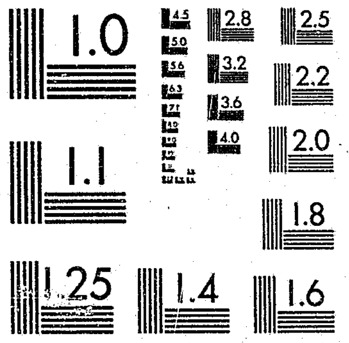


95231

FINAL REPORT

ALTERNATIVES TO ANALYSIS OF COVARIANCE FOR ESTIMATING TREATMENT EFFECTS IN CRIMINAL JUSTICE EVALUATION: COMPARATIVE RESULTS

Rudy A. Haapanen, Ph.D.
California Department of the Youth Authority
4241 Williamsbourgh Drive
Sacramento, CA 95823
(916) 445-1143

with

James C. Cramer, Ph.D.
University of California, Davis
Davis, CA 95616
(916) 752-0809

95231

State of California
GEORGE DEUKMEJIAN,
Governor

Youth and Adult Correctional Agency
N. A. CHADERJIAN,
Secretary

Department of the Youth Authority

PROGRAM RESEARCH AND REVIEW DIVISION

ELAINE DUXBURY
Chief

* * * * *

CARL F. JESNESS, PH.D.
Research Manager III

* * *

RUDY A. HAAPANEN, PH.D.
Principal Investigator

* * *

JAMES C. CRAMER, PH.D.
Consultant

JAMES ROWLAND,
DIRECTOR

GEORGE R. ROBERTS,
CHIEF DEPUTY DIRECTOR

FRANCISCO J. ALARCON,
Deputy Director
ADMINISTRATIVE SERVICES BRANCH

RONALD W. HAYES
Deputy Director
PREVENTION AND COMMUNITY CORRECTIONS BRANCH

WILBUR A. BECKWITH
Deputy Director
PAROLE SERVICES BRANCH

C. A. TERHUNE,
Deputy Director
INSTITUTIONS AND CAMPS BRANCH

95231

ALTERNATIVES TO ANALYSIS OF COVARIANCE FOR ESTIMATING TREATMENT EFFECTS IN CRIMINAL JUSTICE EVALUATION: COMPARATIVE RESULTS

Rudy A. Haapanen, Ph.D.
California Department of the Youth Authority

with

James C. Cramer, Ph.D.
University of California, Davis

July 1982

Preparation of this report was financially assisted through grant number 80-IJ-CX-0045 from the National Institute of Justice, U.S. Department of Justice. Findings, conclusions, and opinions presented in this report are those of the authors, and not necessarily those of the U.S. Department of Justice.

CONTENTS

LIST OF TABLES

LIST OF TABLES (Continued)

LIST OF FIGURES

ABSTRACT

Treatment effect estimates obtained by applying analysis of covariance (ANCOVA) in the nonequivalent control group situation are likely to be biased because the covariates are imperfect measures of the underlying confounding factors, the dependent variable may have a limited range and skewed distribution, and interactions between the variables (including the factors) are often omitted from the model. The present research applied, in addition to ANCOVA, two alternative analytic approaches (LISREL and a factor-loglinear-regression approach) designed to overcome one or more of these problems, to two data sets typical of criminal justice evaluations. Each approach posited one or more general factors (unobserved variables) of which pretest scales were indicators measured with error. Scores on these factors were estimated and used in place of the observed covariates for adjusting outcome scores. The approaches differ in a) the extent to which they can correctly estimate the effects of the underlying factors, and b) their abilities to determine interaction effects in the data and adjust for a limited dependent variable. Results of these comparative analyses showed that with these data, overcoming problems associated with traditional ANCOVA did not lead to substantial differences in treatment effect estimates. These results tend to confirm the commonly-recognized robustness of ANCOVA, using ordinary least squares estimation, and suggest that it can provide reasonably good estimates of treatment effects with criminal justice data. The other methods were found to have certain advantages, however, providing information not obtained with ANCOVA, and did improve estimates slightly. Their use is not discouraged, but their complexity and/or unavailability make them less practical for most evaluative analyses using criminal justice data.

