOVA table for the one-way data shown in figure 3*

Source	df	Sum of Squares	Mean Squares	*F*-test	*P*-value
Between groups	2	7.6	3.8	2.85	.0971
Within groups	12	16.0	1.333		
Total	14	23.6			

If these same data are considered to be observations on a random sample of five subjects with each subject viewing all three stimuli, the single-group repeated measures design (with three commensurate measures) shown in figure 3 results.

When the analyses shown in table 1 (i.e., one-way ANOVA) and in table 2 (i.e., single-group repeated measures ANOVA) are compared, the sum-of-squares total is found to be the same in each table (i.e., 23.6). This is true for two reasons: (1) The scenarios for the two situations changed but the data remained the same, and (2) the sum-of-squares total is calculated as the sum of the squared differences of each observation from the grand mean, and the grand mean does not change.

However, the sum-of-squares total is partitioned differently in the two tables. In the single-group repeated measures ANOVA shown in table 2, the data were analyzed as a two-way ANOVA (a block design) with subjects treated as an additional factor (the blocking factor). Therefore, in this analysis, there is one observation per call only, and the interaction is taken as the estimate of error. This implies that no interaction exists

Subjects	Goal Situation		
	Drinking	Work	Family
S 1	4	6	5
S 2	3	4	3
S 3	2	3	2
S 4	3	4	2
S 5	3	6	4
	M1 = 3	M2 = 4.6	M3 = 3.2

FIGURE 3. *Data for the repeated measures ANOVA*

between the subjects and the treatments (goal situations) so that the plots of the observations across the treatments should be approximately parallel.

In table 2, given a .05 level of significance, the researcher would find a significant difference ($p < .0033$) among the treatments. In this analysis, the sum of squares between treatments (goal situations) remains the same as in the one-way analysis (i.e., 7.6). However, the within-groups sum of

TABLE 2. *ANOVA table for the repeated measures data shown in figure 4*

Source	df	Sum of Squares	Mean Square	*F*-test	*P*-value
Between subjects	4	13.6	3.4		
Within subjects	10	10.0	1.0		
treatments	2	7.6	3.8	12.667	.0033
residual	8	2.4	.3		
Total	14	23.6			

squares (16) from the one-way analysis in table 1 has been partitioned in table 2 into the sum of squares between subjects (13.6) and the residual or error (2.4).

The large reduction in the residual (error) sum of squares (and the accompanying small reduction in the error degrees of freedom [df]) yield an estimate of the error variance in table 2 (.3) that is much smaller than that found in table 1 (1.333) for the one-way design. This reduction in error variance is what makes the repeated measures design more powerful (i.e., more sensitive to treatment differences) than the one-way ANOVA.

In a block design, the units within a block are selected to be homogeneous *within* and heterogeneous *between* blocks. In a repeated measures design, the units are homogeneous within because they are the same unit. Comparisons also may be made with the unit under nonexperimental conditions (e.g., baseline or control group conditions). In this sense, the unit (the subject) acts as its (his or her) own control.

THIS DESIGN LOOKS GOOD! DOES IT HAVE OTHER ADVANTAGES? DISADVANTAGES?

Keppel (1991, pp. 333-336) discusses the following advantages and disadvantages of repeated measures designs. Keppel refers to these designs as "within subjects designs."

Advantages

1. Control of subject heterogeneity (which was just considered).

2. Economy (studies do not require as many subjects and, because the subjects are familiar with the study, its running time may be reduced).

3. " . . . The repeated-measures design has become the most common experimental design with which to study such phenomena as learning, transfer, and practice effects of all sorts" (p. 334).

Disadvantages

1. Carryover effects (Keppel [1991, p. 335] refers to these as "general practice effects, they are effects which affect all treatment conditions equally").

2. Differential carryover effects (these are specific effects that affect "a subject's performance on a later condition one way and on a different condition another way" [p. 335]).

3. Nongeneralizability of the results (the results may not be duplicated by a completely randomized design).

4. The univariate model's assumption of sphericity may be difficult to meet.

CAN ANYTHING BE DONE ABOUT THE DESIGN'S DISADVANTAGES?

Yes! Each disadvantage and how it is handled is addressed in the following sections.

Carryover Effects

Carryover effects may be examined if the experimenter employs what is known as *counterbalancing* in his or her design. For example, consider a sample of students who smoke marijuana and who are trying to break the habit by wearing a drug patch. A study involving two types of prevention interventions was devised to examine the effects of the drug patch on academic achievement. The first type of prevention intervention was to be the wearing of a placebo patch. Counterbalancing would take place if half of the subjects were asked to wear the drug patch first and then given the placebo patch and the other half of the students were asked to wear the placebo patch first, followed by the drug patch.

Differential Carryover Effects

The following comments by Keppel (1991, p. 340) are informative with respect to this disadvantage:

> The most common way of reducing differential carryover effects is to provide sufficient time between sessions to allow the complete dissipation of the preceding treatment condition.
>
> In many cases, the presence of treatmentxposition interactions [differential carryover effects] simply rules out the within-subjects design for the study of a particular phenomenon.
>
> In other cases, however, these interactions have become the object of study, with experiments designed to shed light on the reasons for their occurrence.

Nongeneralizability of the Results

If the researcher is suspicious of this problem, he or she may have to reconsider the treatments in a completely randomized design.

Difficulty in Meeting the Sphericity Assumption

The statistical assumptions for a univariate repeated measures design are:

1. Independence of subjects;

2. Homogeneous variance-covariance matrices across groups (necessary only if there is more than one group of subjects);

3. Multivariate normality;

4. A linear model; and

5. Sphericity (also known as circularity).

The *sphericity assumption* is discussed in detail by Huynh and Feldt (1970) and Rouanet and Lepine (1970). Sphericity is achieved when the variances among all possible pairwise differences of the treatments on the repeated measures factor are equal. The sphericity assumption is almost never met in practice. Rogan and colleagues (1979) indicate that neither the sphericity nor the homogeneity assumption is worth testing.

Sphericity is measured by a parameter denoted by "ε." The range for ε is: $1/(k-1) \leq \varepsilon \leq 1$, where k is the number of repeated measures. The sphericity assumption is met when $\varepsilon = 1$.

Imhof (1962) and Collier and colleagues (1967) present data that show that, if the sphericity assumption is not met, then the Type I error rate (i.e., the probability of rejecting a true null hypothesis, that is, of making a Type I error) is .05 when it actually could be .30.

This problem is handled in practice by multiplying the *F* test's df by an estimate of ε provided by the data; for example,

$$\mathbf{F(df1\ \hat{\varepsilon},\ df2\ \hat{\varepsilon})} \tag{1}$$

This has the effect of reducing the actual level of significance to that of the nominal level of significance.

When ε is less than .75, an estimate provided by Greenhouse and Geisser (1959), denoted by "$\hat{\varepsilon}$," s used. When ε is greater than or equal to .75, an estimate provided by Huynh and Feldt (1976), denoted by "$\tilde{\varepsilon}$," is used. In general, if the researcher is unfamiliar with what the ε is, it is recommended that he or she use the more conservative Greenhouse-Geisser (G-G) $\hat{\varepsilon}$ (Barcikowski and Robey 1984; Muller and Barton 1989).

HOW CAN A REPEATED MEASURES DESIGN BE ANALYZED?

Two approaches commonly are used. The most common is the univariate approach, which was discussed earlier. Another approach is the *multivariate* approach. Both approaches are reported automatically when a repeated measures analysis is requested in standard computer packages like BMDP4V (Dixon 1985), SAS (GLM) (SAS Institute, Inc. 1989), or SPSS (MANOVA) (SPSS, Inc. 1988).

Barcikowski and Robey (1984) and others (e.g., Looney and Stanley 1989) have recommended that both approaches be used in *exploratory* repeated measures analyses. This is because it is possible for either method to detect effects that the other may miss. Robey and Barcikowski (1989) discuss the control of Type I errors when both types of analyses are used.

HOW IS A MULTIVARIATE REPEATED MEASURES ANALYSIS EXECUTED?

A multivariate repeated measures analysis is completed by first transforming the measures and then performing a multivariate analysis on the transformed variables. The transformations are contrasts on the repeated measures. For example, reconsider the study of alcoholics' perceptions. A multivariate analysis could be performed by first transforming the measures into the differences between the drinking and work measures and between the drinking and family measures. The null hypothesis would then be tested on whether the vector of mean difference scores is equal to the null vector (a vector of 0's). Note that this is the multivariate extension of the dependent *t*-test where the null hypothesis is tested on whether the mean of the difference scores is equal to 0.

For the present drug prevention intervention study, the multivariate analysis would be completed on the set of difference scores shown in figure 4. The null hypothesis is that both of these mean differences are equal to 0 in the population.

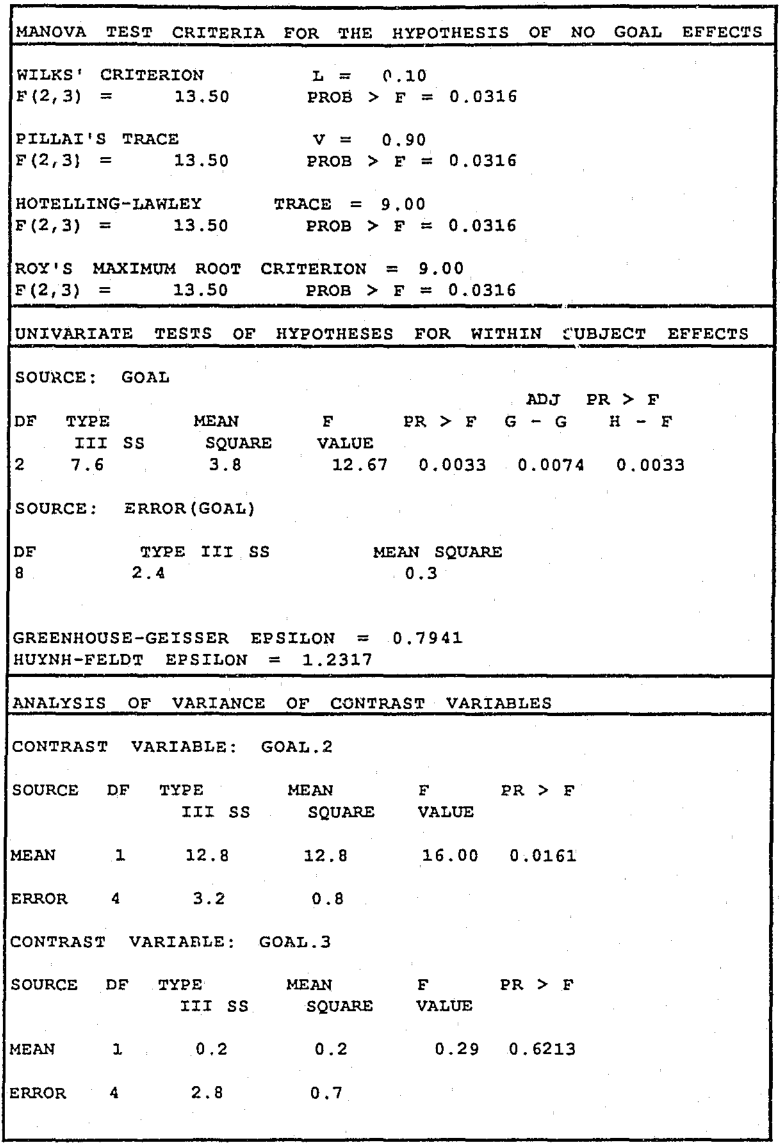
		Goal Differences	
		D-W	D-F
Subjects	S1	-2	-1
	S2	-1	0
	S3	-1	0
	S4	-1	1
	S5	-3	-1
M1 = -1.6			M2 = -.20

FIGURE 4. *Difference scores (contrasts)*

WHAT DOES AN OUTPUT FROM A STANDARD COMPUTER PACKAGE CONTAIN FOR A REPEATED MEASURES ANALYSIS?

The repeated measures output from all of the larger statistics packages (e.g., BMDP4V, SAS [GLM], or SPSS [MANOVA]) may be divided into three parts: omnibus multivariate tests, omnibus univariate tests, and tests on individual contrasts. The authors have selected the output from SAS (GLM) for the single-group repeated measures data shown in figure 3 to illustrate these three parts.

The SAS output for the single-group repeated measures data is shown in figure 5. The output has been partitioned into three parts and slightly modified to fit in the space allowed. The omnibus multivariate tests are output by SAS first, followed by the omnibus univariate tests and the tests on individual contrasts.



```
MANOVA  TEST  CRITERIA  FOR  THE  HYPOTHESIS  OF  NO  GOAL  EFFECTS

WILKS'  CRITERION          L  =   0.10
F(2,3)  =       13.50      PROB  >  F  =  0.0316

PILLAI'S  TRACE            V  =   0.90
F(2,3)  =       13.50      PROB  >  F  =  0.0316

HOTELLING-LAWLEY        TRACE  =  9.00
F(2,3)  =       13.50      PROB  >  F  =  0.0316

ROY'S  MAXIMUM  ROOT  CRITERION  =  9.00
F(2,3)  =       13.50      PROB  >  F  =  0.0316

UNIVARIATE  TESTS  OF  HYPOTHESES  FOR  WITHIN  SUBJECT  EFFECTS

SOURCE:   GOAL
                                                  ADJ    PR > F
DF   TYPE          MEAN         F        PR > F   G - G     H - F
      III  SS        SQUARE      VALUE
2    7.6            3.8          12.67   0.0033   0.0074   0.0033

SOURCE:   ERROR(GOAL)

DF           TYPE  III  SS           MEAN  SQUARE
8           2.4                         0.3


GREENHOUSE-GEISSER  EPSILON   =   0.7941
HUYNH-FELDT  EPSILON   =   1.2317

ANALYSIS   OF   VARIANCE   OF   CONTRAST   VARIABLES

CONTRAST   VARIABLE:   GOAL.2

SOURCE    DF    TYPE          MEAN          F         PR  >  F
                  III  SS       SQUARE       VALUE

MEAN       1       12.8        12.8        16.00    0.0161

ERROR      4        3.2         0.8

CONTRAST   VARIABLE:   GOAL.3

SOURCE    DF    TYPE          MEAN          F         PR  >  F
                  III  SS       SQUARE       VALUE

MEAN       1        0.2         0.2          0.29    0.6213

ERROR      4       2.8          0.7
```

FIGURE 5. *SAS (GLM) three-part output*

Given the single-group data, all of the omnibus multivariate tests in figure 5 yield the same probability values. This would not be true of the tests in more complex designs.

The output for the omnibus univariate tests follows and contains the G-G estimate, $\varepsilon^{\wedge} = .7941$, and the Huynh-Feldt (H-F) estimate, $\varepsilon^{\sim} = 1.2317$. When the H-F estimate is greater than 1, it is reset at 1. These tests indicate that, for these data, the univariate tests are more powerful than the multivariate tests because their probability values of .0033, .0074, and .0033 are smaller than the multivariate test's probability value of .0316. The authors use the probability value based on the G-G estimate (.0074) to interpret this section because they had no prior estimate of ε.

The last part of the output shown in figure 5 contains information on the test of each transformation (i.e., contrast). The results labeled "GOAL.2" test the mean goal differences between the drinking and work measures, and the results labeled "GOAL.3" test the mean goal differences between the drinking and family measures.

WHAT IS THE DIFFERENCE BETWEEN A MULTIVARIATE ANALYSIS AND A REPEATED MEASURES MULTIVARIATE ANALYSIS?

In a multivariate analysis, the focus is on differences between treatments using all of the dependent variables. In a multivariate repeated measures analysis, the researchers usually are interested in differences among the dependent variables.

WHAT ARE SOME COMMON REPEATED MEASURES DESIGNS?

Repeated measures designs usually are described as having *between* and *within* factors. The between factor(s) describe treatments or groups of units (e.g., subjects). The within factor(s) describe the repeated measures.

The drug prevention intervention data set in figure 3 illustrates a single group of subjects with commensurate measures. Therefore, there were no between (or grouping) factors and just one within factor. Repeated measures analyses of commensurate measures frequently are referred to as "profile analyses" (Morrison 1990).

Now consider examples of other repeated measures designs that might be found commonly in drug prevention intervention research. Probably the most common design encountered is one where an initial (baseline) measure is taken, followed by a prevention intervention, a measurement, and another measurement. A diagram of this design might look like the one in figure 6.

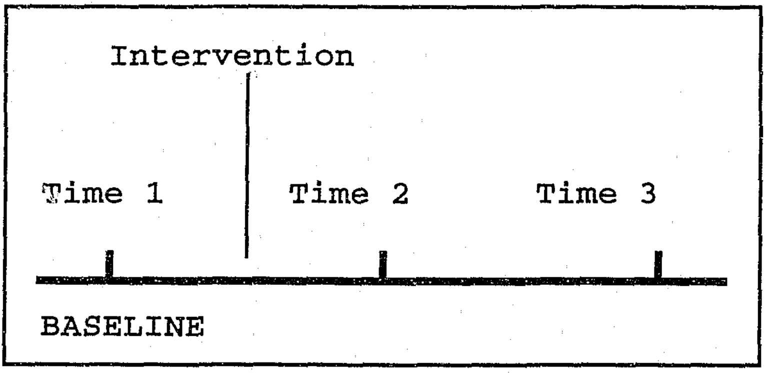


FIGURE 6. *Common intervention study*

In the following nine examples, the authors have used the design illustrated in figure 6 by modifying the scenario about the alcoholics they have been considering. In this new scenario, the procedures will be the same, but the score from each subject at each time will represent the average score across all 30 responses. Here, the scale for the drinking items is reversed so that a high score indicates less drinking, but the total scores still will be computed so that the maximum score is 10. This scenario must be modified slightly for some of the following designs.

Now consider a variety of designs that could be built from the basic repeated measures design illustrated in figure 6. For each design to be considered, the authors have provided:

1. A name for the design;

2. A picture of the design based on an expansion of the preceding scenario;

3. Omnibus generic question(s) answered through a repeated measures analysis of the design;

4. Omnibus example question(s) answered in terms of variables based on an expansion of the preceding scenario;

5. A description of the univariate and multivariate analyses; and

6. An example of a statistical package's between and within input statements using BMDP4V.

EXAMPLES OF REPEATED MEASURES DESIGNS

One-Within Design

Design 1. Our first example is called a *one-within* (or single-group) design. Barcikowski and Robey (1984) provide detailed information on the analysis of this design.

Picture. A picture of the one-within (single-group) design for a single group of five alcoholics is given in figure 7.

Main Question. The omnibus question answered by this design is: "Are there differences among the repeated measures?"

		Time		
		Time 1	Time 2	Time 3
Subjects	S 1	4	6	5
	S 2	3	4	3
	S 3	2	3	2
	S 4	3	4	2
	S 5	3	6	4
		M1 = 3	M2 = 4.6	M3 = 3.2

FIGURE 7. *Design 1: One-within (single-group) design*

Example Question. The omnibus question answered in this example is: "Are there differences in alcoholics' perceptions of themselves at different times?"

Analyses. The univariate repeated measures analysis is a two-way mixed-model analysis with subjects (random) and repeated measures (fixed). The multivariate analysis is a single-group multivariate analysis with contrasts on the repeated measures as dependent variables.

BMDP4V Input Statements. The authors feel that these statements will provide the researcher with a sense of the similar types of statements that are used by this and other programs. The BMDP4V program requires that the factors in a repeated measures design be identified as either BETWEEN or WITHIN. The BMDP4V statements for the first design, shown below, have only a WITHIN set because there is only one group of subjects.

```
/WITHIN     FACTOR = TIME.
            CODES  = 1 TO 3.
            NAMES  = TIME1, TIME2, TIME3.
```

One-Between and One-Within Design

Design 2. The second design is a *one-between and one-within* design (sometimes called a *split-plot* design). Looney and Stanley (1989) provide detailed information on the analysis of this design.

Picture. An example of this design, shown in figure 8, consists of three groups of alcoholics with five alcoholics randomly assigned to each group. A group of alcoholics receives one of three counseling prevention intervention methods (Rogerian, Adlerian, and Eclectic), and each subject is measured at baseline (Time 1) and two times after the introduction of their counseling method.

			Time	
		Time 1	Time 2	Time 3
Counseling method				
Rogerian	S11	4	6	5
	S12	3	4	3
	S13	2	3	2
	S14	3	4	2
	S15	3	6	4
Adlerian	S21	3	4	3
	S22	3	5	3
	S23	3	5	3
	S24	5	7	4
	S25	1	5	2
Eclectic	S31	4	7	5
	S32	7	9	7
	S33	3	5	3
	S34	5	7	4
	S35	5	7	4

FIGURE 8. *Design 2: One-between and one-within design*

Main Questions. There are three omnibus questions that are answered by this design: (1) "Is there an interaction between the between and within factors?" (2) "Given no interaction, are there differences among the repeated measures?" (3) "Given no interaction, are there any differences among the between factor's treatment levels?"

Example Questions. The omnibus questions answered in this example are: (1) "Is there an interaction between the counseling prevention intervention methods and time?" (2) "Given no interaction, are there differences in alcoholics' perceptions of themselves at different times?" (3) "Given no interaction, are there differences in alcoholics' perceptions of themselves among the counseling prevention intervention methods?"

Analyses. The univariate repeated measures analysis is a hierarchical (nested) mixed-model analysis with subjects (random) nested within the levels of the (fixed) between-treatments factor and the subjects crossed with the repeated measures factor (fixed). The multivariate analysis is a one-way multivariate analysis with contrasts on the repeated measures as dependent variables.

BMDP4V Input Statements. The BMDP4V statements for the second design, shown below, have a BETWEEN set to identify the counseling groups and a WITHIN set to identify the times.

```
/BETWEEN   FACTOR = METHOD.
           CODES  = 1 TO 3.
           NAMES  = ROGERIAN, ADLERIAN,
                    ECLECTIC.
/WITHIN    FACTOR = TIME.
           CODES  = 1 TO 3.
           NAMES  = TIME1, TIME2, TIME3.
```

Two-Between and One-Within (Complex Split-Plot) Design

Design 3. In the third design, the authors have added another set of groups so that there are two between factors and one within factor.

Picture. A picture of this design, shown in figure 9, is an expansion of the previous design with the addition of a control set of prevention intervention groups. Subjects in the control set of prevention intervention counseling groups are nonalcoholics who are seeking drug counseling.

			Time		
			Time 1	Time 2	Time 3
Alcoholics	Rogerian	S111	4	6	5
		S112	3	4	3
		S113	2	3	2
		S114	3	4	2
		S115	3	6	4
	Adlerian	S121	3	4	3
		S122	3	5	3
		S123	3	5	3
		S124	5	7	4
		S125	1	5	2
	Eclectic	S131	4	7	5
		S132	7	9	7
		S133	3	5	3
		S134	5	7	4
		S135	5	7	4
Control	Rogerian	S211	3	6	4
		S212	4	8	4
		S213	4	6	4
		S214	5	6	5
		S215	2	5	2
	Adlerian	S221	3	5	5
		S222	6	7	5
		S223	3	5	3
		S224	6	8	6
		S225	4	6	3
	Eclectic	S231	6	6	5
		S232	6	7	7
		S233	2	4	3
		S234	6	8	6
		S235	9	9	7

FIGURE 9. *Design 3: Two-between and one-within design*

Main Questions. Of the seven omnibus questions that can be answered by this design, four consider interactions among the factors, and three focus on differences among the main treatments. The four interaction questions are: (1) "Is there a three-way interaction among the two between factors and the within factor?" (2) "Is there a two-way interaction between between factor A and the within factor?" (3) "Is there a two-way interaction between between factor B and the within factor?" (4) "Is there a two-way interaction between between factor A and between factor B?" The three main effects questions are: (5) "Given no interaction, are there differences among the repeated measures?" (6) "Given no interaction, are there differences among between factor A's treatment levels?" (7) "Given no interaction, are there differences among between factor B's treatment levels?"

Example Questions. The omnibus questions answered in this example are: (1) "Is there a three-way interaction among the type of intervention group, counseling prevention intervention methods, and time?" (2) "Is there a two-way interaction between the type of intervention group and time?" (3) "Is there a two-way interaction between the counseling prevention intervention methods and time?" (4) "Is there a two-way interaction between the type of intervention group and the counseling prevention intervention methods?" (5) "Given no interaction, are there differences in alcoholics' perceptions of themselves at different times?" (6) "Given no interaction, are there differences in alcoholics' perceptions of themselves among the types of intervention groups?" (7) "Given no interaction, are there differences in alcoholics' perceptions of themselves among the counseling prevention intervention methods?"

Analyses. The univariate repeated measures analysis is a hierarchical (nested) mixed-model analysis with subjects (random) nested within the levels of the two (fixed) between-treatments factors and the subjects crossed with the repeated measures factor (fixed). The multivariate analysis is a two-way multivariate analysis with contrasts on the repeated measures as dependent variables.

BMDP4V Input Statements. The BMDP4V statements for the third design, shown below, have two BETWEEN factors that identify the two types of groups (control and alcoholic) and the three counseling groups and a WITHIN set to identify the times.

```
/BETWEEN   FACTOR     = TYPE, METHOD.
           CODES(1)   = 1 TO 2.
           NAMES(1)   = CONTROL, ALCOHOL.
           CODES (2)  = 1 TO 3
           NAMES(2)   = ROGERIAN, ADLERIAN,
                        ECLECTIC.

/WITHIN    FACTOR     = TIME.
           CODES      = 1 TO 3.
           NAMES      = TIME1, TIME2, TIME3.
```

Two-Within (Design On the Variables)

Design 4. In the fourth design, a prevention intervention researcher would be interested in asking questions among his or her repeated measures. These questions are like those asked in a two-way completely randomized design. Bock (1975) provides detailed information on the analysis of this design.

Picture. Figure 10 illustrates a design on the repeated measures by creating nine measures on each alcoholic. Here, the authors have reverted to the repeated measures design discussed at the beginning of this chapter and then have repeated the design at Times 2 and 3 (i.e., after prevention intervention). That is, the subjects received scores on measures of their goals related to drinking, work, and family at three different times. This created a 3 (times)x3 (goals) design on the repeated measures.

Main Questions. The questions among the levels of the repeated measures factor indicate that a factorial design was formed among the treatment levels. The questions are: (1) "Is there an interaction among

	Time 1			Time 2			Time 3		
	Drink	Work	Family	Drink	Work	Family	Drink	Work	Family
S1	4	6	5	3	4	3	4	7	5
S2	3	4	3	3	5	3	7	9	7
S3	2	3	2	3	5	3	3	5	3
S4	3	4	2	5	7	4	5	7	4
S5	3	6	4	1	5	2	5	7	4
S6	3	6	4	3	5	5	6	6	5
S7	4	8	4	6	7	5	6	7	7
S8	4	6	4	3	5	3	2	4	3
S9	5	6	5	6	8	6	6	8	6
S10	2	5	2	4	6	3	9	9	7

FIGURE 10. *Design 4: Two-within (design on the variables)*

the repeated measures main factors?" (2) "Given no interaction, are there differences among the first factor's repeated measures?" (3) "Given no interaction, are there differences among the second factor's repeated measures?"

Example Questions. The omnibus questions answered in this example are: (1) "Is there an interaction between time and focus of the scale?" (2) "Given no interaction, are there differences in alcoholics' perceptions of themselves at different times?" (3) "Given no interaction, are there differences in alcoholics' perceptions of themselves among the different scales?"