CHAPTER I

Criminal Justice Evaluation and Analysis of Covariance

Evaluation research in the field of criminal justice is often faced with the problem of dissimilarities in the characteristics of a group exposed to a particular treatment and the group used for estimating outcomes in the absence of the treatment. The problem may arise as the result of using quasi-experimental (nonequivalent control group) designs or as a consequence of unanticipated problems in implementing the more powerful true-experimental designs involving random assignment to treatment and control groups. A common method of attempting to control for preexisting differences between the groups is through the use of analysis of covariance (ANCOVA). This procedure controls statistically for the effects of those variables in which the groups differ and provides an estimate of the difference in the adjusted group means on the outcome variable. It is a flexible, easy-to-use, and widely-available technique which provides the researcher with easily interpretable estimates of the effects of the treatment or intervention of interest. However, the technique has come under considerable criticism in recent years (Palmer and Carlson, 1976; Reichardt, 1979), most of which has focused on violations of the underlying assumptions of the method and the sensitivity of ANCOVA to problems with the kinds of data typically used in evaluative research.

In response to these criticisms, a number of alternative approaches have been suggested for analyzing data obtained from group-comparison evaluations that, either by design or by chance, fail to employ truly equivalent control groups (e.g., Reichardt, 1979; Kenny, 1979; Blumstein and Cohen, 1979; Linn and Werts, 1977; Sörbom, 1978; Rindskopf, 1981). These techniques are designed to overcome certain of the problems inherent in the use of ANCOVA. Although theoretically superior to ANCOVA, however, these methods have not as yet been demonstrated as providing the increase in accuracy that would call for an abandonment of ANCOVA as the preferred analytical tool.

In the present study, two major alternative methods for analyzing evaluation data were applied, along with ANCOVA, to two different data sets and the results compared. Although the research was not designed to provide definitive answers to questions regarding the relative superiority of the methods, a comparison of the results should shed some light on the importance of certain problems identified with the ANCOVA method as well as introduce the reader to the alternatives and their merits. The two data sets, one from a quasi-experimental evaluation of Youth Service Bureaus (YSBs) and one from an experimental evaluation of a California Youth Authority institutional program, posed a variety of problems for analyzing treatment effects, many of which are common to most evaluations involving criminal justice or delinquency programs. Our results, then, should have implications for the analysis of other evaluation data as well. Of primary interest was whether the results from the various methods differed with respect to their implied conclusions about the effectiveness of the programs. Of additional interest was whether the alternative methods were able to cope more effectively than ANCOVA with specific problems in these data.

In the pages that follow, we outline the logic, assumptions and problems associated with the use of ANCOVA for estimating treatment effects in criminal justice evaluations. In Chapter II, we describe the two alternative methods for analyzing data of this kind and outline the general procedures used.

The following two chapters present the results for the respective data sets. A general discussion of the results and their implications is presented in Chapter V.

The Problem

Evaluations in the field of criminology often take as their primary focus the determination of whether a "treatment" of one kind or another results in behavioral changes among those exposed to it. Although there are a number of possible research designs that can be used for attempting to make such a determination (Campbell and Stanley, 1963), the most commonly-understood designs involve comparing the subsequent behavior of individuals exposed to the "treatment" in question to the behavior of individuals not so exposed. The basic logic of these designs is that the group of nontreated individuals can be used to establish what the behavior of the treated group would have been had they not been given the treatment. The difference between the actual behavior of the treatment group and their "expected" behavior, then, serves as an estimate of the treatment effect.

A central issue in research of this kind is the extent to which the behavior of the group of untreated individuals can legitimately be used to determine the expected behavior of the treatment group in the absence of treatment. The power of research designs for the evaluation of treatment programs is largely a function of how adequately this issue is resolved. The most powerful evaluation design is the "true experimental design" wherein individuals from a single pool of eligibles are assigned, on a random basis, either to the treatment program or to some alternative (preferably no treatment at all). Barring differential levels of attrition or non-response, the no-treatment group (referred to as a "control group") can be expected to be essentially equivalent to the treatment group prior to the treatment.

Assuming that the subsequent experiences of the two groups differ only with respect to the treatment under consideration, the behavior of the control group can, theoretically, serve as a perfect indicator of the expected behavior of the treatment group had they not been treated. Straightforward tests of differences in the subsequent behavior of the groups can be used to determine whether the groups differed in their behavior enough to warrant attributing any "treatment effect" to the program.