Analyses. The omnibus analyses performed for this design are the same as those for the single-group design. The difference occurs in how the contrasts are formed. The contrasts are formed and combined so as to reflect the factorial structure on the repeated measures. The univariate repeated measures analysis is a two-way mixed-model analysis with subjects (random) and repeated measures (fixed). The multivariate analysis is a single-group multivariate analysis with contrasts on the repeated measures as dependent variables.

BMDP4V Input Statements. The BMDP4V statements for the fourth design, shown below, have two WITHIN sets to identify the two within factors time and goal measurements.

```
/WITHIN     FACTOR   = TIME, GOAL.
            CODES(1) = 1 TO 3.
            NAMES(1) = TIME1, TIME2, TIME3.
            CODES (2) = 1 TO 3.
            NAMES(2) = DRINK, WORK, FAMILY.
```

One-Between, One-Within, and Multiple Measures at Each Occasion (Doubly Multivariate or Mixed-Model Univariate or Multivariate)

Design 5. The main feature of this design is that multiple measures are taken at each point in time. The multivariate analyses allow researchers to ask questions wherein more than one dependent variable can be used to explain differences among the levels of the within or between factors. Bock (1975) and Robey (1985) provide detailed information on the analysis of this design.

Picture. Figure 11 contains an example of this design. It is the same as design 2 shown in figure 9 with the addition of a second dependent variable. Therefore, in figure 11, there are two dependent variables, one that measures an alcoholic's perception of his or her goals and a second dependent variable that measures how the alcoholics view their current reality with respect to these goals.

Main Questions. Each of the omnibus questions for this design may be answered either by considering each dependent variable alone or as a set. The decision depends upon whether the variables make sense as a set. The questions are: (1) "Is there an interaction between the between and within factors?" (2) "Given no interaction, are there differences among the repeated measures?" (3) "Given no interaction, are there differences among the between factor's treatment levels?"

Example Questions. If the dependent variables are considered alone, then the example questions would be the same as those presented for design 2 but with a set of questions for each dependent variable. The

			Time		
			Time 1	Time 2	Time 3
Counseling method	Rogerian	S11	4 3	6 6	5 4
		S12	3 4	4 8	3 4
		S13	2 4	3 6	2 4
		S14	3 5	4 6	2 5
		S15	3 2	6 5	4 2
	Adlerian	S21	3 3	4 5	3 5
		S22	3 6	5 7	3 5
		S23	3 3	5 5	3 3
		S24	5 6	7 8	4 6
		S25	1 4	5 6	2 3
	Eclectic	S31	4 6	7 6	5 5
		S32	7 6	9 7	7 7
		S33	3 2	5 4	3 3
		S34	5 6	7 8	4 6
		S35	5 9	7 9	4 7

FIGURE 11. *Design 5: Two-between, one-within, and multiple measures at each occasion*

questions illustrated here are for two variables considered together. The omnibus questions answered in this example when both dependent variables are considered together are: (1) "Is there an interaction between the counseling prevention intervention methods and time when both dependent variables are considered together?" (2) "Given no interaction, are there differences in alcoholics' perceptions of their goals and their perceptions of the reality of these goals at different times?" (3) "Given no interaction, are there differences in alcoholics' perceptions of their goals and their perceptions of the reality of these goals among the counseling prevention intervention methods?"

Analyses. The analysis of the *mixed-model univariate* design is the same as for a split-plot design, except that there is an analysis for each dependent variable. Here, care must be taken to control Type I errors. The analysis of the *mixed-model multivariate* design is the same as for a split-plot design, except that there are multiple measures at each occasion and contrasts are formed on the repeated measures to consider both dependent variables. The *doubly multivariate* analysis also is the same as for a split-plot, except that the contrasts formed on the repeated measures consider both dependent variables.

BMDP4V Input Statements. The BMDP4V statements for the fifth design, shown below, have a BETWEEN set to identify the counseling groups and a WITHIN set to identify the times. In the WITHIN set, the term VARIATES indicates that there is more than one dependent variable, and they are identified as PERCENT and REALITY. The BMDP4V program reports all three of the analyses described above; it is left to the researcher to decide which part of the output to use.

```
/BETWEEN   FACTOR    = METHOD.
           CODES(1)  = 1 TO 3.
           NAMES(1)  = ROGERIAN, ADLERIAN,
                       ECLECTIC.

/WITHIN    FACTOR    = TIME, VARIATES.
           CODES(1)  = 1 TO 3.
           NAMES(1)  = TIME1, TIME2, TIME3.
           CODES(2)  = 1, 2.
           NAMES(2)  = PERCEPT, REALITY.
```

Single-Group Repeated Measures Design With a Dynamic Covariate

Design 6. The main feature of this design is the use of a covariate that allows reduction of the error of the repeated measure at each point in time. This covariate is called dynamic because it is measured at each point in time.

Picture. Figure 12 is an example of this design. It is a modification of design 1 (figure 7) with 2 measures and 10 subjects and the addition of a dynamic covariate. The covariate is a measure of locus of control. Locus of control is a measure of a person's perception of their control over events. People are described as having an internal locus of control if they feel that they are in control of events. People are described as having an external locus of control if they feel that things external to them (e.g., luck) are in control of events.

Main Question. The main question for this design is: "Are the measures, absent the effect of the covariate, different?"

Example Question. The omnibus question answered in this example is: "Are there differences in alcoholics' perceptions of themselves at different times, holding constant their locus of control?"

Analyses. The univariate repeated measures analysis is a two-way mixed-model ANOVA with subjects (random) and repeated measures (fixed). The multivariate analysis has not been defined clearly when a dynamic covariate is involved (Davidson 1980).

Time

Subjects	Time 1		Time 2	
	COV	Goal	COV	Goal
S1	9	4	4	6
S2	8	3	7	4
S3	8	2	7	3
S4	7	3	6	4
S5	9	3	6	6
S6	6	3	5	6
S7	5	4	4	8
S8	5	4	4	6
S9	6	5	3	6
S10	8	2	6	5

FIGURE 12. *Design 6: Single-group design with a dynamic covariate*

BMDP4V Input Statements. The BMDP4V statements for the sixth design, shown below, have a WITHIN set to identify the times and the covariates. In the WITHIN set, the term VARIATES indicates that there is more than one variable measured at each time and the variables measured at each time are identified as COVARIAT and PERCEPT.

```
/WITHIN      FACTORS ARE TIME, VARIATES.
             CODES(1) ARE 1, 2.
             NAMES(1) ARE TIME1, TIME2.
             CODES(2) ARE 1, 2.
             NAMES(2) ARE COVARIAT, PERCEPT.
```

One-Between and One-Within (a 2x2 Split Plot) With a Constant Covariate Design

Design 7. The main feature of this design is the use of a covariate that is measured once prior to prevention intervention. This covariate is called *static* because it is measured only once and remains constant over measures. This covariate allows reduction of the error associated with between differences but has no effect on the error associated with the repeated measures (within) differences. Federer and Meridith (1992) provide detailed information on designs of this type.

Picture. Figure 13 illustrates an example of this design. It is a modification of design 2 (figure 8) with 2 measures and 10 subjects in each of 2 groups and the addition of a static covariate. The covariate is the measure of locus of control described for design 6. The prevention intervention groups are composed of alcoholics receiving Adlerian and Rogerian counseling.

Main Questions. The main questions for this design are: (1) "Absent the effect of the covariate, is there an interaction between the between and within factors?" (2) "Given no interaction, are there differences among the repeated measures?" (3) "Given no interaction, are there differences

			Time			
			Time 1		Time 2	
			COV	Goal	COV	Goal
Counseling method	Rogerian	S11	9	4	9	6
		S12	8	3	8	4
		S13	8	2	8	3
		S14	7	3	7	4
		S15	9	3	9	6
		S16	6	3	6	6
		S17	5	4	5	8
		S18	5	4	5	6
		S19	6	5	6	6
		S110	8	2	8	5
	Adlerian	S21	7	3	7	4
		S22	8	3	8	5
		S23	7	3	7	5
		S24	4	5	4	7
		S25	9	1	9	5
		S26	6	3	6	5
		S27	4	6	4	7
		S28	8	3	8	5
		S29	5	6	5	8
		S210	7	4	7	6

FIGURE 13. *Design 7: One-between and one-within design with a constant covariate*

among the between factor's treatment levels absent the effect of the covariate?"

Example Questions. The omnibus questions answered in this example are: (1) "Holding locus of control constant, is there an interaction between the counseling prevention intervention methods and time?" (2) "Given no interaction and holding locus of control constant, are there differences in alcoholics' perceptions of themselves at different times?" (3) "Given no interaction and holding locus of control constant, are there differences in alcoholics' perceptions of themselves among the counseling prevention intervention methods?"

BMDP4V Input Statements. The BMDP4V statements for the seventh design, shown below, have a BETWEEN set to identify the counseling groups and a WITHIN set to identify the times and the covariate. In the WITHIN set, the term VARIATES indicates that there is more than one variable measured at each time, and the variables measured at each time are identified as COVARIAT and PERCEPT.

```
/BETWEEN   FACTOR IS GROUP.
           CODES ARE 1, 2.
           NAMES ARE ROGERIAN, ADLERIAN.

/WITHIN    FACTORS ARE TIME, VARIATES.
           CODES(1) ARE 1, 2.
           NAMES(1) ARE TIME1, TIME2.
           CODES(2) ARE 1, 2.
           NAMES(2) ARE COVARIAT, PERCEPT.
```

Two-Treatment Two-Period Crossover Design

Design 8. The main feature of this design is that it allows the researcher to test for order (period) and carryover effects prior to examining treatment effects. This design is discussed by Fleiss (1986) and Jones and Kenward (1989).

Picture. Figure 14 contains an example of this design. In this design, five of the alcoholics viewed the drinking slides first, followed by the work slides, and five of the alcoholics viewed the work slides first, followed by the drinking slides. This allowed the authors to test what had been found in previous research (i.e., that there were no order or carryover effects among the slide sets).

Main Questions. The three omnibus questions that are answered by this design are the same as those for a split-plot design; however, in this design, these questions are related to questions stated as: (1) "Is there a difference between the times of administration (i.e., a period effect)?"

			Goal	
			Drink	Work
Presentation order	D - W	S11	4	6
		S12	3	4
		S13	2	3
		S14	3	4
		S15	3	6
	W - D	S21	3	4
		S22	3	5
		S23	3	5
		S24	5	7
		S25	1	5

FIGURE 14. *Design 8: Two-treatment, two-period crossover design*

(2) "Given no period effect and no carryover effect, are there differences between treatments?" (3) "Given no period effect, is there a carry-over effect?" The preceding questions are in the same order that they appear for design 2; therefore, the first question here is answered by a test of the interaction, the second question by a test of the within effect, and the third question by a test of the group differences.

Example Questions. The omnibus questions answered in this example are: (1) "Does the order of presentation of the work or drink slides affect the alcoholics' perceptions of themselves?" (2) "Given that the order of presentation of the slides does not affect the alcoholics' perceptions of themselves and that there are no carryover effects, are there differences in alcoholics' perceptions of the drink and work slides?" (3) "Given that the order of presentation of the work or drink slides does not affect the alcoholics' perceptions of themselves, does having seen the drink slides affect the alcoholics' responses to the work slides and vice versa?"

Analyses. The analyses are the same as those for design 2.

BMDP4V Input Statements. The BMDP4V statements for the eighth design, shown below, have a BETWEEN set to identify the groups that received the slides in different orders and a WITHIN set to identify the slide sets.

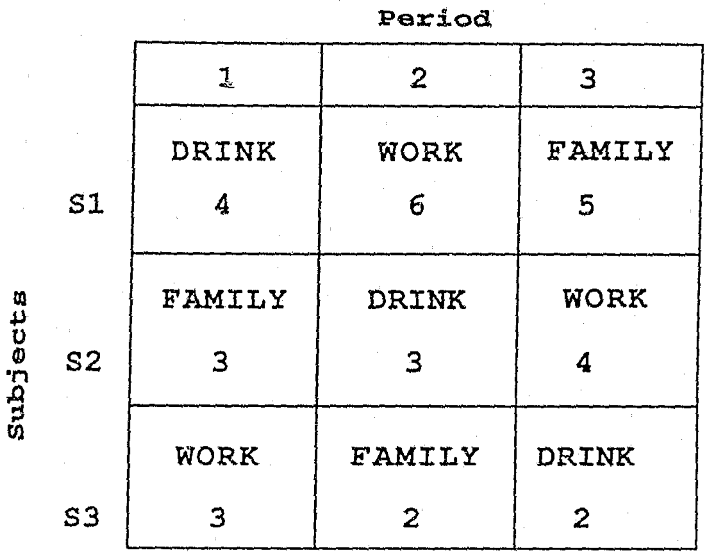
```
/BETWEEN   FACTOR IS GROUP.
           CODES ARE 1, 2.
           NAMES ARE AB, BA.

/WITHIN    FACTOR IS RESPONSE.
           CODES ARE 1, 2.
           NAMES ARE DRINK, WORK.
```

Unreplicated Three-Treatment Repeated Measures Latin Square With No Carryover Effect

Design 9. The main feature of this design is that it allows the researchers to test for differences due to the serial introduction of the treatments (called the period effect). Further examples of the analyses of these designs may be found in Delany and Maxwell (1987), Dowdy and Wearden (1983), and Fleiss (1986).

Picture. Figure 15 contains an example of this design. In this design, three alcoholics view the three different slide sets in three different orders.

Main Questions. The three omnibus questions that are answered by this design are: (1) "Is performance differentially affected by the treatments?" (2) "Are there differences due to the serial introduction of the treatments (period effect)?" (3) "Are there differences among the units?"



Subjects	Period 1	Period 2	Period 3
S1	DRINK 4	WORK 6	FAMILY 5
S2	FAMILY 3	DRINK 3	WORK 4
S3	WORK 3	FAMILY 2	DRINK 2

FIGURE 15. *Design 9: Unreplicated three-treatment repeated measures Latin Square with no carryover effect*

Example Questions. The three omnibus example questions that are answered by this design are: (1) "Are there differences in alcoholics' perceptions of possible treatment outcomes?" (2) "Are there differences in alcoholics' perceptions of possible treatment outcomes due to the serial introduction of the types of slides?" (3) "Are there differences among the alcoholics?"

Analysis. The univariate analysis for a 3x3 Latin Square was used.

BMDP4V Input Statements. The BMDP4V statements, shown below, are appropriate for this design. Notice that this analysis requires no WITHIN statements.

```
/BETWEEN    FACTORS ARE SUBJECT, PERIOD,
            TREATMENT.
            CODES(1) ARE 1 TO 3.
            CODES(2) ARE 1 TO 3.
            CODES(3) ARE 1 TO 3.
```

Why are Repeated Measures Designs Misused?

Repeated measures designs often are misused because investigators are not:

1. Spending time on descriptive analysis of their data, such as examination of:

   a. The structure of the covariance matrix,
   b. Scatterplots for pairs of responses, and
   c. Consideration of the reliability of their instruments;

2. Using the G-G or H-F correction factors in univariate analyses (further details on the correction factors may be found in Cornell and colleagues [1992], Green and Barcikowski [1992*b*], and Robey and Barcikowski [1987]);

3. Using counterbalancing and, when they do, failing to check to see if carryover effects are present;

4. Using both univariate and multivariate analyses in exploratory studies;

5. Using power analysis to help establish sample size (further details on power may be found in Barcikowski and Robey [1985], Green and Barcikowski [1992*a*], Muller and Barton [1989], Muller and colleagues [1992], and Robey and Barcikowski [1984]); and

6. Properly dealing with missing values (further information on this topic can be found in Graham and colleagues [this volume], Little and Rubin [1987], and Schluchter [1988]).

DOES COUNTERBALANCING CONTROL CARRYOVER EFFECTS?

Counterbalancing allows researchers to identify when carryover effects are present, but they have to look for them. Consider the plot shown in figure 16 of response measure by testing time for a design in which counterbalancing was used in studying the benefits of two pills.

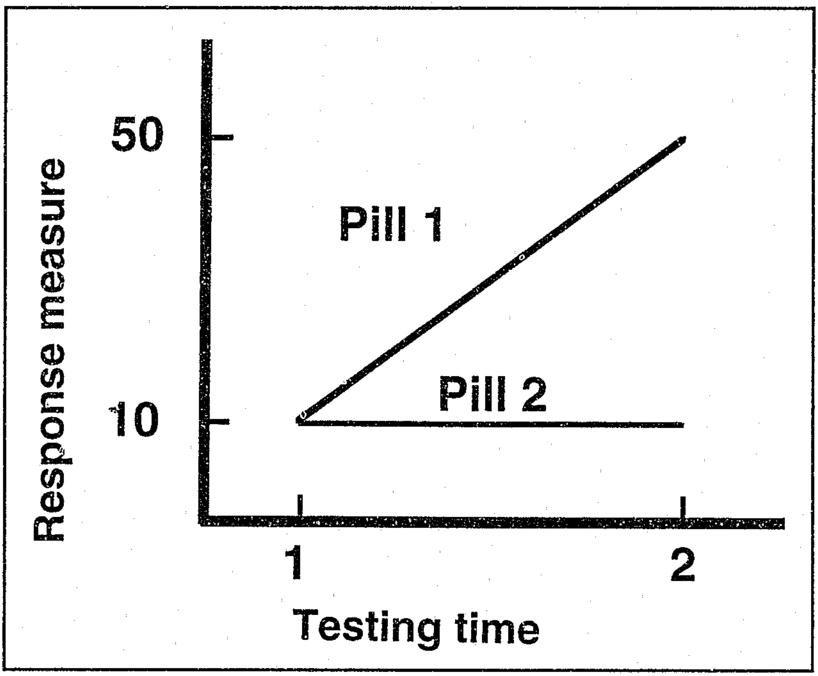


FIGURE 16. *Plot of response measure by testing time*

Here, pill 1 yields a high-response measure only when it is preceded by pill 2. If these data are rearranged for analysis and the fact that counterbalancing was used is ignored, the investigator would conclude that pill 1 was more effective than pill 2.

REFERENCES

Barcikowski, R.S., and Robey, R.R. Decisions in single group repeated measures analysis: Statistical tests and three computer packages. *Am Stat* 38:148-150, 1984.

Barcikowski, R.S., and Robey, R.R. "Sample Size Selection in Single Group Repeated Measures Analysis." Paper presented at the meeting of the American Educational Research Association, Chicago, April 1985.

Bock, R.D. *Multivariate Statistical Methods in Behavioral Research.* New York: McGraw-Hill, 1975.

Collier, R.O., Jr.; Baker, F.B.; Mandeville, G.K.; and Hayes, T.F. Estimates of test size for several test procedures on conventional variance ratios in the repeated measure design. *Psychometrika* 32:339-353, 1967.

Cornell, J.E.; Young, D.M.; Seaman, S.L.; and Kirk, R.E. Power comparisons of eight tests for sphericity in repeated measures designs. *J Educ Stat* 17:233-249, 1992.

Crowder, M.J., and Hand, D.J. *Analysis of Repeated Measures.* London: Chapman and Hall, 1990.

Davidson, M.L. Univariate vs. multivariate tests in repeated-measures experiments. *Psychol Bull* 77:446-452, 1972.

Davidson, M.L. *The Multivariate Approach to Repeated Measures.* BMDP Technical Report No. 59. BMDP Statistical Software, 1980.

Delany, H.D., and Maxwell, S.E. "Alternative Analyses of Latin Square Designs." Paper presented at the meeting of the American Educational Research Association, Washington, DC, April 1987.

Dixon, W.J., ed. BMDP *Statistical Software Manual.* Reprint. Los Angeles: University of California Press, 1985.

Dowdy, S., and Wearden, S. *Statistics for Research.* New York: Wiley, 1983.

Federer, W.T., and Meridith, M.P. Covariance analysis for split-plot and split-block designs. *Am Stat* 46:155-162, 1992.

Fleiss, J.L. *The Design and Analysis of Clinical Experiments.* New York: John Wiley & Sons, 1986.

Games, P.A. Alternative analyses of repeated-measure designs by ANOVA and MANOVA. In: von Eye, A., and Rovine, M.J., eds. *Applied Computational Statistics in Longitudinal Research.* Boston: Academic Press, 1990*b*. pp. 23-38.

Games, P.A. Alternative analyses of repeated-measure designs by ANOVA and MANOVA. In: von Eye, A., ed. *Statistical Methods in Longitudinal Research.* Vol. 1, *Principles and Structuring Change.* Boston: Academic Press, 1990*a*. pp. 81-121.

Gravetter, F.J., and Wallnau, L.B. *Statistics for the Behavioral Sciences.* St. Paul: West Publishing Co., 1985.

Green, S., and Barcikowski, R.S. "Power Estimation in Repeated Measures Analysis of Variance Designs With a Heterogeneous Correlation Matrix." Paper presented at the meeting of the American Educational Research Association, San Francisco, April 1992*a*.

Green, S., and Barcikowski, R.S. "Sphericity in the Repeated Measures Univariate Mixed-Model Design." Paper presented at the meeting of the American Educational Research Association, San Francisco, April 1992*b*.

Greenhouse, S.W., and Geisser, S. On methods in analysis of profile data. *Psychometrika* 24:95-112, 1959.

Huynh, H., and Feldt, L.S. Conditions under which mean square ratios in repeated measurement designs have exact F-distributions. *J Am Stat Assoc* 65:1582-1589, 1970.

Huynh, H., and Feldt, L.S. Estimation of the Box correction for degrees of freedom from sample data in randomized block and split-plot designs. *J Educ Stat* 1:69-82, 1976.

Imhof, J.P. Testing the hypothesis of no fixed main-effects in Scheffe's mixed model. *Ann Math Stat* 33:1085-1095, 1962.

Jones, B., and Kenward, M.G. *Design and Analysis of Cross-Over Trials.* London: Chapman and Hall, 1989.

Keppel, G. *Design and Analysis: A Researcher's Handbook.* 3d ed. Englewood Cliffs, NJ: Prentice-Hall, 1991.

Kirk, R.E. *Experimental Design: Procedures for the Behavioral Sciences.* 2d ed. Belmont, CA: Brooks/Cole, 1982.

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

Looney, S.W., and Stanley, W.B. Exploratory repeated measures analysis for two or more groups. *Am Stat* 43:220-225, 1989.

Maxwell, S.H., and Delaney, H.D. *Designing Experiments and Analyzing Data: A Model Comparison Perspective.* Belmont, CA: Wadsworth, 1990.

Morrison, D.F. *Multivariate Statistical Methods.* 3d ed. New York: McGraw-Hill, 1990.

Muller, K.E., and Barton, C.N. Approximate power for repeated-measures ANOVA lacking sphericity. *J Am Stat Assoc* 84:549-555, 1989.

Muller, K.E.; LaVange, L.M.; Ramey, S.L.; and Ramey, C.T. Power calculations for general linear multivariate models including repeated measures applications. *J Am Stat Assoc* 87:1209-1226, 1992.

Rich, C.E. Repeated measures designs. In: Barcikowski, R.S., ed. *Computer Packages and Research Design With Annotations of Input and Output from the BMDP, SAS, SPSS, And SPSSX Statistical Packages.* Vol. 3, *SPSS and SPSSX.* Lanham, MD: University Press of America, 1983. pp. 567-710.

Robey, R.R. "A Monte Carlo Investigation of Type I Error in the Analysis of Variance for the Single Group Repeated Measures Design With Multiple Measures per Occasion." (Ph.D. diss., Ohio University, 1985). *Diss Abstr Int* 46:05B, 1985.

Robey, R.R., and Barcikowski, R.S. Calculating the statistical power of the univariate and the multivariate repeated measures analyses for the single group case under various conditions. *Educ Psychol Meas* 44:137-143, 1984.

Robey, R.R., and Barcikowski, R.S. "Sphericity Tests and Repeated Measures Data." Paper presented at the meeting of the American Educational Research Association, Washington, DC, April 1987.

Robey, R.R., and Barcikowski, R.S. "Type I Error for the Simultaneous Application of Two Tests for Repeated Measures Data." Paper presented at the meeting of the American Educational Research Association, San Francisco, March 1989.

Rogan, J.C.; Keselman, H.J.; and Mendoza, J.L. Analysis of repeated measurements. *Br J Math Stat Psychol* 32:269-286, 1979.

Rouanet, H., and Lepine, D. Comparison between treatments in a repeated-measurement design: ANOVA and multivariate methods. *Br J Math Stat Psychol* 23:17-163, 1970.

SAS Institute, Inc. *SAS/STAT User's Guide, Version 6.* Vol. 2. 4th ed. Cary, NC: SAS Institute, Inc., 1989.

Schluchter, M.D. Unbalanced repeated measures models with structured covariance matrices. In: Dixon, W.J., ed. *BMDP Statistical Software Manual.* Vol. 2. Berkeley, CA: University of California Press, 1988. pp. 1081-1114.

Stevens, J.P. *Applied Multivariate Statistics for the Social Sciences.* 2d ed. Hillsdale, NJ: Erlbaum, 1992.

SPSS, Inc. *SPSSX: User's Guide.* 3d ed. New York: McGraw-Hill, 1988.

Timm, N.H. *Multivariate Analysis With Applications in Education and Psychology.* Monterey, CA: Brooks/Cole, 1975.

Winer, B.J. *Statistical Principles in Experimental Design.* 2d ed. New York: McGraw-Hill, 1971.

AUTHORS

Robert S. Barcikowski, Ph.D.
Professor of Educational Research and Evaluation
School of Applied Behavioral Sciences and Educational Leadership
College of Education
Ohio University
Athens, OH 45701

Randall R. Robey, Ph.D.
Assistant Professor
Communication Disorders Program
University of Virginia
132 Emmet Street
P.O. Box 9022
Charlottesville, VA 22906-9022

Meta-Analytical Issues for Prevention Intervention Research

Nan Tobler

ABSTRACT

Lack of systematic methods for comparing diversified programs has limited the use of research results. Drug prevention intervention research has a history of mixed or marginal results, a situation that leads to the supposition that programs do not work. Meta-analytical methods have successfully resolved problems of conflicting results and are a cost-effective method for building a knowledge base. Using both qualitative and quantitative methods, meta-analysis applies all of the scientific rigor of primary research to the integration of this research.

Quantitative synthesis is accomplished by computing an effect size, which, unlike significance tests, allows comparisons across studies having varied sample sizes. One advantage for drug prevention intervention research, which seldom shows statistically significant results, is the powerful findings produced when small positive effect sizes are consistent across many studies. Generalizability is possible through meta-analytic aggregation, as a large body of studies contain all the exigencies of real-world research.

Troublesome areas that can distort conclusions are presented to alert readers of literature reviews so they are able to interpret meta-analytic reviews accurately. Specific problematic issues are introduced, such as preexisting differences, combining efficacy and implementation studies, and the use of the weighted effect size with a group of studies that has a large range in sample sizes. Meta-analytic procedures are illustrated by comparing the results of 114 experimental and quasi-experimental school-based adolescent drug prevention programs with a selected subset of 56 higher quality experimentally evaluated programs.

META-ANALYTICAL ISSUES FOR PREVENTION INTERVENTION RESEARCH

The drug prevention field has been marked by conflict on the question of the efficacy of drug prevention programs. Reports have been discouraging and equivocal and have issued the unending call for higher-quality research. However, meta-analytical methods have demonstrated their ability to resolve problems of conflicting results reported in traditional reviews. Cook and colleagues (1992, p. 14) have stated, "No longer is it possible to entertain the pessimistic, simplistic, and energy-sapping hypothesis that 'nothing works'."

The confusion may have resulted from the lack of quantitative and systematic methods for comparing the numerous and varied programs. Traditional literature reviews of prevention intervention[1] research have been narrative and subjective, tending to use limited samples and lacking scientific rigor. The meta-analysis of research refers to a class of methods used to quantitatively integrate and summarize the results of primary[2] research studies.