More common than these true experimental designs in criminal justice evaluation are "nonequivalent control group designs," in which subjects and controls are "matched" on important characteristics or in which outcome for subjects is compared to that of cases from an earlier (nontreatment) time-period, from a comparable area, or from a different setting (e.g., where juveniles diverted by the police are compared to cases referred to probation intake from other local jurisdictions). In these quasi-experimental designs, it can be assumed that the treatment and control groups differ on important variables which may affect behavior independent of treatment. Straightforward comparison of the subsequent behavior of the two groups cannot be used to estimate treatment effects, since the observed difference may be accounted for by preexisting differences between the groups.

In the history of criminal justice evaluation, the establishment of truly equivalent control groups is clearly the exception. In research on delinquency, for example, several recent large-scale evaluations were generally unable to implement random assignment procedures, even though such a method of generating control groups was attempted (Palmer, Bohnstedt and Lewis, 1978; Elliott, Ageton, Hunter and Knowles, 1976; Haapanen and Rudisill, 1980).[1]

[1] This problem appears to be particularly acute with respect to evaluations of such programs as youth service bureaus and diversion projects. Elliott, et al. discusses these problems in relation to their research on page 118 of their manual.

In a review of studies evaluating various forms of intervention with delinquents, Romig (1978) found 170 studies that met such basic methodological criteria as using a matched or randomly assigned control group and measuring outcomes using behavioral indices. Even among these 170 studies, true experimental designs were the exception. In a more general context, Palmer (1978), in his review of the massive survey, The Effectiveness of Correctional Treatment (Lipton, Martinson and Wilks, 1975) found that of 138 studies focusing on recidivism, less than half used random allocation of subjects. Even in studies where true experimental designs are implemented, random assignment procedures may not be carried out as planned, resulting in a selectivity bias similar to (although perhaps not as marked as) that obtained in nonequivalent control group designs.[2] Thus, while true experimental designs are not uncommon to criminal justice research, especially in institutional settings, a large proportion of studies fail to establish truly equivalent control groups.

It is in relation to the relatively weak, quasi-experimental, research designs and those in which experimental design procedures break down that the problem of nonequivalence between treatment and control groups is most apparent. However, the problem can also arise in the most rigorously applied "true-experimental design" situations. Just as the characteristics of any random sample of a larger population are likely to differ from those of that larger population to some degree simply by chance; randomly selected treatment and control groups can also be expected to differ somewhat from one

[2] As examples, in the studies by Haapanen and Rudisill (1980) and Jesness, Allison, McCormick, Wedge and Young (1975), random assignment procedures (implemented at one or more sites) were severely comprised by individuals responsible for the random selection. The control groups from these sites were subsequently treated as nonequivalent control groups.

another, both in terms of preexisting characteristics and in terms of subsequent behavior. This tendency to differ due to "sampling error" is especially true when the diversity of characteristics is wide and/or when the samples are fairly small. The larger the groups, of course, the more likely it is that they will both be reasonably representative of the larger population from which they were drawn and that observed differences in behavior will reflect differences due to the treatment rather than to sampling error. Accordingly, statistical tests used to assess the meaningfulness of these differences in behavior take sample size into account. However, it can happen that by chance, the groups are different enough in important ways that they would be expected to differ in terms of subsequent behavior, making a simple comparison of that behavior a poor indication of a treatment effect. In other words, sampling error alone can result in a situation analogous to having employed a nonequivalent control group design, although the problem is likely to be less severe.

A common method for analyzing data from both true experimental and quasi-experimental designs is analysis of covariance (ANCOVA), which is designed to adjust the observed outcome differences between the groups to take into account the effects of preexisting differences. Using ordinary least squares regression techniques, outcome scores are "predicted" from the characteristics, including treatment, which are felt to influence outcome. The result is an equation that describes outcome as a linear additive combination of the ("independent") variables in the analysis. Each has an unique (or "direct") effect on outcome--the effect of the variable after controlling for the effects of the other variables. The equation is of the following form, where Y is the outcome variable, the Xs are the independent variables, "b" refers to the direct effect of the variable and "a" is the

predicted value of Y when each of the independent variables has a value of zero:

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 \ldots + b_iX_i + e$$

The final term in the equation ("e") refers to the portion of the variation in outcome that is unrelated to the variables in the equation. In terms of prediction, it is "error."

The direct effect (unstandardized regression coefficient) of each variable in the equation refers to the expected, average change in the outcome variable for each change of one unit in the variable having the effect. When a variable coded "1" to indicate treatment and "0" to indicate control group is included in the equation, the coefficient for this variable becomes the estimated treatment effect--the predicted difference between the average outcome levels for the two groups after controlling for the other variables. It may refer to differences in average numbers of arrests, in average rates of reoffending, or in proportions of the groups who recidivate (if a dichotomous variable referring to success/recidivism is used as the outcome variable).

In its basic form, ANCOVA estimates the direct effects as if they are the same at all levels of the independent variable; for example, the difference in outcome between ages thirteen and fourteen is estimated to be the same as between ages seventeen and eighteen. The relationship between the dependent and independent variables, in other words, is estimated as a "linear function." It is in this sense that ANCOVA is said to assume a "linear-additive model" of the dependent variable.

As mentioned above, the direct effects of these variables, being average effects, cannot be used to perfectly account for all the variation in outcome scores, necessitating an error term in the equation. These errors are calculated in ANCOVA by using the developed equation to "predict" values of the criterion variable for all individuals, the difference between the predicted scores and the actual scores being the errors. ANCOVA is designed to minimize these errors, and it is in relation to that goal that the statistical significance of the equation is judged. Basically, statistical significance is based upon the likelihood that a given level of prediction accuracy could occur simply on the basis of chance correlations found in a random sample of a larger population in which there was no predictive power in the variables at all. The level of accuracy required to reach statistical significance is dependent on the size of the sample and the number of variables used. Similar tests of statistical significance are provided for each of the independent variables in the equation. In this case they refer to the significance of the increase in predictive accuracy achieved by adding that variable to those already in the equation, judged once again in relation to that which might be expected simply on the basis of sampling error.

The test of statistical significance of the treatment effect coefficient is generally used as a guide for interpreting the meaningfulness of the predicted difference in outcome between the groups. It is an estimate of whether such a predicted difference would be likely to occur by chance if the groups were equivalent--i.e., simply random samples of a larger population--and there was no treatment effect. Technically, such tests of statistical significance should not be used in this situation, since in the nonequivalent control group design the groups are, by definition, not random samples of a single population. However, in practice, it is often assumed that the group selection process is more-or-less random except for certain known differences, measures of which are included in the analysis. Thus, it is assumed that, by controlling for these differences (the bases for selection

into groups), the groups can be loosely regarded as random samples differing, then, only with respect to the treatment. Still, since this assumption is seldom completely valid, considerable caution is called for in using these tests in interpreting the results.

As suggested above, underlying the use of ANCOVA in nonequivalent control group situations are a number of assumptions regarding the variables, how they are measured, and the relationships between them. These assumptions, along with others, are basic to all ordinary least squares regression analysis. Violations of these assumptions can lead to biases in the estimation of the effects of the independent variables, including treatment, or to unreliability of the tests of statistical significance of these effects.

Four main assumptions are of interest here. At the most general level is the assumption that the variables in the analysis, taken together, constitute the important differences between the groups both before and after treatment. Omission of important variables or inclusion of irrelevant ones can cause the effects of other variables to be estimated incorrectly. Next is the assumption that the measures employed accurately assess individual and group differences in relation to the variables of interest. Inaccuracies can result from using invalid or unreliable measurement devices or from measuring at a single point in time variables that normally fluctuate over time (as, for example, mood or weight). Such inaccuracies will be discussed together under the rubric of "measurement error." Third, the tests of statistical significance of the equation and of the individual effects assume that errors of prediction have certain properties: that they have an equal variance at all values of the independent variables (homoscedasticity), that they are normally distributed, that they average to zero at each point, and that they are not correlated with other terms in the equation. Finally, ANCOVA assumes that the relationships between the dependent and independent variables are correctly specified in the equation. As mentioned previously, the estimates derived by ANCOVA presuppose a linear, additive model of the dependent variable; although only simple modifications are necessary for adding nonlinear and interaction effects, these may be difficult to discover and/or test for importance. Failure to include appropriate nonlinear or nonadditive effects or the inclusion of irrelevant ones, however, can lead to faulty estimates of the direct effects of the other variables, including treatment.

Violations of these assumptions underlying the use of ordinary least squares regression are common to criminal justice evaluation studies, whether they employ true experimental designs or quasi-experimental designs. They are, however, more serious with respect to the latter, since the adjustments sought by using ANCOVA are generally more crucial for arriving at a reasonable estimate of the treatment effect. In the pages that follow, we will discuss the problems that arise in relation to these assumptions and how these problems can effect the estimate of treatment effects using ANCOVA.