This chapter presents a conceptual overview of the meta-analytical approach to research review and explains the various statistical procedures used. The purpose is: (1) to alert readers to the pitfalls of the inappropriate use of these procedures, (2) to enumerate the benefits of using meta-analysis when reviewing a body of literature, and (3) to focus on specific areas that present difficulties and complications for using meta-analysis in the field of drug prevention program research. Illustrations will be made from a recently completed meta-analysis of 120 adolescent drug prevention programs (Tobler 1992*a*).

CONCEPTUAL OVERVIEW OF META-ANALYSIS

Meta-analysis is a conceptual approach, not a single technique. All of the scientific rigor used by the primary researcher is applied to the synthesis of results from primary research (Cook et al. 1992, p. viii). The meta-analytic approach encourages a complete and thorough search for research studies to eliminate review bias. Studies should be included from published, unpublished, and fugitive literature and from public or private sponsorships at the local, State, and national levels. Both failed and successful studies should be included. To eliminate subjective bias, studies are coded systematically for all variables known to affect program success. The integration of quantitative results is accomplished by computing the effect size for each study.

An effect size is defined as the difference between the mean of the experimental group and the mean of the control group divided by the pooled standard deviation. Because the effect size has been standardized, comparisons can be made across programs having varied sample sizes. In meta-analysis, studies are the data points. As in a true experiment, an assumption is made that, if enough data points are included, any problems, idiosyncrasies, and/or threats to internal validity associated with program success or failure will be normally distributed. The data points are effect sizes computed from the summary statistics reported by the primary researcher. A primary researcher performs primary analysis of the original data in a research study. High-quality studies are needed to draw reliable conclusions. Recently. dramatic improvement has been made in primary research methodology and the statistical procedures employed to evaluate results, but the interaction of complex research designs and diverse statistical analyses pose problems in calculating effect sizes.

Meta-analysis "is truly an analysis of the results of statistical analyses" (Hedges and Olkin 1985, p. 13). A succession of papers and books by Glass and colleagues (1981), Hedges and Olkin (1985), Hunter and Schmidt (1990), Light and Pillemer (1984), and Rosenthal (1986) have improved reviewers' ability to quantitatively summarize previous research studies more effectively. The new statistical procedures were developed to be used when combining effect sizes from independent studies, in order to avoid the potential problem of arriving at incorrect conclusions resulting from the aggregation of effect sizes.

Meta-analysis must be viewed as a broad-brush approach that provides an overview of the aggregate results of a body of programs. An analogy that can be used is an oil painting. From a distance, Mona Lisa's smile is lifelike but, on close inspection, all that can be seen are small brush strokes. Comparisons can be made across groups of programs (the painting) but not between individual programs (small brush strokes). In a sense, it is an art form and a scientific endeavor. Asking the right question "is up to the meta-analyst" (Gendreau and Andrews 1990, p. 178). As cited in Altman (1990), Bailar states, "Meta-analyses are partly subjective due to decisions about how to carry them out" and "requires a great deal of deep professional judgement about how and what to combine." For example, Cook states that these decisions include: "(1) the studies included for review; (2) the way effect sizes were computed; and (3) a preference for some types of control groups over others within a few studies" (cited in Ingram 1990, p. 68). However, in contrast to traditional literature reviews, these decisions are articulated clearly so that the consequences can be considered when interpreting the results.

ADVANTAGES OF META-ANALYSIS

Coding

Each variable must have a concrete operational definition, that is, instructions to translate the variable into clearly differentiated categories to enable intercoder reliability. Once categorized, subgroups of programs can be studied as units, and examined and compared against each other and against the whole body of programs. Research reports are coded for treatment components and type of outcome measure, as well as program and client characteristics. Program success is determined not by subjective "gut" feelings but through quantitative measures. For example, the effectiveness of programs evaluated with an experimental research design can be compared to programs evaluated with a quasi-experimental design.

Differences as Sources of Information

Inconsistencies in the magnitude of success are the plight of the adolescent drug prevention field, but these differences can provide valuable information. Pillemer and Light (1980) encourage researchers to examine these inconsistencies.

> When study outcomes disagree, it is tempting to throw up one's hands and assume the research is useless. We believe just the opposite: such conflicts can teach us a lot. Looked at positively, they actually offer an *opportunity* to examine and learn about divergent findings (Light and Pillemer 1984, p. 9).

Program outliers (deviant points) also can inform researchers of essential program differences and reveal what type of program should be offered to whom, at what age, and for how long.

Comparison With a Common Metric

In computing the effect size, the results of each study are converted to a standardized score. The effect size, unlike significance tests, is relatively unaffected by sample size. It is known that small measures of program success can reach statistical significance if the sample size is large enough. Therefore, using a standardized score allows direct comparison of the magnitude of the effect size across programs of varying sample sizes (Cahen 1980). This is a pertinent issue in drug prevention since study samples vary from ten to thousands.

Statistical Significance Unnecessary

Adolescent drug prevention studies seldom show statistically significant results, but meta-analytic techniques use all studies regardless of significance levels. Small positive effect sizes that are consistent across many studies can result in a robust finding (Flay 1985*a*). Rosenthal (1990, p. 133) concurs: "Two .06 results are much stronger evidence against the null than one .05; and 10 *p*'s of .10 are stronger evidence against the null than 5 *p*'s of .05."

Not Rejecting Effective Programs (Type II Error)

Meta-analysis is much less prone to reject programs that are effective. Cooper and Rosenthal (1983) conducted an experiment comparing the conclusions reached by traditional reviewers and meta-analytical reviewers. They determined that the met- analytical reviewers reached a clear conclusion more often than traditional reviewers, with less chance of Type II error.

Resolution of Conflicting Results

Meta-analyses have put to rest many questions in other fields (e.g., whether psychotherapy is effective) allowing the field to concentrate on more specific issues (Light and Pillemer 1984). Traditional reviews of drug prevention programs disagree about their general effectiveness. Reviews reporting mixed or marginal results lead to the supposition that drug prevention programs do not work. Yet, for social and medical treatment programs this is to be expected. "One strong finding from various meta-analyses is that most new treatments have, at best, small to modest effects" (Cook et al. 1992, p. 13). Although not statistically significant, as noted earlier, a number of small effects that are consistent in a direction can represent a strong finding.

In order to resolve conflicting results, two other caveats must be considered. First, the program's success must be analyzed on an independent variable that has the possibility of impacting adolescent drug use, the prime dependent variable for the drug prevention field. Even when the same outcome measures are used, a second caveat arises if the grand mean is obtained for a heterogeneous set of studies. In this case, the good programs are lost in the shuffle as the programs with highly positive results will be counterbalanced by programs with negative results, leaving the reader with the impression that none of the programs work. To avoid this pitfall, the meta-analysts can aggregate the highly successful programs and the unsuccessful programs into two groups and inspect each group for commonalties.

Generalizability

There is no debate about the worth of one good experimental study, yet it is not possible to generalize from a single study, as the results could have occurred by chance. Nor is it possible to implement in a single research study all possible variables that influence program success.

> Research findings are inherently probabilistic (Taveggia 1974), therefore, the results of any single study could have occurred by chance. Only meta-analytic integration of findings across studies can control chance and other artifacts and provide a foundation for conclusions (cited in Hunter and Schmidt 1990, pp. 38-39).

Through meta-analytic aggregation, all the various exigencies of doing research under real-world conditions have a possibility of being included. If the other alternative, multiple replications of primary research efforts (a very costly method for building a knowledge base), is used, the results would not be available for many years.

Prospective

Meta-analysis uses previous research to point out the direction for future research. The meta-analyst can quickly identify areas of systematic bias across all studies in a field. Then, funding sources can be alerted so these problems can be addressed by primary researchers. Also, the meta-analytic process reveals areas of missing data that may have been inadvertently not collected or reported by original researchers. Primary researchers and editors can be encouraged to include this information in future reports.

Meta-analytical results also can be used to notify funding agencies of the substantive areas that lack adequate research studies so that these areas can be targeted in new primary research initiatives. For example, results may show that a certain type of program is successful in small *efficacy*[3] trials, while the identical type of program may show negligible results when used in large-scale *effectiveness*[4] trials. This finding would indicate that, even though very costly, implementation factors should be given top priority when executing large-scale effectiveness trials. Research funding for the smaller-scale efficacy studies then could be limited to either innovative approaches or to refinements of existing efficacious programs for new target populations (i.e., minorities).

CRITICISM OF META-ANALYSIS

Publication Bias

An assumption prevails that negative or nonsignificant results are not published and are found more frequently in unpublished literature, such as reports mandated by funding sources or dissertations. Smith (1980*a*), in examining the findings of 12 meta-analyses, found a 33 percent positive bias favoring reports published in journals when compared to dissertations. Smith's (1980*b*) comparison between published and unpublished literature was less clear but also favored published literature. Rosenthal and Rubin (1980) did a comparison of dissertation and nondissertation literature, finding smaller effect sizes for dissertations. A meta-analysis has greater integrity when all areas of the literature are represented.

Selection Bias (Comprehensiveness)

The choice of studies possibly can bias the results of a meta-analytical review more than any other thing. The term "meta-analysis" implies an exhaustive and comprehensive review to the lay reader. If the set of studies included is not representative of the potential universe of studies, the meta-analysis will suffer selection bias. Publication bias is only one form of selection bias. A more insidious form of bias can result from either a limited number of studies or from the reviewer's selection of studies. Cook and colleagues (1992, p. 289) state, "Where conflicting results exist, an advocate can steer a meta-analysis toward the conclusion sought simply by choosing the subset of studies that reach the favored conclusion."

A limited number of studies is the most problematic area for drug prevention program research. There is no mechanism to obtain easy access to a comprehensive body of well-controlled quantitative studies. Bangert-Drowns (1988) located only 33 studies[5] in his meta-analysis of school-based substance abuse education even though he included grade two through college. Only 14 of these programs had drug use measures and only 10 of the 14 programs targeted adolescents. The combination of Bangert-Drowns' sparse sample over such a range in age precludes any valid or reproducible conclusions. Obtaining and coding a comprehensive set of studies is a formidable task and not for the faint hearted.

Quality of the Data

Eysenck (1978) and Gallo (1978) have criticized meta-analysis for including studies that are poorly designed and/or include a wide variety of questionable outcome measures. Studies need not be rejected because they vary in methodological quality or have different outcome measures or because information is incomplete in some areas. Analyses can be made with only those studies that have outcome measures that tap the same conceptual domain. Those studies missing information on a variable would be excluded only for that variable. For methodological issues, studies with strong experimental design can be compared to those with quasi-experimental designs. Glass and colleagues (1981) found a maximum of 0.1 standard deviation between high- and low-validity experiments. Both Lipsey (1992, pp. 118-121) and Tobler (1992*a*, p. 48) found that random assignment versus nonrandom assignment was not associated with effect size but did verify other design factors that were more highly associated with effect size, such as attrition and initial nonequivalence. A judgment can be made on an empirical basis to include studies with weaker designs if they are not seriously flawed and if these factors can be controlled for in regression analyses.

Apples and Oranges

Many theorists believe heterogeneous studies should not be included in a meta-analysis. This is a valid criticism if the aggregate body of studies is so small that only a few programs of each type are included. If the body of studies is large enough and the independent and dependent variables are given concrete operational definitions, the problem no longer is one of mixing apples and oranges but becomes one of identifying Mcintosh apples from Spies apples.

Identical Methodological Flaws

Meta-analysis is based on the assumption that biases are balanced across studies. For example, it is expected that a large group of studies will have different flaws (i.e., weak data analysis, poor representative sample, or weak internal validity). Some flaws may inflate the effect sizes, but these are counterbalanced by an underestimated effect size for other studies (Cook and Leviton 1983). Ideally, these problems would vary across the body of studies. A meta-analyst can code for methodological issues to determine if the same methodological flaws bias all the studies. Recent reviews of smoking prevention programs report that many past studies contain the same threats to internal validity (Botvin and Wills 1985; Flay 1985*a*, 1985*b*). If a specific problem exists across all studies, then the results must be viewed with this in mind.

STATISTICAL PROCEDURES USED TO INTEGRATE STUDIES

Only methods of integrating studies based on between-group mean differences (i.e., effect sizes) will be included. Correlational measures are not discussed, as they seldom are reported in outcome studies of drug prevention programs. Copper (1984) and Hunter and Schmidt (1990) discuss methods for integrating the correlation coefficients. Also excluded are aggregation methods based on vote-counting or combining probabilities (Rosenthal 1986, p. 102).

Unweighted Effect Size

Glass and colleagues (1981) define effect size as: ES = $(\overline{X}_e - \overline{X}_c)/SD_c$, where ES = effect size, $\overline{X}_e$ and $\overline{X}_c$ are the means for the experimental and control groups, respectively, and SD_c is the control group standard deviation. In drug prevention research, parametric statistics are reported[6] that are computed using the pooled standard deviation. To keep effect sizes comparable, it is more appropriate to use statistics that use the pooled standard deviation, such as Cohen's *d* or its equivalent, Hedges' *g*. Also, the pooled standard deviation tends to provide a better estimate of the population standard deviation (Rosenthal 1986, p. 22).

An effect size of 1 is equivalent to an improvement of one standard deviation for the experimental group when compared to the control. The magnitude of effect sizes would not be expected to exceed 1 or 2 but can range from minus infinity to plus infinity. Practically, Cohen (1977) defined an effect size of 0.2 as small, 0.5 as medium, and 0.8 as large.

Weighted Effect Size

Hedges (1982) and Rosenthal and Rubin (1982) independently formulated a weighted effect size to be used for statistical aggregation. Each study's effect size is weighted by the inverse of its variance:

$$W_i = 1/V_i \tag{1}$$

where W_i = weighting factor of the study and V_i = variance of study. Hedges' formula (1986, p. 739) for the weighting factor of an individual study is:

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

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.

The weighted average (d.) of $d_i \ldots d_k$ effects is given by:

$$d. = \sum_{i=1}^{k} w_i d_i / \sum_{i=1}^{k} w_i \qquad (3)$$

Use of weighted effect sizes is based on the fact that larger samples produce more stable results. Plotting effect size against sample size showed a triangular distribution with the effect sizes of the largest programs varying only slightly, producing a smaller standard deviation (see figure 1). The standard deviation for the 42 smallest programs was .353, compared to .096 for the six largest programs.

Corrections for Bias in Effect Sizes

Hedges and Olkin's (1985) correction factor can be used to obtain an unbiased estimator for small samples under 20. They also provide corrections for measurement error and validity of response measures (see also Hunter and Schmidt 1990). In some cases, these corrections should be made. Lipsey (1992) very appropriately corrected for errors in measurement. Lipsey's study used official reports of arrests, probation violations, and reconvictions as a dependent variable; reliabilities between .20 and .30 would cause a deattenuated mean effect of .20 (Lipsey 1992, p. 98).

Many community-based programs use reactive measures (i.e., reports to therapists and actual behavioral measures); therefore, correction for bias should be planned when computing the effect sizes at the individual study level. On the other hand, corrections for test-retest reliability were not included in Tobler (1992*a*) because confidential self-reports of drug use have test-retest values of .76 (alcohol) to .90 (cigarettes) at 1-year follow-up (O'Malley et al. 1983, p. 813). The unweighted effect would be attenuated by only .028 for alcohol and .008 for cigarettes.[7] Corrections this small are meaningless when compared to the potential errors in computing effect sizes from the varied summary statistics.

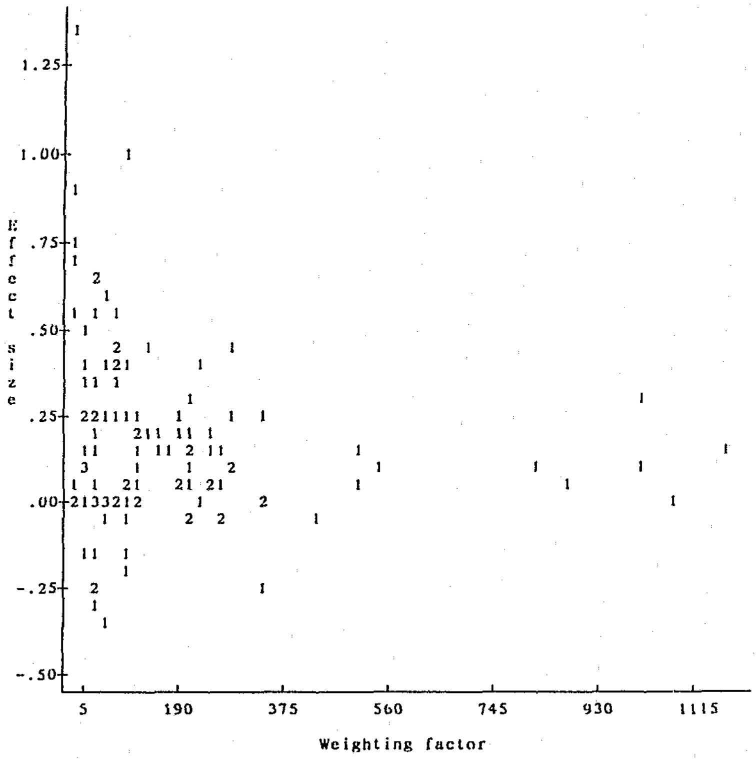


FIGURE 1. *Effect size by the weighting factor*

Tests for Homogeneity of Effect Size

Hedges (1982) and Rosenthal and Rubin (1982) independently developed tests for homogeneity of effect size. The importance of showing that the total set of studies share a common underlying effect size is illustrated by the following: "If half of the studies had a large positive population effect size and half of the studies had a negative population effect size of equal magnitude, then the average—zero—is not representative of the effect size in any of the studies" (Hedges 1986, p. 740). If the overall homogeneity statistic (Q_T) or *total fit* exceeds what can be expected from sampling error alone, then heterogeneity of effect size exists. Although, for social programs, heterogeneity is to be expected (see discussion of mediating variables in the next section), excessive heterogeneity suggests that the studies do not represent a single population parameter. This indicates that widely divergent programs have been included and alerts the meta-analyst to subdivide the total set of programs to obtain groups with more homogeneous effect sizes.

Tests for Between-Group Differences

"Grouping studies into overly broad categories and calculating a mean effect size for each group could wash out real variations in treatment within the categories" (Hedges 1986, p. 745). To avoid this, the researcher usually forms groups of conceptually similar studies and proceeds to test for homogeneity of effect size within the groups. In other words, the meta-analyst subdivides the entire set of programs on a potent variable that would produce a smaller *within-class fit* (Q_W). If the within-class fit indicates homogeneity of effect size, then further analyses are not needed. If further subdivisions do not reduce the heterogeneity of effect size, tests for *between-class fit* (Q_B) should be made. If both between-class fit and within-class fit are rejected, then many mediating or moderating variables may be operating, and continuing to subdivide the studies on other important predictors will quickly reduce the number of studies in the cells to 0 or a number too small to produce reliable results.

These procedures are similar to using analysis of variance (ANOVA) in primary analysis. When using effect sizes, an analogy to ANOVA must be used, as the assumptions of ANOVA may not be met (Hedges and Olkin 1985, p. 148). The relationship for the analogy to ANOVA is:

$$Q_T = Q_B + Q_W \qquad (4)$$

where Q_T = the total fit, Q_B = the between-class fit, and Q_W = the within-class fit. Q_T, Q_B, and Q_W are distributed as chi-square variables (Hedges and Olkin 1985, p. 156). The degrees of freedom (df) associated with Q_T, Q_B, and Q_W are, respectively, *k*-1, *p*-1, and *k-p,* where *p* is the number of groups and *k* is the number of programs.

Multiple Regression Analysis for Effect Sizes

As most social science research involves numerous predictor variables, ANOVA cannot be used effectively. As cited in Cook and colleagues (1992, p. 22), Campbell (1957), Campell and Stanley (1966) and Cronbach (1982) emphasize that "the diversity typically found among people, settings, and historical climates creates a unique context for each study. This study-specific context then somehow transforms the 'meaning' of treatments that, on the surface, appear identical, setting in motion unique causal processes with various of the populations, settings, and times studied." The numerous predictor variables documented in the drug prevention literature suggests that multiple regression methods would be more appropriate. When using ordinary multiple regression procedures with effect sizes, the assumption of homogeneity of variance usually is violated. Hedges and Olkin (1985, pp. 162-188) developed an analog to multiple regression analysis called weighted multiple regression. The primary advantage in using weighted multiple regression procedures is that they can be used to simultaneously examine numerous predictors.

Tests for Model Specification

Finally, the regression equation can be tested for model specification, which determines "whether significant systematic variation in effect sizes remains unexplained" (Hedges 1986, p. 743). In a correctly specified regression model, the proportion of variance accounted for by the residual or error sum of squares (Q_E) would equal what would be expected from sampling error alone. Q_E is tested with the "chi-square distribution with k-p-1 degrees of freedom, where p is the number of predictors not including the intercept" (Hedges and Olkin 1985, p. 174). In this case, the regression sum of squares (Q_R) would account for all the systematic variation in the effect sizes. The meta-analyst must remember that, although tests may indicate that the model has been specified correctly, this does not guarantee that the entire model has not been affected by the same design flaw in each of the individual studies, thereby consistently biasing all the effect size estimates.

Types of Meta-Analysis

Similar to primary research analyses, meta-analysis can be used to investigate relationships or to test a specific a priori hypothesis. The meta-analysis presented in Tobler (1986) is an example of an *exploratory* meta-analysis in which a wide net was cast to include a variety of programs for purposes of identifying relationships. Inferences flowed from program outcomes to types of treatment. In Tobler (1992*b*), a reanalysis[8] was made of 91 programs (a subset of the original 143 programs) that measured change solely on drug use outcome measures. These exploratory meta-analyses (Tobler 1986, 1992*b*) laid the groundwork for the development of specific hypotheses. Tobler (1992*a*) is an illustration of a meta-analysis designed to test specific *hypotheses*. The relationships that evolved in the exploratory meta-analysis were tested with a priori planned comparisons in Tobler (1992*a*). When testing a specific hypothesis, the direction of inference is opposite of that found in an exploratory meta-analysis. "A hypothesis asserts which treatment is most effective: a review then examines empirical evidence to test the hypothesis" (Light and Pillemer 1984, p. 27).

Cook and colleagues (1992) moved meta-analysis to a new level to include explanation. Although making no causal inferences, the methods in Cook and colleagues (1992, p. ix) "explicitly confront the difficult question of how to use meta-analytic techniques to address issues of explanation." This type of meta-analysis has yet to be used in drug prevention program research.

At no time does the meta-analyst manipulate the independent variables, the benchmark of a true experiment. Meta-analyses always are correlational analyses. Yet, Schmidt (1992, p. 1178) states, "The relationships revealed by meta-analysis—the empirical building blocks for theory—can be used in path analysis to test causal theories even when all the delineated relationships are observational rather than experimental."

ILLUSTRATION: META-ANALYSIS OF ADOLESCENT DRUG USE PREVENTION PROGRAMS

The schools have responded to the societywide epidemic of alcohol and other drug (AOD) use by adolescents by implementing drug prevention programs that are intended to delay, retard, or reduce AOD use among teenagers. Although not extensive, there now exist enough research studies to investigate the relative effectiveness of these different types of drug prevention programs using meta-analysis.

The author's latest meta-analysis (Tobler 1992*a*) will be used to illustrate the methodology enumerated earlier. Meta-analytical procedures are emphasized with substantive issues briefly included to illustrate how they influenced meta-analytic decisions. To distinguish between the two meta-analyses, they will be referred to as 1986 and 1992*a*. The 1986 meta-analysis was reported in two published articles: Tobler (1986) and Tobler (1992*b*). Initial analyses of the second and latest meta-analysis were reported in Tobler (1992*a*). The second meta-analysis examines the relative efficacy of varied types[9] of school-based adolescent drug prevention programs (5th-12th grade) for their success in reducing cigarette and AOD use. Types of drug prevention programs were compared for their differential effectiveness with diverse target groups, different drugs, varied program implementations, types of leaders, and on the strength of the research design.

Decisions Reviewers Should Make Before Conducting a Meta-Analysis

Many of the criticisms of meta-analysis can be avoided by thoughtfully selecting studies. Unless the number of evaluation studies is severely limited or an exploratory meta-analysis has not been conducted in the field of interest, the inclusion of widely divergent programs leaves the meta-analyst vulnerable to the problem of mixing apples and oranges. It is through the selection criteria that: (1) the dependent and independent measure‘ are defined, (2) the characteristics of the targeted population are determined, and (3) issues of research design are specified. Knowledge of the substantive and meta-analytical literature informs the meta-analyst about factors to examine for association (not causation) with programs' success and/or failure.

A second type of decision involves the methods to be used to aggregate the selected programs. Any outstanding differences between programs not addressed by their selection must be resolved before aggregation. Despite stringent selection procedures, the final set of drug prevention programs will have used different research methodology, had an extremely varied sample size, targeted different drugs, and included developmentally different ages. Data analysis procedures should be chosen for their ability to address the specific variations among the programs. How these differences are resolved influences the substantive meaning of all the subsequent analyses of program efficacy.

Problems Resolved Through the Selection of Studies

Primary Dependent Variable: Drug Use. Do drug prevention programs work? Two independent research questions are intermingled in this question. The first question is: "How is program success defined?" In other words, "What is the dependent variable?" The second question is: "What type of drug prevention program works?" This question defines the primary independent variable.

Isolating the first question leads to an investigation of the types of outcome measures used to evaluate the effectiveness of drug prevention programs. In Tobler (1986), five categories of outcome measures were identified: knowledge, attitudes, self-reported drug use, skills, and behavior. Each outcome category was distinct and measured the success of the program on conceptually different measures. For example, delaying the onset or reducing current drug use is a much more difficult task than achieving changes with knowledge outcome measures (Bangert-Drowns 1988; Brunvold and Rundall 1988; Tobler 1986). As some outcome measures potentially inflate a program's success and outcome measures are not uniform across the individual programs, the success of a program should not be averaged across different types of measures. This would be a case of mixing apples and oranges. The type of outcome measure contributed the largest increment to R^2 in the regression analyses reported in the 1986 meta-analysis.

Only programs using drug use outcome measures were included in this meta-analysis. This eliminated the confounding effect of including many types of outcome measures. The drug use outcome measures were self-reported paper-and-pencil tests given confidentially in a classroom setting and often accompanied by physical tests (i.e., saliva). The reliability of confidential self-reports of cigarette use has been documented (Murray et al. 1987; Pechacek et al. 1984). The reliability of an adolescent's self-report of licit and illicit drug use was verified by parents and best friends in a study of 8,206 New York State secondary school students (Single et al. 1975). Cited in Oetting and Beauvais (1990, p. 386), Johnston and O'Malley (1985) found that the most reliable self-report questions were lifetime use and use in the past month. Oetting and Beauvais (1990) also cite a number of other studies that validate the self-report measures.

The other dependent variables conceptualized as measures of program success were: knowledge, attitudes and values, refusal skills, generic skills, school-related behaviors, psychological well-being, and nondrug-related measures. Effect sizes also were computed for these variables in addition to the drug use outcomes and will be reported in later analyses to determine their association with decreases in drug use.

Primary Independent Variable. Once the ultimate measure of program success has been determined, the second question can be addressed: "What kind or type of program works?" The past confusion and pessimism about the effectiveness of drug prevention programs may be the result of focusing ONLY on whether they worked. The research was not designed to examine the more important question of which type works with whom. In a landmark review, Schaps and colleagues (1981) defined a topology of drug prevention programs. This topology directed this author's search for quantitative reports to include in the meta-analysis of 143 adolescent drug prevention programs (Tobler 1986). Careful coding for content components helped to eliminate the "black box" regarding program content and resulted in identification of five major types of prevention programs. The definition of the type of program evolved after examining the clustering of content components. In this meta-analysis, the program content was combined with how the content was delivered to more accurately capture what actually happens in the classroom.

Two decisions were made to identify the *type* of program. The first decision about the nature of the program content was made from the various combinations of 30 content items that were grouped into the seven major content areas: knowledge, affective, refusal skills, generic skills, safety skills, extracurricular activities, and others (see table 1). The content areas called extracurricular activities and others occurred very infrequently and subsequently were dropped. Using the five remaining major content areas, a determination was made about which of the various combinations of content best portrayed the program.

The second choice concerned the *process* or delivery method. This was ranked on a continuum beginning with little or no peer interaction (i.e., didactic presentations) and progressively including greater amounts of peer interaction between the group members. Noninteractive programs occupy the first half of the continuum, and Interactive programs occupy the latter half. Four categories of group processes were identified: A, B, C, and D. A group classification system based upon Toseland and Rivas's topology (1984, p. 20-22) was tailored specifically to describe the classroom processes operating in school-based drug prevention programs. A thorough discussion of both the content and the types of groups can be found in Tobler (1992*a*).

Once the decisions about the content and the type of group were made, the two dimensions were combined into the various combinations shown in table 2. The left side of table 2 represents the best choice for content, and the right side is the best choice for the type of group. Twenty-six distinct combinations of content and process actually were located and consolidated into six types of programs. These were collapsed into two overarching groups: Noninteractive programs and Interactive programs. For purposes of this chapter, only the results for the two overarching groups, Noninteractive and Interactive, will be reported.

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

KNOWLEDGE

Knowledge of drug effects
Knowledge of media and social influences
Knowledge of actual drug use by peers (i.e., 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
Support systems/networking with nondrug-using adolescents
Cognitive behavioral skills

GENERIC SKILLS

Communication skills
Coping skills
Social/dating skills
Assertiveness skills
Decision/problem-solving skills
Goal-setting
Identifying alternatives

SAFETY SKILLS

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

EXTRACURRICULAR ACTIVITIES

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

OTHER

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

SOURCE: Tobler (1993)

TABLE 2. *Type of program by content and process*

CONTENT	PROCESS
NONINTERACTIVE	
KNOWLEDGE-ONLY	
Knowledge	Group A
Knowledge	Film/theater
Knowledge	Group A
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 (ECM) training for teachers

** Total community effort supporting the school-based program

SOURCE: Tobler (1993)

Inclusion of Quasi-Experimental Research. Ideally, meta-analysis should include only experimental studies (i.e., random assignment of students), but many well-known and widely used curricular packages have been evaluated only with quasi-experimental designs. A second factor to consider is the difference between small experimental and large quasi-experimental research; in the 1986 meta-analysis, the experimental studies averaged 537 participants, and the quasi-experimental studies had an average of 3,676 participants. Inclusion of both types of research designs balances the rigor of small efficacy studies with the generalizability of large real-world, school-based implementation studies.

One of the problems associated with quasi-experimental research is the possibility of preexisting differences. To alleviate this problem, the quasi-experimental studies should be included only if they have pretest and posttest results for both the treatment and comparison group, so that effect sizes can be computed from change scores.

Inclusion of Only School-Based Programs. Tobler (1992*a*) was limited to the "normative adolescent" and, therefore, addresses only questions about the efficacy of school-based drug prevention programs that included all ethnic groups attending the school. Programs for high-risk youth[10] and/or youth identified as exhibiting abusive or compulsive drug use behaviors were not included, as their drug use etiology necessitates multimodal and markedly different types of prevention programs (Bry 1982; Hawkins et al. 1987; Swisher and Hu 1983; Wall et al. 1981).

Other Selection Criteria. Grades 6 to 12 were included (5th grade also was included if it was in the middle school or if longitudinal research was conducted). Programs with goals of primary prevention or secondary prevention or early intervention were included; treatment programs were excluded. Only studies located in the United States and/or Canada were included. Unpublished as well as published reports were included for the period from 1978 to 1990. As adolescent drug use peaked in 1978, this choice should reflect the downward societal trends (Johnston et al. 1985, 1989).

Selection of a Subset of Higher Quality Experimental Studies. The purpose of selecting a special subset of programs is threefold: (1) to verify the major findings in Tobler (1992*a*) using a reduced set of relevant variables as covariates so that the number of parameters are more in line with the number of cases, (2) to reproduce the results obtained for the 114 programs that included quasi-experimental studies with a selected set of higher-quality experimental studies, and (3) to eliminate studies in which effect size could have been underestimated or overestimated for other methodological reasons.

As stated earlier, many prevention programs are widely disseminated even though they are evaluated with an often questionable, quasi-experimental research design (Klitzner 1988). Many researchers feel that results of programs evaluated with quasi-experimental research designs yield *overestimates* of program effects. Therefore, an analysis was made of a subset of higher quality experimental studies to *empirically* examine this question. This is of prime importance for this field as the number of well-controlled studies is limited.

Other factors can influence the evaluation results of a program besides random or nonrandom assignment. Therefore, studies were included in the higher quality experimental set only if the program: (1) had a delivery intensity of not fewer than 4 hours (i.e., 1 week of classes); (2) administered a posttest not fewer than 3 months after pretest; (3) 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); (4) was not compared to another treatment program; (5) had followed individuals in longitudinal research (i.e., no cross-sectional research); and (6) 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).

Comprehensiveness of the Sample of Studies. An extremely comprehensive literature search was conducted for unpublished and published reports from public or private sponsorships at the local, State, and national levels. The sample of studies built upon the previously located published and unpublished reports in Tobler (1986). A computer search was made of all the relevant data bases with emphasis being placed on the period since 1983. Letters requesting unpublished reports and identification of other persons to contact about research studies were sent to the 60 directors of the National Association of State Alcohol and Drug Abuse Directors. Another 60 members of the National Prevention Network were contacted by letter. Ninety-nine speakers, panelists, and roundtable presenters at the First National Conference on Alcohol and Drug Abuse Prevention, held on August 3-6, 1986, in Arlington, VA, were contacted by letter or phone. The 96 members listed in the Resource Directory of National Alcohol-Related Associations, Agencies, and Organizations were contacted with a letter. All 17 Fiscal Years 84-87 National Institute on Drug Abuse (NIDA) grant recipients were contacted by letter or phone, as well as another 25 researchers. Approximately 75 phone calls were made as a result of recommendations referenced on the returned postcards. Sixteen dissertations, unretrievable through interlibrary loan, were purchased. Searches were made of all traditional literature reviews and of the bibliographies of newly located studies.

Problems To Resolve Before Aggregating Programs

Coding Procedures. A 50-page codebook included over 250 variables related to: (1) treatment components (see tables 1 and 2); (2) participant characteristics (e.g., grade, sex, ethnicity, and socioeconomic class); (3) program characteristics (e.g., year, source of publication, goal, targeted drug, funding, location, number involved, number tested, and research center); (4) implementation factors (e.g., intensity, duration, boosters, leaders, and hours and type of leader training); (5) research methodology (e.g., sampling, assignment, unit of assignment, type of control group, research design, and threats to internal validity); (6) test instrumentation (e.g., reliability, test-retest, internal consistency and reactivity of mea-sure); and (7) data analysis (e.g., unit of data analysis and method of effect size calculation). In coding studies, the main focus was in 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 (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*, appendix 3) and were spot checked by the principal investigator.

Unit of Analysis: A Program. A program is the unit of analysis. In meta-analysis, studies most often are the unit of analysis, with one effect size being reported per study (Bangert-Drowns 1986). However, 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 decision-making 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 also was necessary to ensure that only one effect size was contributed to the overall analyses for 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 ensure independence of a sample of students *all* authors were cross-checked against all other authors in the data base for the purpose of identifying duplicate reports on the same group. Sets of articles or reports then were 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 were reported for a program on the *same* outcome measure, 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 was 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 0 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 (i.e., 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. However, 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, or level of drug use (i.e., nonuser, experimental user, or 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 outcomesx2 sexesx 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 boys and a negative effect for girls, the *mean* effect for the program is 0 and does not accurately portray the program's results. Bangert-Drowns' (1986) study effect method (one effect per program) does not take into account differential results across subpopulations. Because the weighted effect size was used, the weighting factors for the individual subpopulations also were combined into a single weighting factor for the program. However, in this case, the *sum* of the individual subpopulation weights were computed to be used at the aggregate level (see table 3).

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: (1) 1-12 months, (2) 13-24 months, (3) 25-36 months, and (4) 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 followups 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, first analyses were 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 also was found to be nonsignificant. Hand

TABLE 3. *Computation of aggregated effect size per program: First aggregation for subpopulations effect size by level of smoking*

	Time 1		Time 2		Time 3		Time 4	
	ES*	WTF**	ES	WTF	ES	WTF	ES	WTF
Nonsmokers/girls	0.14	52.6	0.32	34.5	-	-	0.07	26.3
Nonsmokers/boys	0.23	41.7	0.19	25.0	-	-	0.12	18.5
Subtotal	0.18	94.3	0.26	59.5	-	-	0.10	44.8
Exper/girls	-0.13	55.5	-0.13	47.6	-	-	-0.16	34.5
Exper/boys	0.18	52.6	0.36	37.0	-	-	0.15	30.3
Subtotal	0.03	108.1	0.12	84.6	-	-	-0.01	64.8
Users/girls	0.34	22.2	0.26	21.3	-	-	0.01	30.3
Users/boys	0.32	38.5	0.11	38.5	-	-	0.22	34.5
Subtotal	0.33	60.7	0.19	59.8	-	-	0.12	64.8
TOTAL	0.18	263.2	0.19	203.9	-	-	0.07	174.4

Example: Time 1 mean effect size
Total Effect Size Time 1 = [Nonsmokers+Experimental+Users]/3

Example: Time 1 sum of the weighting factors
Total WTF Time 1 = WTF Nonsmokers+WTF Experimental+WTF Users

KEY: * ES = Effect size
** WTF = Weighting factor

inspection showed equal numbers of programs reported increases in effect size over time as programs that showed decreases in effect size over time. Third, scatterplots of 118 programs[11] 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 (see table 4). This method maintains the statistical independence for each program.

Choice of Covariate-Adjusted Means. Effect sizes usually are computed on the final *unadjusted* posttest results (Glass et al. 1981; Smith et al. 1980). Unadjusted means can be used only when random assign-ment resulted in truly equivalent treatment and control groups. Undoing the covariate-adjusted scores to obtain the unadjusted means as proposed by Glass and colleagues (1981) and Smith and colleagues (1980) 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 are computed from the *covariate-adjusted means* reported by the researcher. Also, including quasi-experimental (nonrandom assignment) studies necessitates working with change scores, as 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 final consideration, test-retest reliabilities are needed to compute unadjusted posttest scores whether analysis of covariance summary statistics or pretest/posttest means and standard deviations 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) but 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

TABLE 4. *Computation of aggregated effect size per program second aggregation across time*

MEAN EFFECT SIZE PER PROGRAM ACROSS TIME

ES* = [Effect Size Time 1+Effect Size Time 2+Effect Size Time 4]/3

ES = [0.18+0.19+0.07]/3

ES = 0.144

WEIGHTING FACTOR ACROSS TIME

MWF** = [WTF Time 1+WTF Time 2+WTF Time 4]/3

MWTPAT = [263.2+203.9+174.4]/3

MWTPAT = 213.8

KEY: * ES = Effect size per program across subpopulation and time
** WTF = Weighting factor

NOTE: ES and WTF are taken from the example in Table 6.

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 numerically is approximately four times smaller than the number of participants in the program. Twenty-one programs had weighting factors under 25 (i.e., 100 students or fewer), while six programs had weighting factors near or above 1,000 (i.e., 4,000 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.

Use of Homogeneity of Effect Size. Tests for homogeneity of effect size for the entire set of 120 programs showed extreme heterogeneity, nearly six times that expected from sampling error. The windsorized sample of 114 programs was still 3.5 times more heterogeneous than would be expected. The problem of the larger programs overshadowing the smaller ones still existed. To obtain a more homogeneous set of programs, the 120 programs arbitrarily were divided into four size groups based on natural groupings seen on the histogram. Additionally, the sample was reduced to 114 programs after removal of six outliers identified in the regressions.

Q_T, the critical value ($Q_{critical}$), and the ratio of Q_T to $Q_{critical}$ can be found in table 5 for each subset and the total set of 114 programs. Subdividing into four groups further reduced the problem as seen in the reduction of Q_T in the subgroups. More importantly, the conceptual issues of comparing smaller programs, often efficacy studies, to larger implementation studies was alleviated. Each size subset represents an independent meta-analysis: one for the smaller programs, one for the medium, one for the large, and the last for the six extremely large-scale implementations. For each size group, there could be distinctly different findings, as each represents a totally different set of studies. Ideally, the results for the overall set of 114 programs should mirror those in the size groups. If repetition occurs across size groups and then again in the overall analyses of 114 programs, it becomes a very powerful finding. Any single significant findings not replicated should act only as an alert to examine the nature of programs in that subgroup. The results of the analyses by size group are included in Tobler (1992*a*).

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

TABLE 5. *Test for homogeneity of effect size*

Set	Number	Q_{total}	$Q_{critical}$	Ratio $Q_{total}/Q_{critical}$
All	120	895	158	5.7
All * Windsorized	114	534	151	3.5
56 Experimental	56	277	82	3.4
Size One	42	154	71	2.2
Size Two	32	193	53	3.6
Size Three	34	157	56	2.8
Size Four	6	49	17	2.9

KEY: * Six outliers removed from original sample

with change scores is discussed in Tobler (1985). Where parametric statistics were reported, the effect sizes were calculated using formulas documented in Tobler (1992*a*, appendix 3). When reports did not provide the exact *p* value but stated only that the results were significant, a .05 level of significance was assumed, and the corresponding *t* levels were computed. If only a statement of nonsignificance was reported, a *p* value of .50 was assigned, that is, an effect size of 0. This is a conservative method for estimation of effect sizes. Had researchers given the actual *p* value, even though not significant, it would have led to an effect size greater than 0.

DATA ANALYSES

Ordinary and Weighted Least Squares Regressions

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

In order 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 included follows in the next section.

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 leaders. The type of program (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; their dummy coding is explained in the following 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 df for a variable are (*k*-1); therefore, a binary variable uses one df. Three df 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. The equation is:

$$\hat{Y} = a+b_1X_1+b_2X_2+b_3X_3+b_4X_4+[b_{51}X_{51}+b_{52}X_{52}]+ [b_{61}X_{61}+b_{62}X_{62}+b_{63}X_{63}]+b_IX_I \quad (5)$$

where a = regression constant, b = regression coefficient, X_1 = covariate one (X_2 = covariate two, etc.), X_I = primary independent variable, and $\hat{Y}$ = predicted criterion variable.

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 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 importance 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., by partialing out the effect of all the covariates).

Hypothesized Covariates Eliminated Due to Limited Numbers

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 all were eliminated as covariates because only a limited number of programs reported results broken out for this information (frequencies are reported in Tobler 1992*a*).

Hypothesized Covariates Eliminated in Previous Analyses (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 in 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 or another treatment (Tobler 1986, 1992*a*). The reference category was treatments compared to a standard 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 fincings reported in Tobler (1992*a*), which are detailed below.

A decision tree was used in Tobler (1992*a)* that involved three choices: assignment, attrition, and differential dropout (see figure 2). The first decision concerned the method of assignment, random or nonrandom. Ideally, random assignment produces groups that are equivalent on important individual characteristics.

The second choice involved a:trition (i.e., experimental mortality), which usually is not a threat to internal validity unless there is differential dropout from treatment or control groups. The threats to internal validity, however, become more problematic if attrition is extreme. School-based drug prevention studies do not retain students for various reasons: transfer, absenteeism, and dropping out of school. Attrition seldom is the result of a student's choice to leave the program, which would be a threat to internal validity. In school-based programs, if the students were attending school, they would be attending the program. External validity is sacrificed if excessive attrition occurs. The prevention literature documents the higher rates of drug use for dropouts (Johnston et al. 1989). Pirie and colleagues (1988) even found higher rates of smoking among absentees and transfer students. 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). These 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. For studies not reporting a posttest near 12 months, the mean closest to the final posttest was used.

The final decision was whether differential dropout occurred from treatment or control. This information was missing in 61.7 percent of the reports in Tobler (1992*a)*, and these studies were grouped with those reporting differential dropout (a conservative method).

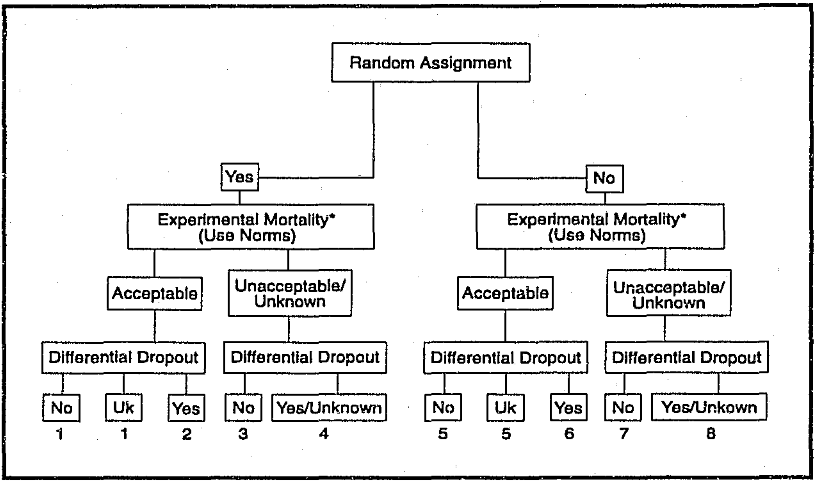


FIGURE 2. *Decision tree for combination of research design and mortality*

KEY: * Use posttest or followup attrition rate nearest a 12-month time interval.

The results of the decision tree (figure 2) for the set of 114 programs are shown in table 6 and provide the background for the conceptualization of the binary variable. Choices #1 to #4 pertain to experimental studies, while choices #5 to #8 are the parallel choices for quasi-experimental programs. The results show that no differences in effect sizes were observed for random (.17) versus nonrandom (.16) assignment. The second choice, acceptable mortality versus unacceptable mortality, showed the largest differences: experimental (.25 to .18) versus quasi-experimental (.12 to -.C3). Finally, the differences between no differential and differential dropout were: experimental (.25 to .16, and .18 to .07) versus quasi-experimental (.12 to .10 and -.03 to .18).Overall, within the group of experimental studies included in the 114 programs, the effect size decreased from choice #1 to choice #4. The results for the quasi-experimental programs showed a reverse pattern, with the poorest design (#8) having the largest effect size (see table 6). As a result of the complex empirical results for experimental design, it was decided that the

TABLE 6. *Effect size by quality of the research design*

EXPERIMENTAL	Choice	Mean	Standard deviation	n's
Mortality acceptable No differential dropout	# 1	0.25	0.33	33
Mortality acceptable Differential dropout	# 2	0.16	0.25	14
Mortality NOT acceptable No differential dropout	# 3	0.18	0.10	10
Mortality NOT acceptable Differential dropout	# 4	0.07	0.25	22
Subtotal Experimental		0.17	0.28	79

QUASI-EXPERIMENTAL	Choice	Mean	Standard deviation	n's
Mortality acceptable No differential dropout	# 5	0.12	0.12	5
Mortality acceptable Differential dropout	# 6	0.10	0.11	3
Mortality NOT acceptable No differential dropout	# 7	-0.03	0.00	1
Mortality NOT acceptable Differential dropout	# 8	0.18	0.26	26
Subtotal Quasi-experimental		0.16	0.23	35

TOTAL 114 PROGRAMS

eight categories were best represented by the binary variable detailed above.

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: (1) smoking programs, (2) alcohol programs, and (3) 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). Later publications will include analyses for cigarettes and alcohol separately. The reference group was smoking programs.

Leaders. Four categories of leaders were entered for this block: (1) teachers, (2) peer leaders, (3) mental health clinicians, and (4) others (e.g., research staff, health educators, and various outside professionals). The reference group was teachers.

RESULTS

Unweighted Versus Weighted Mean Effect Sizes

Comparison of the unweighted mean effect size with the weighted mean effect size shows the unweighted means consistently are higher than the weighted means. However, these higher effect sizes have corresponding larger confidence intervals than those for the weighted means. For the entire set of 114 programs, the unweighted mean (.17) had a confidence interval of .12 to .22, whereas the weighted mean (.13) had a confidence interval of .12 to .15.

A very important observation is that the 56 higher-quality experimental programs had higher means, unweighted (.22) and weighted (.17), than those for the 114 programs. The confidence intervals for the 56 experimental programs were larger for the unweighted effect (.14 to .30) than for the weighted effect (.14 to .19).

Results of the unweighted OLS regressions for the 114 programs showed a total R^2 of 26.3 percent, $F_{(10,103)} = 3.673$, and $p = .0003$. For the 56 experimental programs, the total R^2 was 50.5 percent, $F_{(10,45)} = 4.629$, and $p = .0002$. For the weighted WLS regression analysis, the total R^2 for the 114 programs was 20.5 percent, $F_{(10,103)} = 2.650$, $p = .006$. For the 56 experimental programs, the total R^2 was 32.0 percent, $F_{(10,45)} = 2.1208$, and $p = .042$. Again, the set of 56 programs had total R^2's that were higher than the corresponding total R^2's for the 114 programs.

Noninteractive Versus Interactive Programs

Tables 7(a) (114 programs) and 7(b) (56 programs) give the unadjusted and covariate-adjusted means for the *unweighted* effect size by type of program. Tables 7(c) (114 programs) and 7(d) (56 programs) compare the unadjusted and covariate-adjusted means for the *weighted* effect size by type of program. Examining the covariate-adjusted means by type of program shows that difference between the Interactive programs and Noninteractive programs is substantial: the unweighted (114) programs were .24 compared to .07, the unweighted (56) programs were .31 compared to .03, the weighted (114) programs were .22 compared to .11, and the weighted (56) programs were .26 to .08. However, examining the relationship between the unadjusted mean effect sizes and the covariate-adjusted mean effect sizes for either the Interactive or the Noninteractive programs reveals very small differences between the actual means and the predicted or covariate-adjusted means (approximately .01).

Tables 7(a)-7(d) also shows the n's, unstandardized betas, their standard error, the *t* value, and corresponding *p* values for the comparison of Interactive versus Noninteractive programs (reference group). The betas are higher, for both the unweighted (.28) and weighted (.18) regressions for the 56 experimental programs, than the betas for the unweighted (.18) and weighted (.12) regressions for 114 programs. In all four regressions, the Interactive programs are significantly better than the Noninteractive programs: $p = .002$ for the unweighted OLS regression analysis for 114 programs; $p = .001$ for the unweighted OLS regression analysis for 56 programs; $p = .009$ for the weighted WLS regression analysis for 114 programs; and $p = .015$ for the weighted WLS regression analysis for 56 programs.

Increment to R^2

The increment to R^2, the F change, and the significance of the F change for the independent variable (type of program), as well as any covariates that reached significance in any of the four regressions, can be found in Tables 8(a)-8(d). For the independent variable (type of program), the F change was significant in all four regressions, and the increment to R^2 ranged from 5.6 percent to 13.1 percent.

The F change for the covariate, *sample size*, also was significant in all four regressions. The next largest proportion of variance accounted for after partialing out the effect of the other variables was due to the sample size and ranged from 4.6 percent to 7.5 percent. Compared to the corresponding increments to R^2 for the type of program, these values are lower.

The F change for *targeted drug* was significant in only the two unweighted regressions. However, in the unweighted regression, the increment to R^2 was higher than the corresponding increment to R^2 for the type of program. The extremely high increments to R^2 for targeted drug that occurred in the unweighted OLS regressions were not found when using the weighted WLS regression analysis. This would indicate that higher effect sizes for smoking programs were found for the smaller

TABLE 7(a). *Unadjusted and covariate-adjusted unweighted effect sizes by type of program for 114 programs*

Type of program	n's	Beta	SE	t	Sig. t	Unadj	Cov-Adj
Noninteractive	44	Reference Group				.058	.066
Interactive	70	.18	.06	3.207	.002	.247	.243

TABLE 7(b). *Unadjusted and covariate-adjusted unweighted effect sizes by type of program for 56 experimental programs*

Type of program	n's	Beta	SE	t	Sig. t	Unadj	Cov-Adj
Noninteractive	18	Reference Group				.017	.031
Interactive	38	.28	.08	3.451	.001	.317	.312

TABLE 7(c). *Unadjusted and covariate-adjusted weighted effect sizes by type of program for 114 programs*

Type of program	n's	Beta	SE	t	Sig. t	Unadj	Cov-Adj
Noninteractive	44	Reference Group				.112	.106
Interactive	70	.12	.04	2.681	.009	.217	.221

TABLE 7(d). *Unadjusted and covariate-adjusted weighted effect sizes by type of program for 56 experimental programs*

Type of program	n's	Beta	SE	t	Sig. t	Unadj	Cov-Adj
Noninteractive	18	Reference Group				.062	.079
Interactive	38	.18	.07	2.541	.015	.271	.263

TABLE 8(a). *OLS regression analysis: Unweighted effect size for 114 programs*

Variable	Increment to R^2	F change	Sig. F
Sample size	4.6%	5.333	.023
Special populations	0.3%	0.321	.572
Targeted drug	9.8%	6.340	.002
Type of program	7.4%	10.287	.002

TABLE 8(b). *OLS regression analysis: Unweighted effect size for 56 experimental programs*

Variable	Increment to R^2	F change	Sig. F
Sample size	7.5%	4.344	.040
Special populations	0.2%	0.121	.730
Targeted drug	23.3%	8.806	.001
Type of program	13.1%	11.912	.001

TABLE 8(c). *WLS regression analysis: Weighted effect size for 56 experimental programs*

Variable	Increment to R^2	F change	Sig. F
Sample size	5.1%	2.918	.093
Special populations	2.8%	1.569	.216
Targeted drug	8.2%	2.385	.103
Type of program	9.7%	6.455	.015

TABLE 8(d). *WLS regression analysis: Weighted effect size for 114 programs*

Variable	Increment to R^2	F change	Sig. F
Sample size	5.9%	7.010	.009
Special populations	3.5%	4.234	.042
Targeted drug	1.7%	1.025	.362
Type of program	5.6%	7.190	.009

programs but, when the smoking programs were replicated on a larger scale, the differences no longer were significant.

The finding for special populations was significant in only the weighted WLS regression analysis for 114 programs and had a small increment to R^2 (3.5 percent). The programs that produced the significant results were included in the set of 114 programs but were eliminated from the 56 experimental programs, as they were evaluated with quasi-experimental designs. The F change for the three remaining covariates, *type of control group*, *experimental design*, and *type of leader* were nonsignificant, and their increment to R^2 was below 2 percent.