Omitted variables. In order to completely control for selection bias between the groups, the researcher must enter as covariates measures of either the determinants of group selection (e.g., when selection is based on pretest scores) or all causally relevant variables (i.e., all causes of the outcome criterion other than treatment). If either of these conditions are met, and if the usual assumptions of ordinary least squares (OLS) regression are met as well, the procedure will partial out the effects of group differences and provide the researcher with unbiased estimates of the treatment effects (Reichardt, 1979; Kenny, 1979; Overall and Woodward, 1977). Unfortunately, selection of individuals for intervention is rarely done solely on the basis of test scores or other directly measurable characteristics, and researchers concerned with human behavior never have at their disposal measures of all causally relevant variables.

If a variable that is related both to treatment and to outcome (independent of other variables in the analysis) is omitted from the equation, the estimated relationship between treatment and outcome will be biased, even after controlling for the other variables. To illustrate, if more males are assigned to a treatment program, the omission of sex from the ANCOVA analysis will cause the treatment group variable to act, to some degree, as a measure of sex as well as of treatment. If sex is related to recidivism, the treatment group may appear to have higher (or lower) recidivism even if there is no direct effect of treatment on outcome, since this important variable was not taken into account in estimating the treatment effect. It is possible that sex differences in outcome can be accounted for in the analyses by other variables upon which males and females differ. In this case, sex would be included indirectly, rather than directly. Its inclusion in one way or another, however, is necessary in order to properly adjust the treatment effect estimate to take differences in representation of the sexes into account. Thus, ANCOVA with nonequivalent control groups assumes that all important variables are included in the analysis.

One can never be completely sure, of course, whether this assumption is met, although the problem is minimized with true experimental designs; random assignment to treatment ensures that all important characteristics will, within the confines of sampling error, be equally represented in both groups. The omission of particular variables, then, should not be particularly serious, especially if the samples are fairly large.

In the nonequivalent control group situation, no such tendency toward group equivalence can be expected to operate, even with large samples, and it is important that this issue be carefully considered both in the design and analysis phases of the study. To include all important variables in the analysis, the researcher could include measures of all "causes" of outcome besides treatment (so that their effects can be partialled out of the treatment/outcome relationship). Due to the absence of strong theories of crime and delinquency based upon quantifiable variables, it is extremely unlikely that a researcher will be able to include all causally-relevant variables in the analysis. Alternatively, the researcher could include a direct measure of the basis for selection. For example, if assignment to a particular treatment program were based entirely on some measurable characteristic, such as test scores, one could assume that this measure would serve as a measure of other important group differences as well and that selection for treatment is otherwise random. By controlling for this known difference, the researcher theoretically controls for all important differences due to selection, meeting this assumption of ANCOVA (Campbell and Stanley, 1963; Cook and Campbell, 1979). However, this kind of controlled selection is rare. Nonrandom assignment of individuals to programs offering assistance of various kinds is usually done on the basis of volunteering by the subjects or of subjective assessments of the needs of the individuals, rather than on the basis of scores on some quantified measurement. Such procedures serve virtually to ensure that any comparison group will differ from the treatment group on important dimensions and that only indirect and incomplete measures of the important preexisting differences or basis for selection will be available.

Some bias in the estimate of the treatment effect, then, becomes almost inevitable in the nonequivalent control group situation. The researcher's obligation under these circumstances is a) to attempt to minimize this bias by including measures of as many important differences as possible; b) to understand the limitations of the study and the analysis by identifying possible omissions in the equation; and c) to exercise due caution in interpreting and drawing conclusions from the results. It is not enough, in this regard, simply to investigate possible bases for unreasonable or unexpected findings, since what is expected may, in fact, not be true. The finding of a positive treatment effect, for example, may well be expected and even welcomed and yet merely be the result of having failed to include in the analysis some variable (such as prior delinquent offenses) which biases the treatment effect estimate in favor of clients. The researcher should consider any and all alternative bases for any results of the analysis. It is tempting to assume, however, that because one controls for some of the important differences between the groups, the obtained estimates are closer to their true value, when actually one has merely eliminated some possible alternative explanations for any observed relationship between treatment and outcome. The direction and extent of the bias due to other, omitted variables is still unknown and may seriously bias the estimated treatment effect. Indeed, controlling for only one of two important differences between groups that bias the estimated treatment effect in opposite directions may cause the estimated difference between groups to be farther from the true value than would have been obtained by controlling for neither.