DISCUSSION

Substantive Findings

Meta-analysis can resolve conflicts (see figure 3). The magnitude for the mean effect size of the Interactive programs (n = 70) is considered small but is to be expected for large-scale implementations of social programs. The clear-cut positive direction of these effects cannot be ignored. Essentially, the Noninteractive programs (n = 44) do not prevent, retard, or reduce adolescent drug use. Had the analyses not separated the Interactive from the Noninteractive programs, quite possibly the conclusions

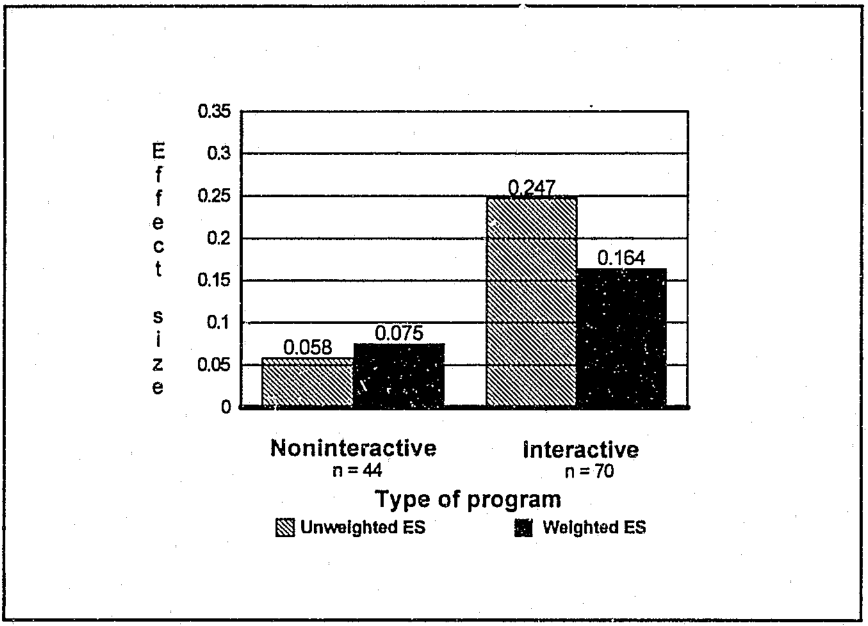


FIGURE 3. *Unweighted and weighted effect sizes by type of program for 114 programs*

SOURCE: Tobler (1993)

would have been equivocal. Instead, it can be concluded that, although not all drug prevention programs work, the Interactive programs are effective.

Another important finding is seen in figure 4. The differences between the two types of programs is even larger for the set of 56 higher quality experimental programs. The inclusion of quasi-experimental studies in the 114 programs did not cause upward positive bias; in fact, both the unweighted and weighted effects are higher for the 56 higher quality experimental studies. An alternative explanation may be that a more stringent selection criterion was used. The selection criteria (see previous section) ruled out many other factors that could affect the magnitude of success, such as posttest results taken at less than 3 months, intensities fewer than 4 hours (1 week of classes), cross-sectional research, and treatment programs compared to another treatment. Perhaps,

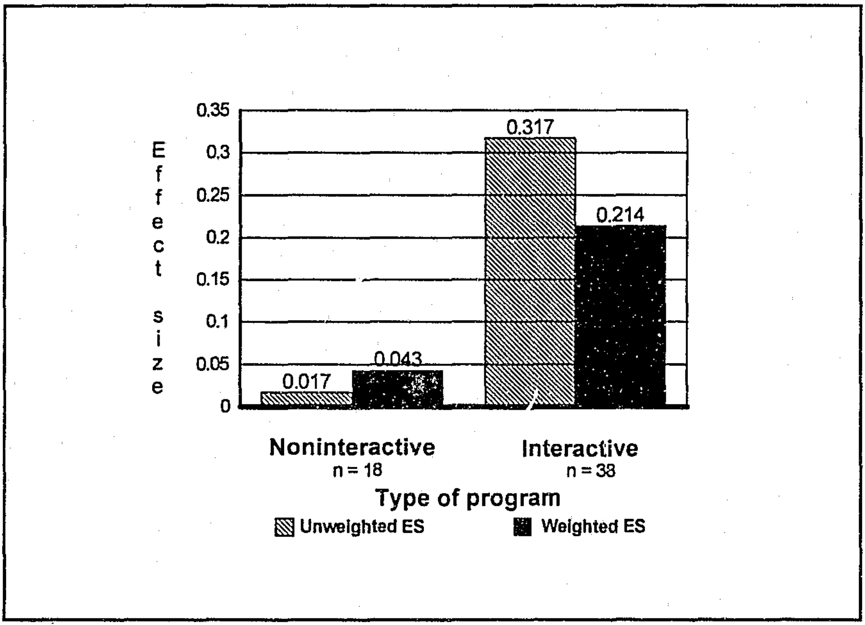


FIGURE 4. *Unweighted and weighted effect sizes by type of program for 56 programs*

SOURCE: Tobler (1993)

the confusion reported in the literature arises from the inclusion of research studies that could not be expected to show program success for a myriad of reasons like those stated above.

The success of the Interactive programs is not without a caveat. It appears that a leveling of the effectiveness occurs when programs are implemented on a large scale. Although the Interactive programs were still superior, the differences between the Interactive and Noninteractive programs become smaller for the larger-scale studies. A potential explanation may be implementation issues; this is a possible direction for new primary research. An Interactive program must include participation by everyone, preferably in small groups. The problem may be when a program is implemented in a regular class situation; without extra leaders, the student would interact only a few times. As the intensity of the average programs is very low (X = 10 hours), an essential part of the Interactive programs may have been missing—that of active involvement, exchange and validation of ideas with their peers, and time enough to practice and truly acquire interpersonal skills.

Replication of Findings

The results reported here replicate the more complex findings by size groups reported in Tobler (1992*a*) without questioning the reliability of the statistical procedures. 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 3 size groups was fewer 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 faced by social scientists of numerous, correlated predictors and limited sample sizes (Pruzek and Lepak 1992).

These findings also replicate similar findings (Tobler 1986, 1992*b*) with an updated set of programs (1978-1990) in which only 39 programs were included from the 1986 meta-analysis. This set of programs contributed R^2 increments ranging from 5.6 percent to 13.1 percent, a much stronger finding than Tobler (1986), in which only 4.2 percent of the total R^2 was accounted for by the type of program.[12]

Meta-Analytical Findings

The very small differences between the actual means and the covariate-adjusted means are noteworthy. Meta-analyses are not designed experiments in which equal numbers of programs are assigned to all categories of the independent variables and then manipulated. Meta-analyses always are observational studies, and the numbers in a particular category are dependent solely on the programs located. Even though six covariates were included, the covariate-adjusted means showed little difference from the actual means, suggesting a relatively balanced data set emerged although it was not preplanned. Little was added or subtracted for *these* covariates to alter the outcome success for Interactive compared to Noninteractive programs.

Previously, in Tobler (1986), when controlling for the experimental design and the type of outcome measure, much larger differences were observed between the actual and covariate-adjusted means. The proportion of variance accounted for by the type of outcome measure was higher than the type of program. This confounding factor was eliminated by including only programs with drug use outcome measures.

A comparison of the unweighted OLS and the weighted WLS regression analyses is included in Shadish (1992, p. 146). He states, "To the best of our knowledge, results of an extensive empirical contrast between these two approaches has not been published before on real data." In this meta-analysis, the results are quite similar;[13] the weighted mean effect was 15 percent smaller than the unweighted mean effect. This compares to Shadish's reduction of 13 percent. The standard errors for the weighted effects in this meta-analysis were 57 percent smaller than those for unweighted effect sizes. This is much larger than Shadish's 26 percent reduction. Shadish concludes by questioning the continued use of unweighted OLS regression analyses for meta-analysis, particularly because he found the homogeneity of variance assumption had been violated in 22 of 28 of his OLS regression analyses. However, Shadish's sample varies from 4 to 119, compared to a range of 20 to 6,000 in this meta-analysis. Therefore, the author feels that *both* methods should be reported and the reader should keep in mind the problems as well as the advantages offered by each method.

The advantage of using the *unweighted* effect size (one effect per program) is that the smaller programs are not overshadowed by the larger programs. Also, it enables comparisons with previous meta-analyses conducted before the formulation of the weighted effect size. A problem in using the unweighted effect is the violation of homogeneity of variance, which makes it unreliable meta-analytically. The *weighted* effect size is meta-analytically sound (effect weighted for sample size) but, in using the weighted effect[14] with *this body of research*, the larger programs overwhelm the smaller ones.

A possible explanation of the complex findings for the experimental design could be that dropouts had higher rates of drug use as reported by 63.3 percent of the programs. Therefore, programs experiencing higher attrition rates would be expected to have higher effect sizes if there was no differential dropout. This might explain the higher effect sizes for choice #8 (the poorest quasi-experimental design), but it does not explain the lower effect sizes for the experimental studies with differential dropout. As preexisting differences occurred in many of the experimental studies, this may be a function of high attrition interacting with preexisting differences. Preexisting differences were not examined in depth. As other findings in Tobler (1992*a*) highlighted, this may be more important than the attrition rates or the differential dropout from treatment or control.

Preexisting differences were addressed in the analyses of the 56 experimental programs by eliminating studies that did not analyze results on change scores (pretest/posttest data). Even in this case, many times the covariates chosen by the primary researcher were not initial pretest drug use but socioeconomic or ethnicity factors.

Meta-Analytical Areas not Addressed in Tobler (1992*a*)

Intraclass correlations were not used for research studies that implemented more than one type of program. No statistical corrections were made for the differences in the magnitude of the effect size that might exist between studies that used a different unit of assignment from data analysis and those that used the same unit of assignment and data analysis. More importantly, no corrections were made for the differences in the magnitude of effect sizes that might exist based on the unit of data analysis: the school, the classroom, or the individual. Very few effect sizes were based on the classroom or school as the unit of analysis and, for those for which df were used to calculate the effect size, the number of classrooms or schools was chosen. No control was made for possible differences in the magnitude of the effect size resulting from a researcher's use of parametric versus nonparametric statistics.

Recommendations for Future Meta-Analysts

Effect sizes should be computed from covariate-adjusted means or change scores whenever possible, as this allows statistical control for preexisting differences. Change scores should be used, even in the case of nonsignificant preexisting differences. A statement of nonsignificant preexisting differences is meaningless for this field, as the program effect sizes also are statistically nonsignificant. When working with small effect sizes, the preexisting differences actually may be greater than the effect size for the success of the program. Meta-analyses should be reported for both the unweighted and weighted effect sizes until enough well-controlled studies exist to separate the disparate studies into two groups (i.e., smaller efficacy studies and larger implementation studies). Analyses of program results should be based on initial level of use (i.e., pretests confidentially administered with ID numbers) with no more than three levels of use (i.e., nonusers, experimental users, and users).

Recommendations for Primary Researchers and Their Funding Sources

All research funding sources should prioritize the use of identical drug use outcome measures like those found in any of the surveys conducted by Johnston and colleagues (1985, 1989) when awarding grants. This does not preclude the use of individualized measures. Primary researchers should report analyses of the adolescents who drop out from the research study. Funding should be placed in the original grant to provide for the extensive procedures necessary to conduct a followup of the dropouts. Three elements of the problem should be reported: (1) the characteristics of the adolescents, particularly drug use levels; (2) the rate of dropout from both the treatment and control conditions; and (3) whether the programs were successful with adolescent nonusers, experimental users, and users.

Primary research studies should be developed, funded, and evaluated for high school age youth, particularly those who voluntarily enter a program (i.e., student assistance programs). A method should be formulated for evaluating the effectiveness of a program in which the behavior has not yet manifested itself. Possibly, longitudinal research could begin in the fourth grade before program implementation. Longitudinal research should not be done with programs of low intensity unless equally intensive boosters are given yearly.

Primary researchers should be documenting what the control condition received as thoroughly as what the treatment group received. The careful examination of the nature and amount of drug intervention activities in the control schools can explain the lack of more substantial findings from excellent programs. Although placebo control groups have been called for and are appropriate for efficacy studies (to assure that extra attention is not the reason for a program's success), prevention programs cannot be withheld, nor will State legislatures allow it. Even before the Drug Free Schools and Communities Act of 1986, there were few schools that did not offer some form of drug education. Finally, Ary and colleagues (1989, p. 15) have called for "assessing the incremental effects of a specific intervention package."

The field is ready for this type of research but it needs the cooperation of State-level school officials working in concert with NIDA officials. Outcome measures would need to be identical. This type of effort cannot be sustained by a *single* university or a *lone* NIDA grantee. If two prevention programs were tested against each other, longitudinal tests could proceed, and the analysis of implementation factors could receive greater emphasis, as both groups would be receiving a prevention program.

Recommendations for Editors

Editors should encourage the following information to be included in all articles being considered for publication: an effect size, exact n's for that effect size, exact *p* values even though nonsignificant, reports of preexisting differences, attrition rates, and analysis of dropouts as just described.

A technical report should accompany and be kept on file for all published journal articles, as this information often has been omitted due to page limitations.

Recommendations for Policymakers

Policymakers must be made cognizant of the fact that successful school-based drug prevention programs address only *one* of the myriad of reasons for adolescent alcohol AOD use, namely peer pressure. Policymakers should not expect drug prevention program effect sizes to be as large as those reported in the 1970s and early 1980s. Since 1979, when adolescent drug use peaked (Johnston et al. 1989), public concern has been reflected in schoolwide policies, community activities, and mass media efforts. There are virtually no schools that can act as a pure control group. Presently, all drug prevention programs are being compared to another treatment; that is, even though a no-treatment control group has been identified, the students in that group are receiving some form of drug prevention (e.g., assemblies and drug prevention week). In addition, many youth have been exposed in the earlier grades to drug prevention programs. Policymakers should be alerted to the problems associated with implementation of large-scale effectiveness trials so they can commit themselves to the level of funding needed for this type of experimental research. Policymakers are in a position to quell the expectations of an anxious general public for immediate and lasting results to be produced by a single implementation of an adolescent drug prevention program (i.e., the wished-for silver bullet).

Congressional leaders should consider funding a staff to establish a permanent protocol for a continuous meta-analysis that would be updated as each new generation of prevention programs is completed. A data base should be established in which all funded researchers would have to report the minimum information necessary for a meta-analysis. After a few years, no matter how much a researcher would like to supply the information, it is impossible, too costly, or too time consuming to retrieve.

CONCLUSIONS

Meta-analysis is a research tool that, when used correctly, can help resolve the conflicts in drug prevention intervention research. It does not replace high-quality primary research but affords a method of aggregating the present primary research and, therefore, lends generalizability to a set of studies previously judged to have internal validity. The pitfalls of the inappropriate use of meta-analysis and the advantages of using meta-analysis were illustrated with a set of adolescent drug prevention programs.

NOTES

1. Prevention intervention research will be referred to as "prevention research" for purposes of brevity.

2. Primary research refers to the primary analysis of the original data in a research study (Glass et al. 1981, p. 21).

3. "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).

4. 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).

5. Programs that targeted cigarettes were excluded.

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

7. The unweighted mean effect size for the 114 programs was used to estimate this value. In actual practice, each drug would be corrected at the study level.

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

9. Type of program is defined as the intersection of program content with the group process used to implement the program.

10. "High-risk youth" is defined as an individual who: is a school drop-out; 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 (Anti-Drug Abuse Act 1986).

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

12. Type of program replaces the term "modality" used in Tobler (1986). Interactive programs were called peer programs in Tobler (1986).

13. Lipsey's (1992) findings for his meta-analysis of 397 juvenile delinquency programs also report a higher unweighted mean effect size (ES = .172). He found a weighted mean effect size equal to .103.

14. The weighted WLS regression statistics are tested at the program level and not at the level of the number of cases (i.e., individuals).

REFERENCES

Altman, L. New method of analyzing health data stirs debate. *Science Times, The New York Times*, August 21, 1990.

Anti-Drug Abuse Act. P.L. No. 99-570, 100 Stat. 3207, 1986.

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, 1899 Willamette Street, Eugene, OR 97401.)

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

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

Botvin, G., and Wills, T. Personal and social skills training: Cognitive-behavioral approaches to substance abuse prevention. 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. pp. 8-49.

Brunvold, W., and Rundall, T. A meta-analysis and theoretical review of school based tobacco and alcohol interventions. *Psychol Health* 2:55-73, 1988.

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.

Cahen, L.S. Meta-analysis—A technique and promises. *Evaluation in Education.* Vol. 4. Great Britain: Pergamon Press Ltd., 1980. pp. 37-39.

Cohen, J. *Statistical Power Analysis for the Behavioral Science.* New York: Academic Press, 1977.

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

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.

Cook, T., and Leviton, L. Reviewing the literature: A comparison of traditional methods with meta-analysis. In: Light, R., ed. *Evaluation Studies Review Annual.* Vol. 8. Beverly Hills: Sage Publications, 1983. pp. 59-82.

Cooper, H., and Rosenthal, R. Statistical versus traditional procedures for summarizing research findings. In: Light, R., ed. *Evaluation Studies Review Annual.* Vol. 8. Beverly Hills: Sage Publications, 1983. pp. 59-82.

Copper, H. *Integrating Research: A Guide for Literature Reviews. Applied Social Research Methods Series.* Vol. 6. Beverly Hills: Sage Publications, 1984.

Eysenck, H. An exercise in mega-silliness. *Am Psychol* 33(5):517, 1978.

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.

Gallo, P. Meta-analysis—A mixed metaphor? *Am Psychol* 33(5):515-517, 1978.

Gendreau, P., and Andrews, D. A tertiary prevention: What the meta-analyses of the offender treatment literature tells us about "what works." *Can J Criminol* 32:173-184, 1990.

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

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

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

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.* National Institute on Drug Abuse. DHHS Pub. No. (ADM)87-1537. Washington, DC: Supt. of Docs., U.S. Govt. Print. Off., 1987. pp. 81-131.

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

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

Ingram, L. An overview of the desegregation meta-analyses. In: Wachter, K., and Straf, M., eds. *The Future of Meta-Analysis.* New York: Russell Sage Foundation, 1990. pp. 61-70.

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. Washington, DC: Supt. of Docs., U.S. Govt. Print. Off., 1985.

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.

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.

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.

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

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.

Pillemer, D., and Light, R. Benefiting from variation in study outcomes. In: Rosenthal, R., ed. *New Directions for Methodology of Social and Behavioral Science.* Vol. 5, *Quantitative Assessment of Research Domains.* San Francisco: Jossey-Bass, Inc., 1980. pp. 1-11.

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.

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

Rosenthal, R., and Rubin, D. Summarizing 345 studies of interpersonal expectancy effects. In: Rosenthal, R., ed. *New Directions for Methodology of Social and Behavioral Science.* Vol. 5, *Quantitative Assessment of Research Domains.* San Francisco: Jossey-Bass, Inc., 1980. pp. 79-95.

Rosenthal, R., and Rubin, D. Comparing effect sizes of independent studies. *Psychol Bull* 92(2):500-504, 1982.

Schaps, E.; DiBartolo, R.; Moskowitz, J.; Palley, C.; and Churgin, S. A review of 127 drug abuse prevention program evaluations. *J Drug Issues* 11(1):17-43, 1981.

Schmidt, F. What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. *Am Psychol* 47(10):1173-1181, 1992.

Shadish, W. Do family and marital psychotherapies change what people do? A meta-analysis of behavioral outcomes. 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. 129-208.

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. Integrating studies of psychotherapy outcomes. In: Rosenthal, ed. *New Directions for Methodology of Social and Behavior Science: Quantitative Assessment of Research Domains.* No. 5. Washington, DC: Jossey-Bass, Inc., 1980*b*. pp. 47-61.

Smith, M. Publication bias and meta-analysis. *Eval Educ* 4:22-24, 1980*a*.

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

SPSS, Inc. *SPSS Reference Guide.* Chicago: SPSS, Inc., 1990.

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. pp. 141-153.

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*. National Institute on Drug Abuse. 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*. Proceedings report from a March 18-20, 1993, conference. San Diego: UCSD Extension, University of California, 1993. pp. 71-86.

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

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.

ACKNOWLEDGMENT

This research was supported in part by National Institute on Drug Abuse grant R01DA05505 and New York State Division of Substance Abuse contract C002429.

AUTHOR

Nan Tobler, M.S.W., Ph.D.
School of Social Welfare
State University of New York at Albany
135 Western Avenue
Richardson Hall
Albany, NY 12222

Mailing Address:

Box 246
Sand Lake, NY 12153
(518) 674-3559

Dynamic Systems-Modeling as a Means To Estimate Community-Based Prevention Effects

Barry M. Kibel and Harold D. Holder

ABSTRACT

The best applications of prevention programming now are believed to be community-supported strategies, comprehensive in scope, implemented in stages, and delivered by public and private agencies and organizations. Despite the flurry of program activity, research directed toward comprehensive, community-based prevention programming remains a largely uncharted domain. Communities need to know what will work in their specific contexts. The essential question communities are asking themselves and consultants is, "Which *mix* of interventions will yield maximum reduction in alcohol and other drug (AOD) abuse in *our* community?"

Computer-based dynamic simulation modeling holds promise in helping provide an answer to this question. These models are designed to replicate the historic patterns and dynamics of target communities with regard to substance availability, use, and misuse and then to simulate future patterns and dynamics under alternative assumptions and intervention mixes. Researchers can use these models to construct structural relationships that reflect alternative theories or explanations for important processes. The models afford community planners and decision-makers with mechanisms for asking "what if" questions regarding alternative intervention mixes and, consequently, for determining which politically acceptable mix of feasible interventions will yield the most reduction in AOD-related problem behaviors. In this chapter, a model under development by the Prevention Research Center of the Pacific Institute for Research and Evaluation is described that focuses on alcohol use and

misuse in communities and allows the testing of a wide range of prevention options.

THE CONTEXT: COMMUNITY-BASED PREVENTION

The alcohol and other drug (AOD) prevention field has exhibited dramatic growth in sophistication during the past three decades. This is best illustrated through a brief review of school-based prevention strategies. During the 1960s, the belief prevailed that providing information on the negative aspects of AOD use, misuse, and abuse would deter experimentation and decrease risky use. While these efforts had some effect on AOD knowledge and attitudes, these programs were shown essentially to have no effect on subsequent drug use by school-aged populations (Stewart and Klitzner 1992). The weight of evidence from numerous evaluation studies suggested that knowledge-based programs, *when implemented in isolation*, had little lasting impact on drug use, and there even might be negative or boomerang effects (i.e., awakening curiosity about drugs).

During the 1970s, the central focus for primary prevention aimed at school-aged populations shifted away from knowledge-based curricula toward affective education approaches (Orlando 1992*a*). These were based on exercises aimed at increasing self-confidence, improving self-concept, and adapting positive values and beliefs. During the late 1970s and throughout the 1980s, affective education was augmented with life-skills education and resistance training. Life-skills approaches emphasized the development of personal competencies, communication and decision-making skills, as well as skills in resisting pressures to use tobacco, drugs, and alcohol. Resistance training approaches were more specific in their focus on teaching and rehearsing skills to resist peer and other social pressures to use drugs. A comprehensive evaluation of the long-term effects of these efforts has not been performed. However, studies of selected programs employing strong research designs have noted some delays in initial experimentation with drugs but few sustained effects (Moskowitz 1989).

The inability to sustain long-term effects has been explained as follows (Orlando 1992*a*):

> Interventions that focus their positive influence on one specific context or element of an individual's life—for example, the school context for adolescents—will eventually fail if the summative effects of all other negative influences are greater. Thus, given that students spend a relatively small portion of their time in school, the role of family, friends, organizations, groups, mass media, and other nonschool influences should not be ignored.

This experience with school-aged populations can be generalized to the entire population: While programs aimed at individuals in some immediate settings (e.g., workers in the workplaces) can be part of a comprehensive prevention strategy, they are incomplete. The potential drug user/misuser/abuser plays a range of life roles (e.g., worker, husband or wife, father or mother, friend, and neighbor) within a variety of settings, each potentially with its own unique codes of behavior regarding the use of substances. These codes have been referred to elsewhere as "cultural recipes" (Maloff et al. 1979). These codes influence how different substances will be used at what levels, for what purposes, by which persons, at which times, and in which contexts (e.g., a worker might go for a couple of drinks with the guys after work). The individual may likely receive a range of mixed, but equally compelling, messages regarding substance use while moving through his or her life spaces and assuming these varied life roles. Accordingly, the effects from single-component prevention interventions (e.g., programs at the workplace) can likely be neutralized by the multiplicity of counterforces operating in the individual's other life spaces. The conclusion is that prevention strategies must have multiple components to address all influential aspects of the individual's life.

A natural extension of this logic is that prevention strategies are most effective when focused on the community at large rather than on specific individuals at risk. For purposes of AOD prevention, a "community" can be viewed as a set of persons with adverse behaviors or associated risks with respect to alcohol and drugs that the prevention is intended to reduce or eliminate. Alternatively, a "community" can be viewed as a set of persons engaged in shared social/cultural/political/economic processes that the prevention is intended to modify in order to reduce risky behaviors associated with alcohol and drugs. Both perspectives are aimed at risk reduction. The difference, stated succinctly, is that the first perspective focuses on behaviors of individuals at risk, while the second perspective focuses on systemwide behaviors collectively affecting these individuals, as well as others in the community.

The first approach (referred to as the "catchment-area perspective") is used commonly in health problem prevention. It follows a straightforward model: Find the persons at risk, then educate or serve them in an appropriate manner to reduce the *individual risk* to each person so identified. The second approach (referred to as the "community-systems perspective") is used less commonly due, perhaps, to its greater conceptual complexity. AOD-related problems are outcomes of processes driven and sustained by the community at large. These potentially affect all members of the community while producing adverse effects in certain groups more than in others (due to individual and environmental factors that contribute to disproportionate exposure or increased susceptibility). Through appropriate interventions affecting these processes, the intention is to reduce the *collective risk*. Both the catchment area and the community-systems perspectives deserve consideration in designing community prevention programs.

The catchment-area approach to treatment and prevention is useful particularly when some of the following criteria prevail:

1. *An individual's problem*. The targeted condition or behavior is contained (or can be contained) within individuals and can be treated as an individual condition or state (e.g., coronary heart disease, lung cancer, alcoholism, or drug dependency). Even if the condition potentially is transmittable, such as would be the case with Acquired

Immunodeficiency Syndrome (AIDS) or polio, still it is contained within affected individuals.

2. *A subpopulation-specific problem.* Those with problems or who are at risk of these problems can be identified by type (e.g., gender, age, ethnicity, religion, occupation, or residency) and can be prescribed appropriate services and opportunities.

3. *A recurring or continuous problem.* The condition or behavior is chronic or persistent; that is, it remains with the individual over a sustained period of time rather than occurring intermittently or infrequently.

4. *A tightly bounded problem.* The condition or behavior, while potentially influenced by environmental processes, appears to be defined largely within the context of the individual, the immediate family, and close social contacts, such as the peer group. The condition is disruptive largely of the individual's life and the immediate social network but usually does not directly affect the lives of others within the greater community.

A community-systems perspective to prevention differs from the catchment-area perspective in several important ways: (1) rather than addressing a single problem behavior or condition, a potentially wide-ranging set of problem behaviors are considered simultaneously, (2) rather than focusing on individuals at risk, the entire population within the community is studied in concert, and (3) rather than basing prevention strategies on direct causal linkages, interventions are considered that affect aspects of the behavioral environment, promote changes in decisions, and, thus, indirectly contribute to shifts in behavior of the population away from problem-causing contexts.

It is this type of "systemic thinking" that has led to the current community-based focus for prevention. The influences of environment and lifestyle on increased risk of cardiovascular disease and cancer were recognized fully in the mid-1970s by public health professionals.