Implied by the foregoing discussion is the value of causal modeling as an evaluation research tool. Causal models of outcome and the selection process set the stage for deciding what variables to measure and include in the analyses and for understanding the nature of the study's limitations. If certain variables known or hypothesized to be important for adjusting the treatment effect estimate are omitted from the analysis (because they cannot be, or simply are not, measured), a causal model should aid the researcher in not only pinpointing them but also in speculating about their possible effects on the treatment effect estimate. This kind of information, in turn, sets the context for interpreting the results that are obtained.

Measurement error. ANCOVA assumes that the variables used to adjust the estimate of the treatment effect are accurate measures of the important differences between the groups. If there are errors in these measurements, the treatment effect estimate generated by ANCOVA will be in error also. On the one hand, a measure may be "invalid" in the sense that it measures something other than the variable of interest. This is systematic measurement error. What is measured may be closely related to what is intended to be measured, however, and it makes sense to speak of degrees of invalidity. For example, if one were to measure prior delinquent behavior in terms of delinquency-related police contacts, one would technically be using an invalid measure, since what is being measured directly is the behavior of police officers, rather than the behavior of the individual. However, police contacts and delinquent behavior are undoubtedly related, giving the measure some degree of validity. To the extent that a particular variable of interest is not validly measured, the researcher is faced with a situation analogous to having omitted an important variable from the analysis.

On the other hand, a measure may be employed that does directly measure the variable of interest, but not very well, resulting in random measurement error. This kind of inaccuracy can result from unreliability or from

measuring at a single point in time a variable that normally fluctuates over time. Unreliable measures are those that are affected by things unrelated to the variable being measured. Responses to an attitude questionnaire, for example, may be affected by the individual's momentary mood or state of physical health and would be expected to differ to some degree if obtained again at another point in time. Still, these influences are generally assumed to be random, so that, in general, the measure will provide a valid indication of the variable in question. The effect of the unreliability is that it adds to the normally-expected variation in the characteristic a certain amount of simply random variation as well. A similar increase in random variation is obtained when a variable, such as mood or emotional attachment, is measured at a single point in time. For most people, these characteristics tend to fluctuate around some "normal" midpoint, and even if they could be measured with perfect reliability they would not be the same if measured at another point in time. Technically, the problem here is one of validity, since the variable of interest is more likely to be a person's overall, average level of mood or emotional attachment, while what is being measured is momentary mood or feeling of attachment. Since the errors of measurement tend to be random, however, the effect on the measure at the aggregate level is much the same as that of unreliability: random variation is added to the normal variation found in the population.

For ANCOVA, the importance of this kind of random measurement error is that it tends to reduce (or "attenuate") the correlations between the variables. ANCOVA, however, uses these correlations as if they reflect the true association between them, resulting in biased estimates of direct effects (see Reichardt, 1979; Cohen and Cohen, 1975). In general, random measurement error in the dependent, or outcome, variable will cause errors in estimating the statistical significance of the effects of the independent variables, while errors of this kind in an independent variable can cause errors in estimating both the effects themselves and their statistical significance.

Random measurement error in the dependent variable is, by definition, uncorrelated with anything, making it absolutely unpredictable. The greater the amount of this error, then, the lower the possible accuracy of the equation, regardless of the true association between outcome and the covariates. Since the statistical tests focus on this prediction accuracy, they may provide misleading estimates of statistical significances. The ability of ANCOVA to estimate the direct effects (unstandardized regression coefficients) of the covariates, however, is not affected.

Random measurement error in one of the independent variables (or "covariates") results in an underadjustment for the effects of that variable. Since the correlations used by ANCOVA underestimate the true associations between this variable and others, its effects cannot be completely taken into account in estimating the direct effects of the other independent variables, including treatment. As random error increases, the correlations become smaller and smaller, so that eventually the effect is similar to simply omitting the variable altogether. As with omitted variables, the effect on the treatment effect estimate can be either up or down, depending on the true relationships between this variable, treatment, and outcome. Suppose, for example, that the clients of a program have a higher recidivism rate than comparisons, but actually have a lower recidivism rate than would be expected considering that they had more prior offenses. With ANCOVA, a positive treatment effect should be obtained if the effect of prior offenses is adequately taken into account in estimating the treatment effect. Underadjustment for the effects of prior offenses could result in the clients still appearing to have a higher recidivism rate (if the underadjustment is large) or to be no different from comparisons (if the underadjustment is moderate). Again, if the true effect of the variable with error counterbalances the effects of other variables in the equation, failure to properly adjust for it can cause, in essence, an overadjustment for the other effects. Group differences would appear to have more of an effect on outcome differences than was actually the case.

The same logic applies to the case where there is random measurement error in several, or even all, of the independent variables. As random error increases, the result is to increasingly fail to take into account the effects of these variables (these "group differences"). The more important the variable, the greater the bias that would result from its omission from the equation; hence, differential amounts of random error among the independent variables will result in varying amounts of bias in the treatment effect estimate. Further, since any particular omission may bias the estimate in favor of clients or comparisons, the resultant treatment effect estimate may be biased in either direction. In short, although it is clear that random measurement error in the independent variables will cause the treatment effect estimate obtained with ANCOVA to be biased, the extent and direction of the bias may be difficult to predict.

The important variables upon which treatment and control groups are likely to differ are often difficult to measure, raising the possibility of both systematic and random measurement errors. Such variables as age, ethnicity and sex can be measured with relative accuracy, but such theoretically important variables as family relations, attitudes about the law, or problems in school are less amenable to direct, accurate measurement. Commonly, these kinds of variables are measured by means of questionnaires or interviews wherein respondents are forced to choose among a limited number of alternatives in describing their lives, their attitudes, and their social situations. Scales constructed from these kinds of responses can be expected to contain considerable random error. Further, as measures of important preexisting differences between the groups, such scales may also be invalid to varying degrees. They may focus on the wrong attitudes, beliefs, perceptions, substantive areas, etc., or they may be constructed from questions that capitalize on the respondent's tendencies to answer particular kinds of questions in particular ways, adding "method variance" to the true variance. Regardless of the origin, the existence of measurement error in the covariates will bias the estimate of the treatment effect obtained by ANCOVA. Measures employed in juvenile and criminal justice evaluations are prone to these kinds of errors.

Classical approaches to the problem of measurement error in the independent variables advocate correcting for the resultant attenuation of the correlations among the fallible measures and others in the analysis (Kenny, 1979; Cohen and Cohen, 1975). A major problem with this approach is the likely absence of information regarding the reliabilities of the covariates. If standard psychometric measures are used, this information may be readily available, but where ad hoc questionnaires are used, information regarding reliabilities must be estimated. With large samples, these estimates can be made through common psychometric procedures for some kinds of data. In many or most cases, however, no empirically-based estimates are possible and the researcher would have to use "guesses" (perhaps high and low ones for comparison). Still, in the absence of a better alternative, such an analytic

strategy which takes possible measurement error into account would be preferable to those that ignore the issue altogether. Widely varying and inconsistent results, which might be expected when different values for the reliabilities are used and when the actual treatment effect is probably small, would merely serve to underscore the problems associated with data of this kind.

An alternative approach to the problem of measurement error is to employ multiple measures, or "indicators," of the variables that cannot be measured without error (Blalock, 1974; Kenny 1979; Jöreskog and Sörbom, 1979; Carmines and McIver, 1981). In some cases, existing measures may simply be conceivable as measuring a single variable in different ways. These indicators may be combined to form single measures, using factor loadings or by adding them together. The resultant measure, although still fallible, would have the effect of reducing the number of fallible measures that have to be contended with. When fewer measures are used, the analytic strategy outlined above may be more feasible. It is important to carefully consider the implications of such a practice, however, since the combination of variables may reduce the validity of the analysis by imposing on the data a faulty conception of how certain variables interrelate. For example, two variables may both be positively related to outcome and to one another, but their individual effects (controlling for each other) may actually be opposite. By merely adding the measures together, their independent effects would be assumed to be simply additive, and the result may be an erroneous estimate of their joint effect. Since the effects of neither of the variables would be accurately taken into account, the treatment effect estimate could actually be biased. This procedure is best used only when there is a strong basis for believing that the measures are indicators of the same underlying variable or that their independent effects are actually additive.