Communitywide interventions were developed to reduce these risks. In like manner, community-based prevention strategies now are being developed to modify the environments and contexts within which substance use occurs, so as to reduce or eliminate harmful effects of such use. The intent is to foster communities where use of illegal substances has been eliminated and where use of legal, but potentially risky, substances is regulated through formal controls or social norms. Federal agencies like the Center for Substance Abuse Prevention (CSAP) and the National Institute on Alcohol Abuse and Alcoholism (NIAAA) are leading the support for these efforts through their community partnership grants and community trials projects, respectively. State-level initiatives, such as California's Friday Night Live and Club Live programs, are attempting to expand prevention activities aimed at youth beyond the school day to include social and community-service activities. The Robert Wood Johnson Foundation also is supporting major research in community-based prevention and treatment coordination through its Fighting Back initiative.

While varying greatly with respect to goals and implementation approaches, these community-based prevention efforts share some common characteristics. They each emphasize community responsibility and ownership of the prevention strategy. They each promote inclusion, whereby participation by a widening set of individuals and groups is actively encouraged. They each depend for their continued survival on the tacit or active support of the official leadership of the greater community of which they are a part. Most significant for research purposes, they each involve multifaceted strategies that collectively affect the individual users (e.g., changing behaviors), the substances being used (e.g., use of warning labels), and the conditions or contexts of use (e.g., restricting availability). These strategies recognize that (1) program interventions that prove successful with certain age or cultural groups are not guaranteed to be applicable to other groups; (2) no single intervention will reach all groups with equal measure, hence the need for varied approaches; and (3) single-component interventions fail to account for the complex set of factors that frustrate or negate well-intentioned efforts directed toward only one sensitivity point of a network of social systems.

The best applications of prevention programming now are believed to be *community-supported strategies, comprehensive in scope, implemented in stages, and delivered by public and private agencies and organizations.* To illustrate, an approach to community trials developed by the Prevention Research Center (PRC) under a grant from NIAAA and focused on alcohol-related trauma includes five interrelated components: (1) efforts to raise public awareness, (2) beverage server and owner/manager training, (3) school- and community-based efforts to discourage sales and access to alcohol by underage populations, (4) increased focus on enforcement of drinking and driving laws, and (5) use of local zoning powers and other regulatory controls to reduce availability of alcohol. Trial-site communities have formed broad-based coalitions to consider how best to implement these five components in concert.

RESEARCH QUESTIONS BEING ADDRESSED

Despite the flurry of program activity, research directed toward comprehensive, community-based prevention programming remains a largely uncharted domain. Communities need to know what will work in their specific contexts. Too often, decisions are made about prevention strategies based on research or hearsay evidence of success in communities that may not be comparable. Furthermore, there is little available research on multicomponent intervention. Yet, the essential question that communities are asking themselves and consultants is: *Which mix of interventions will yield maximum reduction in AOD-related trauma in <u>our</u> community*?

The challenge for researchers is heightened by both the multicomponent character of the strategies and the selection and implementation of their components through communitywide, participatory planning processes rather than through research hypotheses and controlled experiments.

With regard to the latter, a group of researchers at Toronto's Addiction Research Center noted the following (Giesbrecht et al. 1991):

1. Researchers and community members often have divergent priorities. The former are concerned with increasing the body of relevant knowledge. The latter are concerned with developing programs that match local conditions and address perceived needs.

2. Community members are prone to accept local "truths" and discard or distrust research propositions that conflict with these beliefs. Hence, they may not agree that a certain intervention does not work until they try it for themselves. They also may reject an intervention proposed by researchers because it does not sound like it has a chance of working locally.

3. Researchers may carry their own baggage into the community and consciously or otherwise embed these within their assumptions. For example, they may argue that the intervention has to be implemented in a specific way to permit comparisons across treatments. The communities, for their part, may insist on putting their own particular twist on the intervention to make it appear, or actually be, locally relevant.

4. Researchers may assume that community members possess the requisite knowledge and insights to grasp research that recommends a particular approach and fail to take the time to explain the approach so that it is embraced locally. Therefore, rather than admit to confusion, community members may counter with expressions of impatience and discard potentially valuable research.

A solution to these challenges, according to the Toronto group, lies in increasing involvement of community members in the actual research effort. This sentiment is echoed in a recently released CSAP monograph on culturally sensitive evaluation that calls for local research that respects local cultures and accommodates local realities when testing or evaluating new initiatives (Orlando 1992*b*).

A methodological approach that holds promise in addressing challenges associated with community-based prevention, while also taking account of the multicomponent character of the strategies, is developing, testing, and experimenting with computer-based simulation models of these communities. These models are designed to be tested against the historical patterns and dynamics of target communities with regard to substance use and misuse and then to simulate future patterns and dynamics under alternative assumptions and interventions. Researchers can use such models to explore structural relationships that reflect alternative theories or explanations for important processes (e.g., the relationship between the availability of a drug and its subsequent consumption by one or more groups within the community). Community groups can use the models to explore the anticipated impacts of alternative mixes of interventions prior to finalizing their plans for actual implementation.

CONCEPTUAL OVERVIEW

Under this approach, any community interested in exploring its AOD prevention options could have a computer model of the community at their disposal that replicates the unique dynamics of the community with respect to all substances of interest. The model would simulate the sets of behavioral and causal relationships needed to describe fully and accurately the dynamics of substance availability, use by different groups within the population, and resulting problem behaviors. The model further would permit the testing of changes in key economic and demographic parameters, national and local cultural norms, and public pressures and regulatory controls that moderate AOD use, misuse, and abuse.

Computer-modeling often has been used in other areas of research, including business, economics, health care, retail sales, and defense. However, it rarely has been used as a part of the development of the science of substance abuse. In this chapter, dynamic computer-modeling is a technique for developing causal understandings of the complex community system of which AOD use and abuse is a part. In this way, computer-modeling is a part of the data analytical tools available to researchers and to planners.

One of the goals of science is to enable people to understand the complex systems of which they are a part. The science of substance abuse is the search for tools and techniques that assist them in this goal. Dynamic systems-modeling is unique compared to the other data analytical tools described in this monograph. Traditionally, statistical techniques have been applied in the field of substance abuse to learn about empirical relationships between variables. For example, cross-sectional data analysis techniques may be used to derive findings from a school survey or a community survey. The results of this analysis, of course, are limited to the data utilized. Even if a nationally representative survey is conducted, the results are generalizable only to the time period of the survey itself. Such results describe neither what the situation was 5-10 years prior to this survey nor what it will be 5-10 years in the future.

Time series or longitudinal statistical techniques provide information about changes over time in variables under study. Time series analysis following Box and Jenkins (1976), McCleary and colleagues (1980), and Tiao and Box (1981) provides a statistical technique for establishing patterns and cycles in time series data. Such techniques have been used successfully in alcohol policy analyses (Blose and Holder 1987; Holder and Blose 1987; Wagenaar 1986; Wagenaar and Holder 1991*a*). However, such techniques primarily are a means to establish the historical patterns and cycles of the time series itself and do *not* identify the underlying causal relationships that produced the series. Thus, researchers are not provided with any scientific understanding that suggests how to intervene to improve things in the future.

Dynamic systems-modeling is both a perspective of the real world and a data analytical technique for developing a scientific understanding. In this perspective, community substance abuse is a system that is dynamic over time (i.e., it changes and contains many feedback loops whereby the results of a chain of cause and effect, in turn, influence these earlier causal factors). As a research tool, the model is a statement of the causal relationships between and among many factors. The preferred method to state these relationships is mathematical. Thus, the systems computer model actually is a series of mathematical equations that describe the dynamic interaction of a large number of variables over time.

Unlike many statistical techniques, the computer model is not a "curve-fitting" approach. The computer model is not loaded with historical data and then used to estimate the future in the way a regression equation is used. Rather, a computer model, as a structured set of relationships that have been expressed mathematically, is loaded with only initializing data and then started. Using this initializing data only, the model runs to simulate a period of time, say 20 years of history. The results of the model then are compared with known historical data (i.e., a time series). Thus, the model is tested against historical data, not loaded with these data. An acceptable model of a complex community system can recreate a required historical benchmark. Only when the model can pass its numerous historical benchmark tests is it judged ready to undertake experiments with the future.

Since the model is an explicit statement of a theory, it can be used to test hypotheses. If a model has not been validated empirically, it still can be used to examine the implications and perturbations of a theory before the theory is examined empirically. For example, if a researcher develops a explanatory theory of the relationship among the price of illicit drugs, changes in patterns of drug sales, and changes in drug use, a model could be developed to explore this theory before it is tested in the real world. If the model has been validated empirically already, it can be used to test hypotheses about possible changes to reduce AOD problems in the community. For example, a validated model could be used to examine the possible changes in drug use and drug-related problems in a community with various price levels for these drugs.

How, then, does a dynamic causal model get developed? What is its relationship to more traditional statistical techniques? In brief, the model is composed of a series of causal relationships. These relationships usually are derived from a variety of sources: (1) published scientific literature—the preferred source, (2) statistical analyses of data conducted specifically to examine variables and relationships for the model that have not been explored in the scientific literature, and (3) expert judgment. During model development, the same types of statistical techniques for data analyses described in this monograph are used.

Ideally, all interactive relationships necessary to develop a causal model should be tested empirically and published in research papers based upon peer review. Unfortunately, this seldom is the case. While much of the necessary research often exists, some critical variables may not have been sufficiently studied previously. If a data base is available on which empirical studies can be based, then unique analyses to assist model development are undertaken. Such data analyses have the advantage of being related directly to the design of the model. When neither of the prior two approaches is possible, expert judgment is required. Such judgment then can be tested during model validation using sensitivity-testing. In other words, if an estimate of a parameter value is not available from scientific research, then the model can be tested using a range of possible variables, that is, through sensitivity-testing.

An Alcohol Use Model

PRC is taking some preliminary steps toward a policy tool for AOD prevention through the development of a model of community alcohol use. Alcohol is a logical choice as a substance to model following the systems perspective. While it is true that heavily dependent users (i.e., alcoholics) have the greatest individual risk rates for most alcohol problems, their numbers are so small that they contribute only modestly to most alcohol-related problem areas. Infrequent and moderate users of alcohol, who are neither currently nor ever likely to be dependent on alcohol, account for the majority of alcohol-involved trauma, such as auto crashes, falls, and drownings. Young people, in particular, account for a disproportionately large number of alcohol-related problem events, such as traffic crashes and accidental injuries.

The probability that any one drinker at any specific time will incur an alcohol-involved trauma or death usually is quite low. For example, the chance of an alcohol-impaired driver being stopped and arrested by the police is estimated to be 1 in 2,000 events on the average. Most alcoholics who drink heavily throughout their life will never be involved in an alcohol-related traffic crash or have an encounter with the police. On the other hand, a young 18-year-old with limited driving and drinking experience may cause a serious auto crash with small amounts of alcohol in the blood system. Hence, many alcohol-related problem events, while ultimately assignable to individuals who had too much to drink, can better be interpreted as stochastic events (i.e., time dependent and probabilistic).

Furthermore, these events are not predictable in terms of individual characteristics alone. Many alcohol problems are the cumulative result of the structure and flow of complex social, cultural, and economic factors within the community system. The dynamics of alcohol use and associated problems change as new members enter and others leave, as alcoholic beverage marketing and promotion evolve, and as social and economic conditions (including employment and disposable income) change. No single prevention program, no matter how good, can sustain its impacts unless system-level changes are effected (Holder and Wallack 1986). These changes aim at lowering the odds of adverse events; that is, they induce shifts in individual decisions and risky behavior through relevant changes in the social, economic, and, in some cases, physical environments of the community system.

A personal computer-based model of alcohol use and abuse currently is being developed and tested at PRC (refer to figure 1). The model recreates the systems dynamics of a targeted community with regard to alcohol retail activity, alcohol consumption patterns, drinking and driving behavior, social norms, and regulatory controls. Published research findings, survey data, and results from secondary data analyses are used to define and mathematically specify relationships among variables within and across the model subsystems. Annual outcomes generated by the model include the distribution of consumption by age-sex groups,

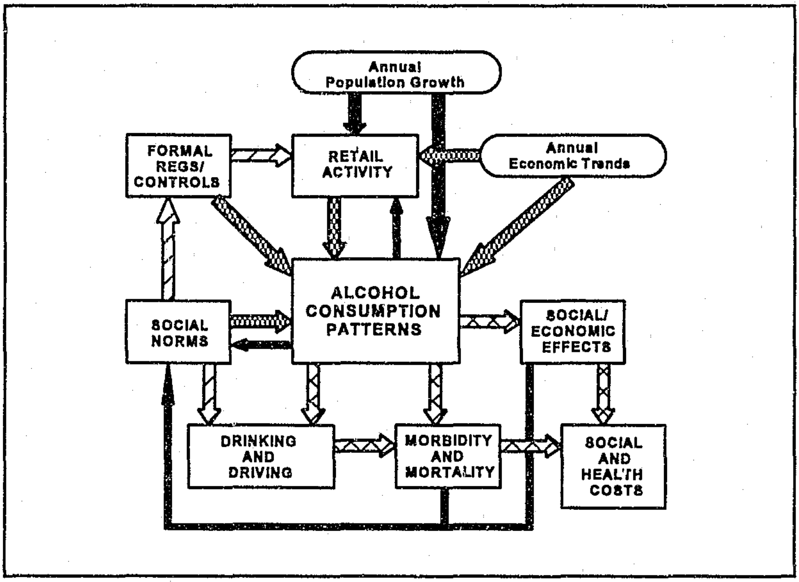


FIGURE 1. *Causal model of alcohol use and alcohol-related trauma*

alcohol retail sales, new licenses for the on-premises and off-premises sale of alcoholic beverages, driving under the influence (DUI) arrests and convictions, driver fatalities, injury crashes, measures of mortality and morbidity, and the socioeconomic consequences of problem drinking. Once congruence between model results and historic data for the period from 1970 to the end of 1991 has been achieved, the model can be used to simulate likely impacts of alternative prevention interventions over the period from 1992 to the end of 2001 and beyond.

Through an iterative process of design, congruence-testing, and redesign, the model is being refined toward the future point when it can be offered to researchers and prevention specialists as a reliable, comprehensive tool for understanding complex community dynamics and for estimating impacts of interventions intended to reduce alcohol-related problems. The current version of the model has replicated alcohol use patterns successfully at the national level and for one State (California). Tests now are underway at the county level, using San Diego County, CA, as the test site. Future plans call for pilot-testing the model in

12 representative American communities that currently are engaged in prevention planning and have identified alcohol use and abuse as a local problem needing to be addressed.

Design Considerations

There are many parts of the community system contributing to the use of alcohol. While planners and community leaders intuitively may appreciate that alcohol-involved problems are impacted by many diverse factors, they generally do not have at their disposal the tools or technology to translate intuitive understandings into concrete relationships. Computer-modeling is a research and policy-evaluation technique that has been used for at least three decades to investigate problems and changes in problem indicators as a result of system-level shifts. However, this technique rarely has been applied to understanding alcohol use and to systematic study of the potential impact of varied prevention strategies on reducing alcohol-related trauma. Some researchers (Cook et al. 1973; Holder 1974; Schlenger et al. 1976) have applied computer models to various aspects of alcohol misuse and abuse. For example, Summers and Harris (1978) conducted a computer simulation of the general deterrence of driving while intoxicated, but this was not applied to specific communities. The community use of alcohol, in total, has not been explored previously through such models.

Studying a community from a systems perspective demands the accumulation and synthesis of considerable amounts of disparate information about that community. The design process begins with the compilation and articulation of rules that succinctly describe how the population and its environment behave under diverse circumstances and conditions. These behavioral rules are converted to mathematical and algorithmic forms that allow a large number of variables to interact with one another over the time period being simulated. Baseline data are gathered, and the model is calibrated to replicate historical patterns and dynamics. Congruence-testing of the model-generated results against benchmarks (actual historical data not used for initialization) determines how well the model "fits" reality. Once calibrated, data from different communities are introduced to determine how well the same model structure can replicate these new patterns.

The authors refer to this type of modeling as "causal modeling" to distinguish it from statistical modeling approaches:

- *Statistical models* are derived primarily through curve-fitting exercises, whereby actual data are used to compute a mathematical expression(s) that most closely approximates the relationships existing across variables.

- *Causal models*, in contrast, begin with the compilation and articulation of rules (i.e., a working theory) that succinctly describe how the population and its environment behave under diverse conditions. These behavioral rules are converted to algorithmic forms that allow a large number of variables to interact over time. Some actual data are used to set the initial conditions for the model, after which the behavioral rules of the model generate values for all variables over time.

The PRC model is now in its third generation of development. The first generation of the PRC model, begun in 1980, explored the general approach for describing a community through causal modeling with regard to its alcohol use. The model was tested for congruence at the national level and preliminarily tested at the local level in three communities (Alameda County, CA; Washington County, VT; and Wake County, NC). Second-generation modeling focused on congruence-testing of two subsystems: "alcohol consumption" and "drinking and driving," in San Diego County. The second-generation model subsequently was used to examine a range of interventions. Third-generation modeling, completed in November 1992, focused on design and congruence-testing of the eight subsystems arrayed in figure 1. Again, San Diego County was the target community. Historical data for the period from 1970 to the end of 1991, based on national, State, and local surveys and data sets, were used to refine and calibrate the model. Additional data for San Diego County, not used in calibrating the model, were compared with model-generated outcomes to test for congruence. In addition to testing the performance of the overall eight-subsystem model, results from each subsystem and components within subsystems were tested individually (e.g., alcohol consumption behaviors of 18-20-year-old males) against results reported in published papers or generated from local surveys. The range of data used for calibrating and testing the model is illustrated in table 1.

Once congruence has been established, a series of prevention interventions will be posited, and their likely effects will be simulated over a 10-year period (1992-2001). The general procedure is to (1) begin with a congruent model, (2) program the model to allow users to alter policy-sensitive variables, (3) run the model with these changes, and (4) compare model outcomes generated using different values for these policy variables. Where possible, the forecasting capability of the model is assessed by contrasting simulated predictions against results obtained when similar policy variables actually were changed in other communities. Where the policy is unique and previously never implemented, or where adequate research results are unavailable, the most closely relevant research findings are selected as points of comparison with model forecasts.

Model Summary

As was illustrated in figure 1, the third-generation model consists of eight interacting subsystems. These are described next in brief.

Consumption Subsystem. The single most critical dynamic in the model is the causal relationships that result in shifting patterns of alcohol consumption over time. The population is assigned to 14 age-sex groups, each of which is tracked and modified separately through the model dynamics. Seven age categories are defined for each gender; these are pooled age groups that exhibit similar drinking patterns or that are likely to be affected similarly by specific interventions (e.g., 18-20-year-old males). These age groups are: 13-17, 18-20, 21-24, 25-34, 35-49, 50-64, and 65 and older.

TABLE 1. *Selected variables for congruence-testing*

SUBSYSTEM	MODEL VARIABLES	SYSTEM MEASURES	DATA SOURCES
CONSUMPTION	Alcohol sales	Annual retail sales by beverage	State Dept. of Revenue
	Age-sex consumption	Local consumption by beverage by age and sex group	Target county*
RETAIL	Alcoholic Beverage Control (ABC) licenses	License counts	State ABC Board
	Alcohol sale permits	Permit counts	State ABC Board
DRINKING & DRIVING	Driver fatalities	Annual number of fatalities	State Dept. of Transportation
	Injury crashes	Annual number of crashes	State Dept. of Transportation
SOCIAL NORMS	Alcohol-related arrests/convictions	DUI, underage sales, public intoxication	Local law enforcement/ criminal justice
	Public pressure/ public concern	Newspaper content analysis counts	On-line/local newspaper
MORBIDITY & MORTALITY	Nontraffic injuries	Annual number of nontraffic injuries	State Dept. of Public Health
	Alcohol-related deaths	Alcohol mortality for selected International Classification of Diseases codes	State Dept. of Public Health
SOCIAL/HEALTH SERVICES	Alcoholism treatment	Annual admissions to treatment	State Div. of Mental Health or Alcohol authority
SOCIAL/ECONOMIC CONSEQUENCES	Alcohol-related family violence	Child abuse/neglect with alcohol involvement	State Dept. of Social Services

KEY: * Test sites should have at least one community consumption survey.

The distribution of average daily drinking behavior of each of the 14 groups, at any point in time, is defined through a lognormal probability-density function. This functional form is unimodal with a peak close to 0 and a strong right skew. That is, within any age-sex group, most drinkers consume at modest levels, and smaller and smaller percentages are found to drink at increasingly larger levels. In addition, a certain percentage of each group (roughly one-third) are known to be abstainers (Clark and Hilton 1991).

As illustrated in figure 2, shifts in these distributions (in the direction of more or less drinking) are triggered by changes in five stimulus factors. These are disposable income, alcohol beverage prices (Cook 1981; Cook and Tauchen 1982; Levy and Sheflin 1983; Ornstein 1980; Saffer and Grossman 1987), alcohol availability (Gruenewald et al. 1993; Holder and Blose 1987; Holder and Wagenaar 1990; MacDonald 1986; Rabow and Watts 1982; Room 1987; Wagenaar and Holder 1991*b*), social norms (Johnson et al. 1990; Linsky et al. 1985; Room 1989; Treno et al. 1993), and enforced minimum drinking age.

Percentage changes in each of the stimulation factors are translated into corresponding changes in average consumption (Brenner 1975; Skog 1986). The net effect of these changes determines the new value for average daily consumption and, in turn, causes a shift in the overall consumption pattern for the group.

Retail Subsystem. This subsystem focuses on the availability of alcohol for on-premises or off-premises consumption (MacDonald and Whitehead 1983; Moskowitz 1989; Ornstein and Hanssens 1985; Rush et al. 1986; Saltz 1987). Depending on a State's Alcoholic Beverage Control (ABC) laws, retail establishments may obtain licenses for the sale of alcohol for consumption at the location of the establishment (e.g., bars, pubs, restaurants, or arenas) or for the sale of alcohol in containers for consumption elsewhere (e.g., wine shops, liquor stores, supermarkets, or convenience stores). The model uses population growth and economic-indicator data (e.g., average disposable income) to explain and predict the

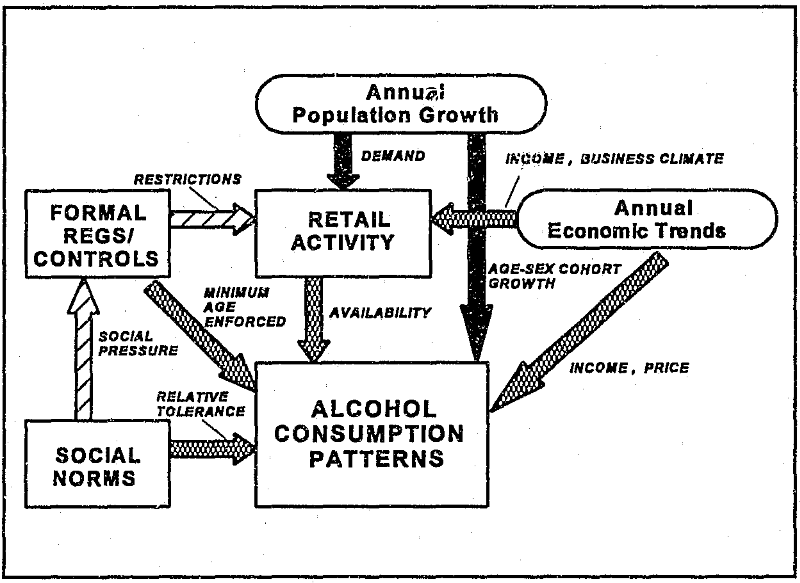


FIGURE 2. *Causal model of alcohol use and alcohol-related trauma: Factors affecting alcohol consumption patterns*

number and types of outlets that are licensed and receive permits to sell alcoholic beverages.

Formal Regulations and Controls Subsystem. This subsystem reproduces the effects of interventions introduced by State or local regulatory agencies during the time period being simulated to influence alcohol retail sales or consumption activity. For example, the number of new licenses of a given type might have been restricted as a means of curbing availability, or ABC enforcement activities and the severity of penalties for sales to minors might have been increased as a means of reducing consumption by underage drinkers. Local zoning options also might have been exercised as a means of lowering densities of establishments in targeted areas (Wittman and Hilton 1987). This subsystem also is used to test future policy options.

Social Norms and Public Pressure Subsystem. In the model, social norms act as stimulus factors that influence levels of alcohol consumption through both positive and negative feedback: "positive" meaning that increases in consumption of alcohol over time are associated with *increased acceptance* for alcohol use, and "negative" meaning that more consumption leads to more drinking-related problems and consequently to *less* social acceptance (Atkin 1987; Atkin et al. 1983; Haskins 1985; Partanen and Montonen 1988; Saffer 1991; Smart 1989; Treno et al. 1993). Ethnic and other sociocultural determinants of drinking behavior (e.g., numbers of college students or military populations) also are accounted for through this subsystem (Caetano 1987*a,* 1987*b,* 1988; Caetano and Mora 1988; Connors et al. 1989; Corbett et al. 1991; Markides et al. 1988).

Drinking and Driving Subsystem. As illustrated in figure 3, "drinking and driving" is one of the four outcome-related components of the simulation model. The distribution of driving events (i.e., trips by vehicle from an origin to a destination) at varying blood alcohol concentration (BAC) levels are computed for the community's population groups (Beitel et al. 1975; Foss et al. 1990; Hingson et al. 1990; Homel 1988; Jonah and Wilson 1983; Levy et al. 1989; Lund and Wolfe 1990; Perrine and Foss 1990; Ross 1982; Ross et al. 1984; Snortum et al. 1986; Voas and Hause 1987; Voas and Williams 1986; Worden et al. 1989). The distribution of such events then is mapped into numbers of driver fatalities and injury crashes. The driving-event distribution by BAC level is derived from considerations of (1) the driving-behavior patterns of the community; (2) the population distribution by age, sex, and consumption levels; (3) the legal driving limit associated with BAC; (4) law enforcement activity; (5) public activity to discourage drinking and driving through pressure and education; (6) perceived risk of being arrested for driving under the influence of alcohol; and (7) perceived risk of being convicted of driving under the influence of alcohol.

Mortality and Morbidity Subsystem. This subsystem employs group-specific risk rates linked to levels of alcohol consumption to convert numbers of persons in the age-sex groups into annual cases of alcohol-associated deaths, illnesses, and nontraffic injuries (Cherpitel 1988, 1989*a*, 1989*b*; Hingson and Howland 1987; Hingson et al. 1988; Holder

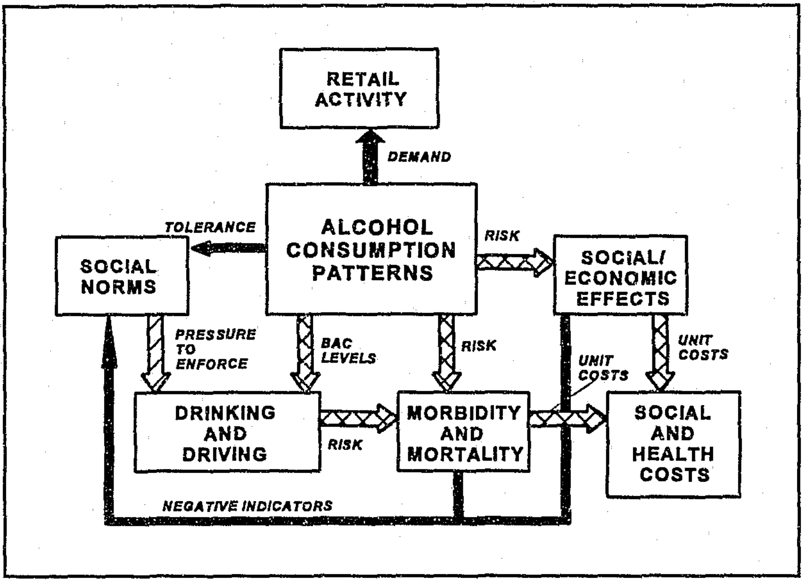


FIGURE 3. *Causal model of alcohol uεe and alcohol-related trauma: Effects of changes in alcohol consumption patterns*

1989; Howland and Hingson 1987, 1988). In the model, increases in these numbers can trigger social activity and, consequently, formal regulatory activity aimed at reducing consumption and/or behaviors associated with alcohol-related problems.

Social and Economic Consequences Subsystem. The consequences of drinking on the family (e.g., numbers of alcohol-related incidents requiring law enforcement intervention) and on the workplace (e.g., alcohol-related accidents) are handled in much the same way as the calculation of mortality and morbidity (i.e., as risk rates applied to each of the groups) (Joksch 1985; Roman 1990). The intent is to reflect that alcohol misuse within a community increases the likelihood (in actuarial terms) of problems such as these. The model is designed to permit a community to select alternative social and economic outcomes for tracking based on local interest and the availability of community indicators for calibrating the model.

Social and Health Services Subsystem. This subsystem reflects the demand for social and health services related to drinking. The general approach is to interrelate levels of consumption with risks of problems and, in turn, with need for treatment. The model is designed to permit a community to select and track increases (or decreases) in alcoholism treatment services, general health care services, and social services. The model yields simulated statistics on patients, treatment facilities, and costs, such as number of new patients, distribution of patients by treatment mode, waiting-list size, average time in treatment, average treatment costs, and insurance benefits received by patients (Hallan and Holder 1986*a*, 1986*b*; Holder 1974).

Model Complexity: An Illustration

For *face validity*, the model must replicate real-system behaviors in reasonable and believable ways (Naylor and Finger 1967; Pidd 1988). That is, researchers and practitioners reviewing how the model works should agree that the components and structures upon which the model is based reflect accepted theories and known behaviors. To illustrate this point, the drinking and driving subsystem will be previewed briefly. The focal point of the drinking and driving subsystem is the derivation of driving events by BAC levels for each of the 14 population groups being tracked. Once derived, risk rates (obtained from national and local data sources, including roadside survey data and coroner reports) are applied to these event counts to compute annual numbers of injury crashes and traffic fatalities.

Two important factors influencing driving events by BAC levels are (1) the actual volume of driving events (i.e., one-way driving trips between an origin and a destination) in the community on days and at times when drinking is most likely to occur, and (2) the legal BAC limit. In the model, drinking and driving events are a function of miles traveled, number of licensed drivers, and the distribution of alcohol beverage consumption by age and sex. National roadside surveys (Farris et al. 1977; Sterling-Smith 1976; Voas and Hause 1987), coupled with local survey data, have been used to estimate driving-event volumes and associated probabilities of driving with different levels of BAC (Foss et al. 1990; Voas and Hause 1987).

Underlying the derivation of driving events by BAC levels is a body of research (Borkenstein 1975; Ross 1982; Ross et al. 1984) linking drinking and driving behavior to the perceived risk of arrest and conviction for DUI. This literature suggests that: (1) perceived risk, rather than actual risk, affects drinking and driving behavior (the actual risk of DUI detection by police is quite small—one arrest in every 2,000 drinking and driving events); (2) large-scale changes in perceived risk are necessary before significant behavior changes can be expected; (3) in the short run, the gap between actual and perceived risk may increase temporarily due to factors like increased enforcement and public announcements of enforcement intentions; but (4) in the long run, perceived risk approaches and approximates actual risk (i.e., pe-ple moderate their perceptions based on experiences and information).

The complete mathematical specification of the subsystem requires more than 20 equations. To calibrate specific variables, he model utilizes (1) published research, (2) data derived through reanalysis of existing data bases, and, in the absence of reliable data, and (3) expert judgment that reflects research findings. To illustrate this process with one variable within the drinking and driving subsystem, consider public perception of DUI (PPDUI) enforcement risk. The concept of public perception of enforcement was introduced (Ross 1982) as an explanatory intervening variable for understanding frequency of driving after or concurrent with drinking. Within any specific community, perception is a dynamic variable: That is, when publicity of DUI enforcement increases, risk estimates are likely to increase, at least for a time, then drop as the public learns from experience that the actual risk of arrest is much lower than its perception (Ross et al. 1984).

Based upon analyses by Homel (1988) and Ross (1982) involving changes in driver behavior resulting from changes in perceived risk, an equation was developed that describes the new perceived risk as a function of the perceived risk in the previous year modified by increases or decreases in actual enforcement and by activities aimed at influencing perceived risk (such as publicity regarding roadside stops on the weekends). The initial value of PPDUI (defined as an index from 0.0 to 1.0) is obtained from (1) community surveys of perceived risk, or from (2) content analysis of media coverage of drinking and driving events, as a proxy for public attention given to this behavior. The enforcement capacity index of the community for DUI is computed (using local enforcement data) as the percentage change in the number of officer hours per time period. Using the equation, annual values for perceived risk are obtained and used to moderate the number of community-specific drinking and driving events.

ADVANTAGES AND LIMITATIONS

A single, general model structure is being posited to capture the principal systems dynamics of communities with respect to alcohol use. This structure is sufficiently general to apply to any community, yet sufficiently detailed to capture the uniqueness of specific communities through initial data loadings. One criticism of the application of computer-modeling to urban land use and transportation planning in the 1970s was that planners in local jurisdictions were trying to deal with the same conceptual and methodological issues while building essentially unique models for each jurisdiction (Kain 1978). A more recent criticism of planning tools with presumed universal applicability is that they are too general, being based on data structures (e.g., spreadsheet layouts) rather than on problem structure (Klosterman 1986). The approach used here has the virtue of not requiring a new model to be built from scratch for each local application while remaining rich in research-based understanding of community behaviors and dynamics.

Type of Effort Demanded

Those interested in expanding the model to include other drugs, or in building separate models focusing on other substances, may well wonder about the difficulties associated with such undertakings. As suggested above, the current modeling effort has been underway for several years and has involved a series of iterations and major design refinements. There are more than 1,000 pages of documentation of work that has been associated with this project. In the next few paragraphs, some of the more important lessons from this experience are reviewed.

Those familiar with statistical models may have trouble grasping how a model comprised of literally 100 variables can be made accurate. The key lies in the nesting of groups of variables. Small clusters of variables are used to explain and predict how specific behaviors or processes perform. These are calibrated and refined independently of the remainder of the model. Sets of clusters then are tested, calibrated, and refined together. Only then is the entire model run and further refined. This process is roughly analogous to that used in creating a portrait or detailed landscape; the areas are first blocked out, then each is attended to on its own terms while the total composition is always kept in mind. In this way, many variables can be introduced and, yet, reliability and stability can be retained.

Perhaps most difficult and time consuming has been the creation of good theory. A working simulation model of a community requires a coherent and consistent understanding of how key processes and behaviors interact over time. Clues to such understanding are available only in fragments in the literature, and there are many missing links. One advantage of current computer technology is that it is relatively easy to build conceptual models and explore relationships between variables. However, deciding on which variables and relationships to retain and, equally critically, what weights to use to quantify these relationships is far from a simple challenge.

Data availability is a constant source of frustration. It has proven difficult to find adequate reliable data to quantify relationships between variables or to serve as benchmarks for use in calibrating subsystems of the model. Although alcohol is a legal substance that is regulated relatively closely, it requires considerable effort and ingenuity to compile national and State data bases of retail activity for use in a model at the current level of sophistication. Alcohol-consumption data are notoriously unreliable (Pernanen 1974; Toneatto et al. 1992). It is typical that results of surveys of drinking behavior account for no more than half of the alcohol known to be consumed by the target population.

The community-systems perspective, as has been discussed here, demands far more than simple study of individual drinking or other drug use patterns. It forces consideration of the larger social and economic context in which community AOD-related problems are embedded. Contrary to the implicit assumption in much prevention research that heavy and problem drinkers are the core of the problem, the use of alcohol within the total community—including its retail price, availability, and community values about acceptable and unacceptable drinking—is, in fact, central to the problem and its solutions. In the end, both researchers and community planners will have to extend their thinking about effective countermeasures beyond those factors that have been considered traditionally. Until such thinking is a regular part of efforts to reduce alcohol-involved problems, prevention activities will operate in a hit-or-miss manner without substantially reducing risks to the community.

Getting Communities To Use the Model

In most applications, models are used to evaluate alternative policies or provide forecasting estimates as feedback to policymakers or planners. This is not the primary role of this dynamic model. Rather, the model is a tool to help the community understand the nature of the complex factors associated with alcohol use and its related problems. Further, it can be an intellectual vehicle for accumulating and synthesizing the best available research in the field, in a sense, serving as an evolving research platform and theory-building mechanism.

Taking a model developed primarily as a research tool and applying it to a community context raises some issues. Such a model may prove to be more complex and require more data than generally are considered necessary for effective community AOD prevention applications. Furthermore, the complexity of a sophisticated, computer-based model might create a considerable gap between model designers and community participants (Brewer 1983; Pugh 1977). To what extent do policymakers need to grasp the complexity of the model's design? To what extent do they need to evaluate the assumptions and extrapolations from research and data bases outside their community to use the model appropriately?

Too frequently, research of value to local decision-makers does not reach them and/or is not understood and used (Giesbrecht et al. 1991; Langendorf 1985). Furthermore, even when the research is valued, it rarely is packaged in forms that can be applied by local practitioners lacking specialized training. However, the type of model under development is less abstract than other mathematically based approaches and should be easier for lay persons to grasp and use, since the mathematics driving the models is hidden beneath the surface logic, which can be explained to the lay person using easy-to-read flow diagrams. When presented with a model of their community with a relatively high degree of accuracy (and furthermore a model that can be used easily to test alternative prevention strategies), it is anticipated that the community will use the model and be interested in understanding it further.

If community leaders or planners are not able (or are unwilling) to undertake the difficult thinking necessary to improve their understanding of their own community system, then the community's selection of prevention interventions will have limited long-term effectiveness. The long-term reduction of alcohol-involved problems requires that community prevention planners get involved personally in some of the same conceptual work and supporting longitudinal research required for a useful computer model, even if computer-modeling technology itself is not central to this involvement.

EMPIRICAL EXAMPLE

As discussed earlier in this chapter, alcohol-related trauma is related to average drinking behavior (i.e., as the average increases, trauma increases in relative proportion). It also was noted that five stimulus factors appear to be associated with shifts in average levels of alcohol consumption: personal income, beverage prices, availability, social norms, and minimum drinking age enforced. The last factor impacts only the first two age groups in the model (13-17-year-olds and 18-20-year-olds), while the first four factors affect all groups. Understanding how each of these factors influences average drinking levels is key to grasping the dynamics of alcohol use within a community and also to choosing the types of interventions that are most likely to reduce alcohol-related trauma. The following presentation describes work done to calibrate the consumption subsystem of the model (i.e., to determine how the five stimulus factors influence average drinking levels within a community).

Based on independent research conducted at PRC (Gruenewald et al. 1993), initial values for elasticities of income, price, and availability were at hand. The elasticity parameters indicate the percentage change in average consumption associated with a 1-percent change in the given stimulus factor. For example, an elasticity of 2.0 means that a 1-percent change in the stimulus factor results in a 2-percent change in average consumption; an elasticity of 0.5 means that a 1-percent change in the stimulus factor results in a one-half-percent change in average consumption. Elasticity measures also may be negative. For example, an elasticity of -0.5 means that a 1-percent increase in the stimulus factor results in a one-half-percent decrease in average consumption. The elasticity parameter associated with income is positive (i.e., income and consumption move in the same direction), as is the elasticity parameter for availability. However, the elasticity parameter for price is negative (i.e., as price increases, consumption decreases, and vice versa).

No estimates for elasticity parameters for enforced minimum drinking age or social norms were available in the literature. It was necessary to estimate these based on selected research findings and analysis of historical data trends. There have been numerous studies of the impact of State laws raising the minimum drinking age from 18 years of age or older to

21 years of age. These provided a rich source of information for establishing a rough initial estimate for the elasticity measure associated with this stimulus factor.

The combined effects of income, price, availability, and minimum drinking age would suggest a continuing upward increase in average drinking throughout the study period (i.e., from 1970 to the end of 1991). Actual national and State (i.e., California) data reflect a decline in average consumption beginning in the early 1980s and continuing through the end of 1991. In accordance with the model formulation, this decline must ɔe accounted for primarily through changes in social norms. Test-ing of different values against these national and State trends led to the selection of an initial value for the elasticity parameter.

With these five elasticity parameters as a starting point, the model was run using national data as inputs that led to the results depicted in figure 4. These elasticity measures became part of the model design and were retained in subsequent tests of the model at State and county levels.

The model next was reinitialized using California data. The results appear in figure 5. As can be seen, the model estimates were very close to the actual reported data, never differing by more than 5 percent across the 18-year period for which data were available. The model values did peak, however, 3 years later than the actual values (i.e., in 1982 rather than 1979), suggesting the need for further refinements.

The consumption subsystem of the model next was tested using data from a representative set of California counties. However, a problem was encountered. There have been few reliable surveys conducted of alcohol consumption at the county level in California. Hence, no benchmark data exist against which to test the model at this level of generalization. As a rough proxy, it was assumed that average consumption in the counties mirrored that of the State as a whole. The results of this exercise appear in figure 6. As can be seen, Santa Clara County estimates matched the actual State pattern very closely. Kern County estimates consistently

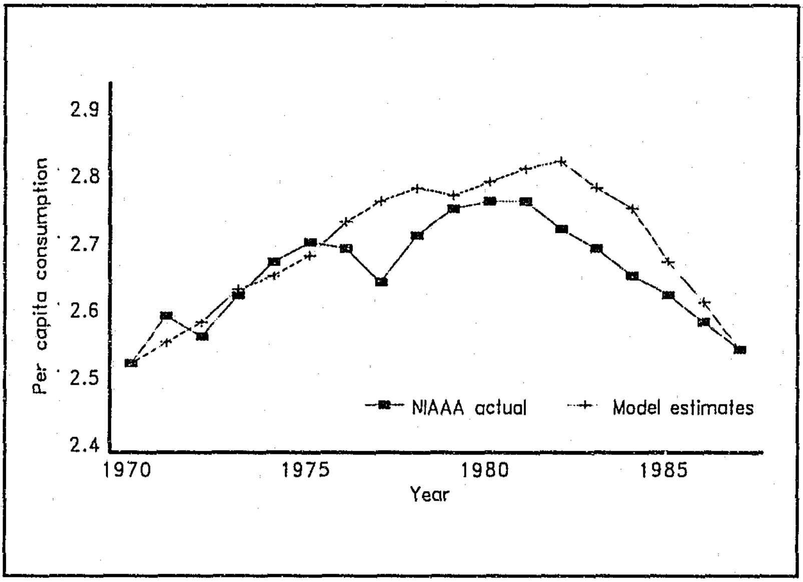


FIGURE 4. *Test of the model against national data: Average consumption, 1970-1987*

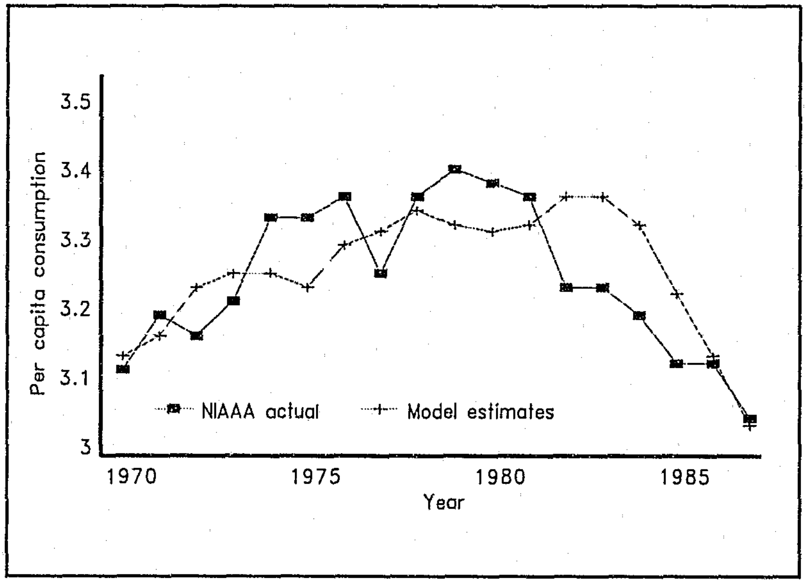


FIGURE 5. *Test of the model against California data: Average consumption, 1970-1987*

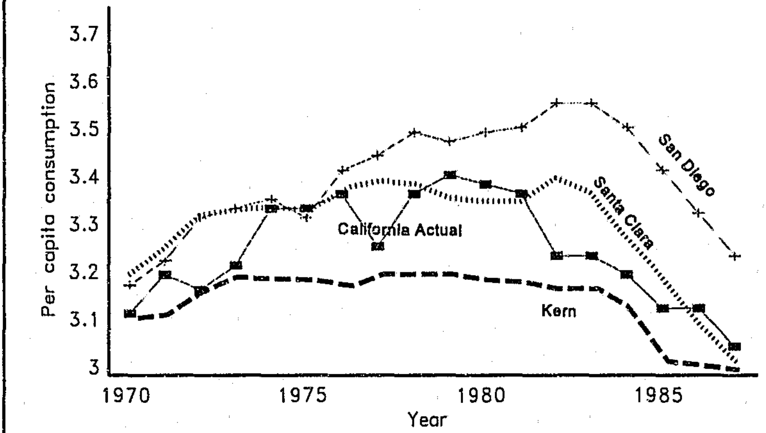


FIGURE 6. *Model estimates for selected California counties: Average consumption, 1970-1987*

were lower than the State pattern, while San Diego County estimates consistently were higher than the State pattern.

San Diego County provides some unique modeling challenges in (1) having a large military population, leading to disproportionately large numbers of young male adults in the county; (2) being close to the Mexican border with implications for both consumption and retail sales patterns; and (3) having a diverse ethnic mix among its populations, reflecting different drinking cultures. To test if the model estimates for consumption are accurate (i.e., the county's population does in fact drink at higher levels than the State population), risk rates are being applied to these estimates and matched against traffic fatalities and other morbidity and mortality data for which county-specific historic data exist. Additional tests will explore how special characteristics of the community can be simulated within the social norms and public pressure subsystem, which feeds back to affect consumption patterns (as discussed above).

The above results in replicating historic trends in alcohol consumption were produced with the sensitivity weights (i.e., elasticities) arrayed in table 2. Thus, for example, a 3-percent change in licenses to sell spirits between 2 years, say 1987 and 1988, translates in the model into a 0.75-percent (3x0.25) increase in average consumption of spirits in 1988. The model then adjusts the consumption distributions for each of the 14 age-sex groups to account for this increase by recomputing the parameters of the associated lognormal distributions. This, in turn, leads to increases in trauma associated with consumption of alcohol.

TABLE 2. *Relationship between stimulus factors and average consumption*

For each 1% change in this stimulus factor:	Corresponding % change in average consumption:
income	0.10 (beer)
income	0.40 (wine)
income	0.20 (spirits)
beer price	-0.03 (beer)
wine price	-0.03 (wine)
spirits price	-0.20 (spirits)
beer availability	0.10 (beer)
wine availability	0.50 (wine)
spirits availability	0.25 (spirits)
social norms	0.35 (all)
minimum drinking age	0.50 (beer, 13-20-year-olds)
minimum drinking age	0.50 (wine, 13-20-year-olds)
minimum drinking age	0.50 (spirits, 13-20-year-olds)

The model allows two basic paths for intervention to reduce alcohol trauma. The first way is by affecting the values of any of the above stimulus factors. The second is by reducing the risk probabilities associated with consumption. An example of the first type of intervention might be increased newspaper coverage of drinking-related trauma, leading to increased social concerns and a positive increase in the social norms index. This increase would translate to a corresponding change in consumption (based on the 0.35 elasticity value). An example of an intervention influencing the risk rates would be suspending drivers licenses for those convicted of DUI offenses. While this approach may reduce consumption indirectly, it has a direct relationship to risk by lowering the number of risky drivers on the road.

Communities will differ with regard to which stimulus factors or risk rates can be changed most readily and at what economic and political costs. The model affords community planners and decision-makers with a mechanism for asking "what if" questions regarding alternative interventions and, consequently, for determining which politically acceptable mix of feasible interventions will yield the highest reduction in alcohol-related trauma. Once determined, the community still must engage in an appropriate planning process (e.g., following a collaborative model) to articulate the implementation specifications for desired interventions. The computer model provides a tool enabling this process to occur in a more informed, less speculative manner.

REFERENCES

Atkin, C.K. Alcoholic-beverage advertising: Its content and impact. In: Holder, H.D., ed. *Control Issues in Alcohol Abuse Prevention: Strategies for States and Communities*. Greenwich, CT: JAI Press, 1987. pp. 267-287.

Atkin, C.K.; Neuendorf, K.; and McDermott, S. The role of alcohol advertising in excessive and hazardous drinking. *J Drug Educ* 13:313-325, 1983.

Beitel, G.A.; Sharp, M.C.; and Glauz, W.D. Probability of arrest while driving under the influence of alcohol. *J Stud Alcohol* 36(1):109-116, 1975.

Blose, J.O., and Holder, H.D. Liquor-by-the-drink and alcohol-related traffic crashes: A natural experiment using time-series analysis. *J Stud Alcohol* 48:52-60, 1987.

Borkenstein, R.F. Problems of enforcement, adjudication and sanctioning. In: Israelstam, S., and Lambert, S., eds. *Alcohol, Drugs and Safety.* Proceedings of the Sixth International Conference on Alcohol, Drugs, and Traffic Safety. Toronto: Addiction Research Foundation of Ontario, 1975.

Box, G.E.P., and Jenkins, G.M. *Time Series Analysis: Forecasting and Control.* London: Holden-Day, Inc., 1976.

Brenner, H.M. Trends in alcohol consumption and associated illnesses: Some effects of economic changes. *Am J Public Health* 65:1279-1292, 1975.

Brewer, G. Some costs and consequences of large scale social systems modeling. *Behav Sci* 28:166-185, 1983.

Caetano, R. Acculturation and drinking patterns among U.S. Hispanics. *Br J Addict* 82:789-799, 1987*a*.

Caetano, R. Alcohol use and depression among U.S. Hispanics. *Br J Addict* 82:1245-1251, 1987*b*.

Caetano, R. A comparative analysis of drinking among Hispanics in the United States, Spaniards in Madrid, and Mexicans in Michoacán. In: Harford, T., and Towle, L., eds. *Cultural Influences and Drinking Patterns: A Focus on Hispanic and Japanese Populations*. Rockville, MD: National Institute on Alcohol Abuse and Alcoholism, 1988. pp. 237-311.

Caetano, R., and Mora, M.E.M. Acculturation and drinking among people of Mexican descent in Mexico and the United States. *J Stud Alcohol* 49(5):462-471, 1988.

Cherpitel, C. Alcohol consumption and casualties: A comparison of two emergency room populations. *Br J Addict* 83:1299-1307, 1988.

Cherpitel, C. Breath analysis and self-reports as measures of alcohol-related emergency room admissions. *J Stud Alcohol* 50(2):155-161, 1989*a*.

Cherpitel,C. A study of alcohol use and injuries among emergency room patients. In: Giesbrecht, N.; Gonzalas, R.; Grant, M.; Osterberg, E.; Room, R.; Rootman, I.; and Towle, L., eds. *Drinking and Casualties: Accidents, Poisonings and Violence in an International Perspective.* London: Tavistock, 1989*b*. pp. 288-299.

Clark, W.B., and Hilton, M.E. *Alcohol in America: Drinking Practices and Problems.* Albany, NY: State University of New York Press, 1991.

Connors, G.J.; Maisto, S.A.; and Watson, E.W. Initial drinking experiences among black and white male and female student drinkers. *Int J Addict* 24(12):1173-1182, 1989.

Cook, I.J.; Dixon, R.T.; Holder, H.D.; Kennedy, F.D.; Sawyer, L.L.; Schlenger, W.E.; and Williams, R.B. *Costs for Alternative Public Inebriate Services: Atlanta, Georgia.* Raleigh, NC: The Human Ecology Institute, 1973.

Cook, P.J. The effect of liquor taxes on drinking, cirrhosis and auto accidents. In: Moore, M.H., and Gerstein, D.R., eds. *Alcohol and Public Policy: Beyond the Shadow of Prohibition.* Washington, DC: National Academy Press, 1981. pp. 255-285.

Cook, P.J., and Tauchen, G. The effect of liquor taxes on heavy drinking. *Bell J Econ* 13(2):379-390, 1982.

Corbett, K.; Mora, J.; and Ames, G. Drinking patterns and drinking-related problems of Mexican-American husbands and wives. *J Stud Alcohol* 52(3):215-223, 1991.

Farris, R.; Malone, T.; and Kirkpatrick, M. *A Comparison of Alcohol Involvement in Exposed and Injured Drivers.* Final Report (DOT Report No. 802-555). Washington, DC: National Highway Traffic Safety Administration, 1977.

Foss, R.D.; Voas, R.B.; Beirness, D.J.; and Wolfe, A.C. *Minnesota 1990 Statewide Drinking and Driving Roadside Survey.* Final Report (Contract 525493). St. Paul: State of Minnesota Department of Public Safety, 1990.

Giesbrecht, N.; Hyndman, B.K.; Bernardi, D.R.; Coston, N.; Douglas, R.R.; Ferrence, R.G.; Gliksman, L.; Goodstadt, M.S.; Graham, D.G.; and Loranger, P.D. "Community Action Research Projects: Integrating Community Interests and Research Agenda in Multicomponent Initiatives." Paper presented at 36th International Institute on the Prevention and Treatment of Alcoholism, Stockholm, June 2-7, 1991.

Gruenewald, P.J.; Ponicki, W.R.; and Holder, H.D. The relationship of outlet densities to alcohol consumption: A time series cross-sectional analysis. *Alcohol Clin Exp Res* 17(1):38-47, 1993.

Hallan, J.B., and Holder, H.D. Analysis of insurance benefit plans for alcoholism treatment through computer simulations. (Part I.) *Comput Psychiatry Psychol* 8(1):12-15, 1986*a*.

Hallan, J.B., and Holder, H.D. Analysis of insurance benefit plans for alcoholism treatment through computer simulations (Part II.) *Comput Psychiatry Psychol* 8(2):12-15, 1986*b*.

Haskins, J.B. The role of mass media in alcohol and highway safety campaigns. *J Stud Alcohol* 10:184-191, 1985.

Hingson, R., and Howland, J. Alcohol as a risk factor for injury or death resulting from accidental falls: A review of the literature. *J Stud Alcohol* 48(3):212-219, 1987.

Hingson, R.W.; Howland, J.; and Levenson, S. Effects of legislative reform to reduce drunken driving and alcohol-related traffic fatalities. *Public Health Rep* 103(6):659-667, 1988.

Hingson, R.H.; Howland, J.; Schiavone, T.; and Damiata, M. The Massachusetts Saving Lives Program: Six cities widening the focus from drunk driving to speeding, reckless driving, and failure to wear safety belts. *J Traffic Med* 18:123-132, 1990.

Holder, H.D. *Alternative Approaches to the Public Inebriate Problem in Metropolitan Areas: A Summary of Findings for Atlanta, Georgia and Baltimore, Maryland.* Raleigh, NC: The Human Ecology Institute, 1974.

Holder, H.D. Drinking, alcohol availability and injuries: A systems model of complex relationships. In: Giesbrecht, N.; Gonzalas, R.; Grant, M.; Osterberg, E.; Room, R.; Rootman, I.; and Towle, L., eds. *Drinking and Casualties: Accidents, Poisonings and Violence in an International Perspective.* London: Associated Book Publishers, 1989. pp. 133-148.

Holder, H.D., and Blose, J.O. Impact of changes in distilled spirits availability on apparent consumption: A time series analysis of liquor-by-the-drink. *Br J Addict* 82(6):623-631, 1987.

Holder, H.D., and Wagenaar, A.C. Effects of the elimination of a state monopoly on distilled spirits' retail sales: A time-series analysis of Iowa. *Br J Addict* 85:1615-1625, 1990.

Holder, H.D., and Wallack, L. Contemporary perspectives of preventing alcohol problems: An empirically-derived model. *J Public Health Policy* 7(3):324-339, 1986.

Homel, R. *Policing and Punishing the Drinking Driver: A Study of General and Specific Deterrence.* New York: Springer-Verlag, 1988.

Howland, J., and Hingson, R. Alcohol as a risk factor for injuries or death due to fires or burns: Review of the literature. *Public Health Rep* 102(5):475-483, 1987.

Howland, J., and Hingson, R. Alcohol as a risk factor for drownings: A review of the literature. *Accid Anal Prev* 20(1):19-25, 1988.

Johnson, R.C.; Nagoshi, C.T.; Danko, G.P.; Honbo, K.M.; and Chau, L.L. Familial transmission of alcohol use norms and expectancies and reported alcohol use. *Alcohol Clin Exp Res* 14(2):216-220, 1990.

Joksch, H.C. Review of the major risk factors. *J Stud Alcohol* 10:47-53, 1985.

Jonah, B.A., and Wilson, R.J. Improving the effectiveness of drinking-driving enforcement through increased efficiency. *Accid Anal Prev* 15(6):463-481, 1983.

Kain, J. The use of computer simulation models for policy analysis. *J Urban Anal* 5:175-189, 1978.

Klosterman, R.E. An assessment of three microcomputer software packages for planning analysis. *Am Plann Assoc J* 52:199-202, 1986.

Langendorf, R. Computers and decision making. *Am Plann Assoc J* 51:422-433, 1985.

Levy, D.; Shea, D.; and Asch, P. Traffic safety effects of sobriety checkpoints and other local DWI programs in New Jersey. *Am J Public Health* 79(3):291-293, 1989.

Levy, D., and Sheflin, N. New evidence on controlling alcohol use through price. *J Stud Alcohol* 44:920-937, 1983.

Linsky, A.S.; Strauss, M.A.; and Colby, J.P., Jr. Stressful events, stressful conditions and alcohol problems in the United States: A partial test of Bale's theory. *J Stud Alcohol* 46:72-80, 1985.

Lund, A.K., and Wolfe, A.C. *Changes in the Incidence of Alcohol-Impaired Driving in the United States, 1973-1986.* Arlington, VA: Insurance Institute for Highway Safety, 1990.

MacDonald, S. The impact of increased availability of wine in grocery stores on consumption: Four case histories. *Br J Addict* 81(3):381-387, 1986.

MacDonald, S., and Whitehead, P.C. Availability of outlets and consumption of alcoholic beverages. *J Drug Issues* 4:477-486, 1983.

Maloff, D.; Becker, H.S.; Fonaroff, A.; and Rodin, J. Informal social controls and their influence on substance use. *J Drug Issues* 2:161-184, 1979.

Markides, K.S.; Krasue, N.; and Mendes de Leon, C.F. Acculturation and alcohol consumption among Mexican Americans: A three-generation study. *Am J Public Health* 78(9):1178-1181, 1988.

McCleary, R.; Hay, R.A.; McDowall, D.; and Meidinger, E.E. *Applied Time Series Analysis for the Social Sciences.* Beverly Hills, CA and London: Sage Publications, 1980.

Moskowitz, J.M. The primary prevention of alcohol problems: A critical review of the research literature. *J Stud Alcohol* 50(1):54-88, 1989.

Naylor, T.H., and Finger, J.M. Verification of computer simulation models. *Manag Sci* 14(2):B.92-B.101, 1967.

Orlando, M.A. The challenge of evaluating community-based prevention programs: A cultural perspective. In: Orlando, M.A., ed. *Cultural Competence for Evaluators.* CSAP Cultural Competence Series I. DHHS Pub. No. (ADM)92-1884. Washington, DC: Supt. of Docs., U.S. Govt. Print. Off., 1992*a*. pp. 1-22.

Orlando, M.A., ed. *Cultural Competence for Evaluators.* CSAP Cultural Competence Series I. DHHS Pub. No. (ADM)92-1884. Washington, DC: Supt. of Docs., U.S. Govt. Print. Off., 1992*b*.

Ornstein, S. Control of alcohol consumption through price increase. *J Stud Alcohol* 412:807-818, 1980.

Ornstein, S., and Hanssens, D.M. Alcohol control laws and the consumption of distilled spirits and beer. *J Consum Res* 12:200-213, 1985.

Partanen, J., and Montonen, M. "Alcohol and the Mass Media." EURO Reports and Studies 108. Copenhagen: World Health Organization, Regional Office for Europe, 1988.

Pernanen, K. Validity of survey data on alcohol use. In: Gibbins, R.J.; Israel, Y.; Kalant, H.; Popham, R.E.; Schmidt, W.; and Smart, R.G., eds. *Research Advances in Alcohol and Drug Problems.* Vol. 1. New York: Wiley, 1974. pp. 355-374.

Perrine, M.W., and Foss, R.D. *The 1990 Roadside Survey of Drinking and Driving in Northeastern Ohio: Final Report for 1990.* North Canton, OH: The Human Ecology Institute, 1990.

Pidd, M. *Computer Simulation in Management Science.* New York: John Wiley and Sons, 1988. pp. 9-10.

Pugh, R.E. *Evaluation of Policy Simulation Models.* Washington, DC: Information Resources Press, 1977.

Rabow, J., and Watts, R.K. Alcohol availability, alcoholic beverage sales and alcohol-related problems. *J Stud Alcohol* 44:767-801, 1982.

Roman, P.M., ed. *Alcohol Problem Intervention in the Workplace: Employee Assistance Programs and Strategic Alternatives.* New York: Quorum Books, 1990.

Room, R. Alcohol monopolies in the U.S.: Challenges and opportunities. *J Public Health Policy* 8(4):509-530, 1987.

Room, R. Cultural changes in drinking and trends in alcohol problem indicators: Recent U.S. experience. In: Waahlberg, R., ed. *Prevention and Control/Realities and Aspirations.* Proceedings of the 35th International Congress on Alcoholism and Drug Dependence. Geneva: International Council on Alcohol and Addictions, 1989.

Ross, H.L. *Deterring the Drinking Driver: Legal Policy and Social Control.* Lexington, MA: D.C. Heath and Company, 1982.

Ross, H.L.; Klette, H.; and McCleary, R. Liberalization and rationalization of drunk driving laws in Scandinavia. *Accid Anal Prev* 16(5/6):471-487, 1984.

Rush, B.; Gliksman, L.; and Brook, R. Alcohol availability, alcohol consumption, and alcohol-related damage. *J Stud Alcohol* 47:1-10, 1986.

Saffer, H. Alcohol advertising, bans and alcohol abuse: An international perspective. *J Health Econ* 10:65-79, 1991.

Saffer, H., and Grossman, M. Beer taxes, the legal drinking age, and youth motor vehicle fatalities. *J Leg Stud* 16:351-374, 1987.

Saltz, R. The roles of bars and restaurants in preventing alcohol-impaired driving. *Eval Health Prof* 10:5-27, 1987.

Schlenger, W.; Haywood, B.; and Hallan, J. *Simulation Study of the Impact of Occupational Programs.* Raleigh, NC: The Human Ecology Institute, 1976.

Skog, O.-J. An analysis of divergent trends in alcohol consumption and economic development. *J Stud Alcohol* 47(1):19-25, 1986.

Smart, R.G. Does alcohol advertising affect overall consumption: A review of empirical studies. *J Stud Alcohol* 49(4):314-323, 1989.

Snortum, J.R.; Burger, D.E.; and Hauge, R. Deterring alcohol-impaired driving: A comparative analysis of compliance in Norway and the United States. *Justice Q* 3(2):139-165, 1986.

Sterling-Smith, R.S. *Psychosocial Identification of Drivers Responsible for Fatal Vehicular Accidents in Boston.* DOT Report No. 801-95. Washington, DC: National Highway Traffic Safety Administration, 1976.

Stewart, K., and Klitzner, M. "Alcohol and Other Drug Problem Prevention From a Public Health Perspective." Background paper prepared for the Center for Substance Abuse Prevention Planning Work Group Meeting, 1992.

Summers, L.G., and Harris, D.H. *The General Deterrence of Driving While Intoxicated. Vol. I, System Analysis and Computer-Based Simulation.* NTIS No. PB-288 112. Washington, DC: Supt. of Docs., U.S. Govt. Print. Off., 1978.

Tiao, G.C., and Box, G.E.P. Modeling multiple time series with application. *J Am Stat Assoc* 96:802-816, 1981.

Toneatto, T.; Sobell, L.C.; and Sobell, M.B. Predictors of alcohol abusers' inconsistent self-reports of their drinking and life events. *Alcohol Clin Exp Res* 16(3):542-546, 1992.

Treno, A.J.; Parker, R.N.; and Holder, H.D. Understanding U.S. alcohol consumption with social and economic factors: A multivariate series analysis, 1950-1986. *J Stud Alcohol* 54:146-156, 1993.

Voas, R.B., and Hause, J.M. Deterring the drinking driver: The Stockton experience. *Accid Anal Prev* 19:81-90, 1987.

Voas, R.B., and Williams, A.F. Age differences of arrested and crash-involved drinking drivers. *J Stud Alcohol* 47(3):244-248, 1986.

Wagenaar, A.C. Preventing highway crashes by raising the legal minimum age for drinking: The Michigan experience six years later. *J Saf Res* 17:101-109, 1986.

Wagenaar, A.C., and Holder, H.D. A change from public to private sale of wine: Results from natural experiments in Iowa and West Virginia. *J Stud Alcohol* 52(2):162-173, 1991*b*.

Wagenaar, A.C., and Holder, H.D. Effects of alcoholic beverage server liability on traffic crash injuries. *Alcohol Clin Exp Res* 15:942-947, 1991*a*.

Wittman, F.D., and Hilton, M.E. Uses of planning and zoning ordinances to regulate alcohol outlets in California cities. In: Holder, H.D., ed. *Control Issues in Alcohol Abuse Prevention: Strategies for States and Communities.* Supplement 1. Greenwich, CT: JAI Press, 1987. pp. 337-366.

Worden, J.K.; Flynn, B.S.; Merrill, D G.; Waller, J.A.; and Haugh, L.D. Preventing alcohol-impaired driving through community self-regulation training. *Am J Public Health* 79(3):287-290, 1989.

ACKNOWLEDGMENT

Research and preparation of this paper were supported in part by National Institute on Alcohol Abuse and Alcoholism Research Center grant AA06282 to the Prevention Research Center of the Pacific Institute for Research and Evaluation.

AUTHORS

Barry M. Kibel, Ph.D.
Senior Research Scientist
Pacific Institute for Research and Evaluation
121 West Rosemary Street
Chapel Hill, NC 27516

Harold D. Holder, Ph.D.
Director
Prevention Research Center
and
Senior Scientist
Pacific Institute for Research and Evaluation
2532 Durant Avenue
Berkeley, CA 94704

National
Institute on
Drug
Abuse
Research
MONOGRAPH SERIES

While limited supplies last, single copies of the monographs may be obtained free of charge from the National Clearinghouse for Alcohol and Drug Information (NCADI). Please contact NCADI also for information about availability of coming issues and 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 are also 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 1975-1993 and those not listed, write to NIDA, Community and Professional Education 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.)
NTIS PB #80-118755/AS (A09) $27.00
NCADI #M26

42 THE ANALYSIS OF CANNABINOIDS IN BIOLOGICAL FLUIDS. Richard L. Hawks, Ph.D., ed.
NTIS PB #83-136044/AS (A07) $27.00
NCADI #M42

50 COCAINE: PHARMACOLOGY, EFFECTS, AND TREATMENT OF ABUSE. John Grabowski, Ph.D., ed.
NTIS PB #85-150381/AS (A07) $27.00
NCADI #M50

52 TESTING DRUGS FOR PHYSICAL DEPENDENCE POTENTIAL AND ABUSE LIABILITY. Joseph V. Brady, Ph.D., and Scott E. Lukas, Ph.D., eds.
NTIS PB #85-150373/AS (A08) $27.00
NCADI #M52

53 PHARMACOLOGICAL ADJUNCTS IN SMOKING CESSATION. John Grabowski, Ph.D., and Sharon M. Hall, Ph.D., eds.
NTIS PB #89-123186/AS (A07) $27.00
NCADI #M53

54 MECHANISMS OF TOLERANCE AND DEPENDENCE. Charles Wm. Sharp, Ph.D., ed.
NTIS PB #89-103279/AS (A19) $52.00
NCADI #M54

56 ETIOLOGY OF DRUG ABUSE: IMPLICATIONS FOR PREVENTION. Coryl LaRue Jones, Ph.D., and Robert J. Battjes, D.S.W., eds.
NTIS PB #89-123160/AS (A13) $36.50
NCADI #M56

61 COCAINE USE IN AMERICA: EPIDEMIOLOGIC AND CLINICAL PERSPECTIVES. Nicholas J. Kozel, M.S., and Edgar H. Adams, M.S., eds.
NTIS PB #89-131866/AS (A11) $36.50
NCADI #M61

62 NEUROSCIENCE METHODS IN DRUG ABUSE RESEARCH. Roger M. Brown, Ph.D., and David P. Friedman, Ph.D., eds.
NTIS PB #89-130660/AS (A08) $27.00
NCADI #M62

63 PREVENTION RESEARCH: DETERRING DRUG ABUSE AMONG CHILDREN AND ADOLESCENTS. Catherine S. Bell, M.S., and Robert J. Battjes, D.S.W., eds.
NTIS PB #89-103287/AS (A11) $36.50
NCADI #M63

64 PHENCYCLIDINE: AN UPDATE. Doris H. Clouet, Ph.D., ed.
NTIS PB #89-131858/AS (A12) $36.50
NCADI #M64

65 WOMEN AND DRUGS: A NEW ERA FOR RESEARCH. Barbara A. Ray, Ph.D., and Monique C. Braude, Ph.D., eds.
NTIS PB #89-130637/AS (A06) $27.00
NCADI #M65

69 OPIOID PEPTIDES: MEDICINAL CHEMISTRY. Rao S. Rapaka, Ph.D.; Gene Barnett, Ph.D.; and Richard L. Hawks, Ph.D., eds.
NTIS PB #89-158422/AS (A17) $44.50
NCADI #M69

70 OPIOID PEPTIDES: MOLECULAR PHARMACOLOGY, BIOSYNTHESIS, AND ANALYSIS. Rao S. Rapaka, Ph.D., and Richard L. Hawks, Ph.D., eds.
NTIS PB #89-158430/AS (A18) $52.00
NCADI #M70

72 RELAPSE AND RECOVERY IN DRUG ABUSE. Frank M. Tims, Ph.D., and Carl G. Leukefeld, D.S.W., eds.
NTIS PB #89-151963/AS (A09) $36.50
NCADI #M72

74 NEUROBIOLOGY OF BEHAVIORAL CONTROL IN DRUG ABUSE. Stephen I. Szara, M.D., D.Sc., ed.
NTIS PB #89-151989/AS (A07) $27.00
NCADI #M74

78 THE ROLE OF NEUROPLASTICITY IN THE RESPONSE TO DRUGS. David P. Friedman, Ph.D., and Doris H. Clouet, Ph.D., eds.
NTIS PB #88-245683/AS (A10) $36.50
NCADI #M78

79 STRUCTURE-ACTIVITY RELATIONSHIPS OF THE CANNABINOIDS. Rao S. Rapaka, Ph.D., and Alexandros Makriyannis, Ph.D., eds.
NTIS PB #89-109201/AS (A10) $36.50
NCADI #M79

80 NEEDLE SHARING AMONG INTRAVENOUS DRUG ABUSERS: NATIONAL AND INTERNATIONAL PERSPECTIVES. Robert J. Battjes, D.S.W., and Roy W. Pickens, Ph.D., eds.
NTIS PB #88-236138/AS (A09) $36.50
NCADI #M80

82 OPIOIDS IN THE HIPPOCAMPUS. Jacqueline F. McGinty, Ph.D., and David P. Friedman, Ph.D. eds.
NTIS PB #88-245691/AS (A06) $27.00
NCADI #M82

83 HEALTH HAZARDS OF NITRITE INHALANTS. Harry W. Haverkos, M.D., and John A. Dougherty, Ph.D., eds.
NTIS PB #89-125496/AS (A06) $27.00
NCADI #M83

84 LEARNING FACTORS IN SUBSTANCE ABUSE. Barbara A. Ray, Ph.D., ed.
NTIS PB #89-125504/AS (A10) $36.50
NCADI #M84

85 EPIDEMIOLOGY OF INHALANT ABUSE: AN UPDATE. Raquel A. Crider, Ph.D., and Beatrice A. Rouse, Ph.D., eds.
NTIS PB #89-123178/AS (A10) $36.50
NCADI #M85

87 OPIOID PEPTIDES: AN UPDATE. Rao S. Rapaka, Ph.D. and Bhola N. Dhawan, M.D., eds.
NTIS PB #89-158430/AS (A11) $36.50
NCADI #M87

88 MECHANISMS OF COCAINE ABUSE AND TOXICITY. Doris H. Clouet, Ph.D., Khursheed Asghar, Ph.D., and Roger M. Brown, Ph.D., eds.
NTIS PB #89-125512/AS (A16) $44.50
NCADI #M88

89 BIOLOGICAL VULNERABILITY TO DRUG ABUSE. Roy W. Pickens, Ph.D., and Dace S. Svikis, B.A., eds.
NTIS PB #89-125520/AS (A09) $27.00
NCADI #M89

92 TESTING FOR ABUSE LIABILITY OF DRUGS IN HUMANS. Marian W. Fischman, Ph.D.; and Nancy K. Mello, Ph.D., eds.
NTIS PB #90-148933/AS (A17) $44.50
NCADI #M92

94 PHARMACOLOGY AND TOXICOLOGY OF AMPHETAMINE AND RELATED DESIGNER DRUGS. Khursheed Asghar, Ph.D.; Errol De Souza, Ph.D., eds.
NTIS PB #90-148958/AS (A16) $44.50
NCADI #M94

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.
NTIS PB #90-237660/AS (A99) $67.00
NCADI #M95

96 DRUGS OF ABUSE: CHEMISTRY, PHARMACOLOGY, IMMUNOLOGY, AND AIDS. Phuong Thi Kim Pham, Ph.D. and Kenner Rice, Ph.D., eds.
NTIS PB #90-237678/AS (A11) $36.50
NCADI #M96

97 NEUROBIOLOGY OF DRUG ABUSE: LEARNING AND MEMORY. Lynda Erinoff, ed.
NTIS PB #90-237686/AS (A11) $36.50
NCADI #M97

98 THE COLLECTION AND INTERPRETATION OF DATA FROM HIDDEN POPULATIONS. Elizabeth Y. Lambert, M.S., ed.
NTIS PB #90-237694/AS (A08) $27.00
NCADI #M98

99 RESEARCH FINDINGS ON SMOKING OF ABUSED SUBSTANCES. C. Nora Chiang, Ph.D. and Richard L. Hawks, Ph.D., eds.
NTIS PB #91-141119 (A09) $27.00
NCADI #M99

100 DRUGS IN THE WORKPLACE: RESEARCH AND EVALUATION DATA. VOL II. Steven W. Gust, Ph.D.; and J. Michael Walsh, Ph.D., eds.
GPO Stock #017-024-01458-3 $8.00
NCADI #M100

101 RESIDUAL EFFECTS OF ABUSED DRUGS ON BEHAVIOR. John W. Spencer, Ph.D. and John J. Boren, Ph.D., eds.
NTIS PB #91-172858/AS (A09) $27.00
NCADI #M101

102 ANABOLIC STEROID ABUSE. Geraline C. Lin, Ph.D. and Lynda Erinoff, Ph.D., eds.
NTIS PB #91-172866/AS (A11) $36.50
NCADI #M102

106 IMPROVING DRUG ABUSE TREATMENT. Roy W. Pickens, Ph.D.; Carl G. Leukefeld, D.S.W.; and Charles R. Schuster, Ph.D., eds.
NTIS PB #92-105873(A18) $50.00
NCADI #M106

107 DRUG ABUSE PREVENTION INTERVENTION RESEARCH: METHODOLOGICAL ISSUES. Carl G. Leukefeld, D.S.W., and William J. Bukoski, Ph.D., eds.
NTIS PB #92-160985 (A13) $36.50
NCADI #M107

108 CARDIOVASCULAR TOXICITY OF COCAINE: UNDERLYING MECHANISMS. Pushpa V. Thadani, Ph.D., ed.
NTIS PB #92-106608 (A11) $36.50
NCADI #M108

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.
NTIS PB #92-106616 (A08) $27.00
NCADI #M109

111 MOLECULAR APPROACHES TO DRUG ABUSE RESEARCH: VOLUME I. Theresa N.H. Lee, Ph.D., ed.
NTIS PB #92-135743 (A10) $36.50
NCADI #M111

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.
NTIS PB #92-155449 (A15) $44.50
NCADI #M112

113 ECONOMIC COSTS, COST EFFECTIVENESS, FINANCING, AND COMMUNITY-BASED DRUG TREATMENT. William S. Cartwright, Ph.D., and James M. Kaple, Ph.D., eds.
NTIS PB #92-155795 (A10) $36.50
NCADI #M113

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.
NTIS PB #92-146216 (A16) $44.50
NCADI #M114

115 METHAMPHETAMINE ABUSE: EPIDEMIOLOGIC ISSUES AND IMPLICATIONS. Marissa A. Miller, D.V.M., M.P.H., and Nicholas J. Kozel, M.S., eds.
NTIS PB # 92-146224/II (AO7) $27.00
NCADI #M115

116 DRUG DISCRIMINATION: APPLICATIONS TO DRUG ABUSE RESEARCH. R.A. Glennon, Ph.D., T.U.C. Jarbe, Ph.D., and J. Frankenheim, Ph.D., eds.
NTIS PB # 94-169471 (A20) $52.00
NCADI #M116

117 METHODOLOGICAL ISSUES IN EPIDEMIOLOGY, PREVENTION, AND TREATMENT RESEARCH ON DRUG-EXPOSED WOMEN AND THEIR CHILDREN. M. M. Kilbey, Ph.D. and K. Asghar, Ph.D., eds.
GPO Stock #O17-024-01472-9 $12.00
NTIS PB #93-102101/LL (A18) $52.00
NCADI #M117

118 DRUG ABUSE TREATMENT IN PRISONS AND JAILS. C.G. Leukefeld, D.S.W. and F. M. Tims, Ph.D., eds.
GPO Stock #017-024-01473-7 $16.00
NTIS PB #93-102143/LL (A14) $44.50
NCADI #M118

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
NTIS PB #92-214956/LL (A12) $36.50
NCADI #M120

121 BUPRENORPHINE: AN ALTERNATIVE TREATMENT FOR OPIOID DEPENDENCE. Jack D. Blaine, Ph.D. ed.
GPO Stock #017-024-01482-6 $5.00
NTIS PB #93-129781/LL (A08) $27.00
NCADI #M121

123 ACUTE COCAINE INTOXICATION: CURRENT METHODS OF TREATMENT. Heinz Sorer, Ph.D., ed.
GPO# 017-024-01501-6 $6.50
NTIS PB #94-115433/LL (A09) $27.00
NCADI #M123

124 NEUROBIOLOGICAL APPROACHES TO BRAIN-BEHAVIOR INTERACTION. Roger M. Brown, Ph.D., and Joseph Fracella, Ph.D., eds.
GPO #017-024-01492-3 $9.00
NTIS PB #93-203834/LL (A12) $36.50
NCADI #M124

125 ACTIVATION OF IMMEDIATE EARLY GENES BY DRUGS OF ABUSE. Reinhard Grzanna, Ph.D., and Roger M. Brown, Ph.D., eds.
GPO# 017-024-01503-2 $7.50
NTIS PB # 94-169489 (A12) $36.50
NCADI #M125

126 MOLECULAR APPROACHES TO DRUG ABUSE RESEARCH VOLUME II: STRUCTURE, FUNCTION, AND EXPRESSION. Theresa N.H. Lee, Ph.D., eds.
NTIS PB # 94-169497 (A08) $27.00
NCADI #M126

127 PROGRESS AND ISSUES IN CASE MANAGEMENT. Rebecca Sager Ashery, D.S.W., ed.
NTIS PB # 94-169505 (A18) $52.00
NCADI #M127

128 STATISTICAL ISSUES IN CLINICAL TRIALS FOR TREATMENT OF OPIATE DEPENDENCE. Ram B. Jain, Ph.D., ed.
NTIS PB #93-203826/LL (A09) $27.00
NCADI #M128

129 INHALANT ABUSE: A VOLATILE RESEARCH AGENDA. Charles Wm. Sharp, Ph.D., Fred Beauvais, Ph.D., and Richard Spence, Ph.D., eds.
GPO #017-024-01496-6 $12.00
NTIS PB #93-183119/LL (A15) $44.50
NCADI #M129

130 DRUG ABUSE AMONG MINORITY YOUTH: ADVANCES IN RESEARCH AND METHODOLOGY. Mario De La Rosa, Ph.D., Juan-Luis Recio Adrados, Ph.D., eds.
GPO #017-024-01506-7 $14.00
NTIS PB # 94-169513 (A15) $44.50
NCADI #M130

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 #017-024-01505-9 $14.00
NTIS PB # 94-169521 (A15) $44.50
NCADI #M131

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# 017-024-01502-4 $23.00
NTIS PB #94-115508/LL (A99)
NCADI #M132

133 SIGMA, PCP, AND NMDA RECEPTORS. Errol B. De Souza, Ph.D., Doris Clouet, Ph.D., and Edythe D. London, Ph.D., eds.
NTIS PB # 94-169539 (A12) $36.50
NCADI #M133

134 MEDICATIONS DEVELOPMENT: DRUG DISCOVERY, DATABASES, AND COMPUTER-AIDED DRUG DESIGN. Rao S. Rapaka, Ph.D and Richard L. Hawks, Ph.D., eds.
GPO #017-024-01511-3 $11.00
NTIS PB # 94-169547 (A14) $44.50
NCADI #M134

135 COCAINE TREATMENT: RESEARCH AND CLINICAL PERSPECTIVES. Frank M. Tims, Ph.D. and Carl G. Leukefeld, D.S.W., eds.
GPO #017-024-01520-2 $11.00
NTIS PB # 94-169554 (A13) $36.50
NCADI #M135

136 ASSESSING NEUROTOXICITY OF DRUGS OF ABUSE. Lynda Erinoff, Ph.D., ed.
GPO #017-024-01518-1 $11.00
NTIS PB # 94-169562 (A13) $36.50
NCADI #M136

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 #017-024-01519-9 $13.00
NTIS PB # 94-169570 (A15) $44.50
NCADI #M137

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. 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. VOLUME II: ABSTRACTS. Louis S. Harris, Ph.D., ed.
NCADI #M141

NIH Publication No. 94-3599
Printed 1994