The multiple indicator approach can also be applied in such a way that the measures themselves, rather than some a priori combination of them, are used in the analysis. A similar assumption is made concerning how the measures interrelate and jointly relate to the outcome measure, but the conceptual "model" and the estimation procedure are quite different than those embodied in ANCOVA. Such a procedure is employed in the present research, and a full discussion of it will be presented later in the report.

Limited, skewed dependent variables. The use of measures of subsequent criminal or delinquent behavior as outcome variables pose certain problems for the estimation and interpretation of treatment effects with ANCOVA. These measures generally have a limited range of possible values, constrained at the lower end to be zero or above, and have skewed distributions--the most common value ordinarily being zero. Such constraints on the values of the dependent variable in ANCOVA can lead to problems in estimating the statistical significance of obtained coefficients and this difficulty, in turn, can lead to problems in determining not only the importance of the treatment effect estimate obtained but also which variables to include in, or omit from, the equation. Important variables may be omitted or irrelevant variables may be included, leading to actual biases in the estimate of the treatment effect.

These kinds of variables present problems for statistical tests in ANCOVA due to their potential effects on the distribution of prediction errors. Statistical tests in ANCOVA rest on the assumption that errors of prediction are normally distributed, and that they have a zero mean and equal variances at each value of the covariates (the assumption of "homoscedasticity"). Only if the errors are the same at each value does it make sense to consider the statistical significance of a particular effect of a variable. If the errors are not normally distributed or if the amount of error differs at different values (heteroscedasticity) the overall estimate of the error variance is not a good basis for assessing the significance of the effect. Such deviations would suggest that the estimated significance of the variable is true for some values but not for others. The use of the statistical tests under these conditions may lead to faulty judgements concerning the effects of variables in the model, including treatment. With limited, skewed outcome measures, most errors of prediction will be calculated relative to the observed value of zero. For these cases the amount of error that can be obtained will be directly related to the absolute value of the predicted score: the farther the predicted score is from zero, the larger the error variance can be. Thus, the assumption of equal error variances at different values of the independent variables is likely to be violated. These errors will also not be normally distributed and will not have, in general, means of zero. When the outcome variable is dichotomized as success/failure, it can be shown algebraically that the error variance is a function of the value of the covariates (Goldberger, 1964).

The outcome variables used in criminal justice evaluation research can also lead to poorly distributed error terms simply because they are unlikely to be linearly related to the independent variables. Employing a linear, additive equation to the prediction of outcome scores assumes, at least on an intuitive level, that the difference between any two values of the dependent variable are the same (that is, that the difference between "no subsequents" and "one subsequent" is of the same nature as the difference between four subsequents and five). Given changes in the values of the independent variables can then be simply related to constant changes in the outcome variable. In actuality, these differences in outcome are not likely to be of equal importance, and the independent variables may be related to the outcome differently at different levels of outcome. If a linear additive model were assumed, the amount of error would again be related to the predicted level of outcome.

The violation of these technical assumptions of ANCOVA can lead to inconsistent estimates of the error variance and even of the direct effect estimates themselves. In the case of a dichotomous dependent variable (and to a lesser extent when numbers of offenses are used), heteroscedasticity can lead to biases in the estimate of both the error variance for the equation and the standard errors associated with the direct effect coefficients for the independent variables. The nature of the bias depends on the relationships among all the variables in the equation, and there are no general rules for determining its direction or size (Goodman, 1976). Since the standard errors provide a basis for assessing the statistical significance of the effects of the independent variables, the researcher may have difficulty interpreting the results of the analysis, including the confidence to be placed in a given treatment effect estimate. The treatment effect estimates, themselves, however, remain unbiased.

Poor estimation of statistical significances can lead, however, to mistakes in judgement concerning the importance of certain covariates for adjusting the treatment effect estimate. If the statistical tests used for making such judgements are based on underestimates of standard errors, the significance of certain variables may be overestimated, and the researcher may retain variables that should be omitted from the equation, leading to faulty estimates of treatment effects. Conversely, if the standard errors are biased upwards, a truly important variable may not appear so on the basis

of the statistical tests. Such variables may be mistakenly omitted, again causing biased treatment effect estimates.

Several conditions influencing the extent of these problems caused by heteroscedasticity serve to mitigate somewhat their potential effect on ANCOVA results in most applications (Goodman, 1976). First, the amount of heteroscedasticity in the case of binary dependent variables is related to the proportions of the sample in the two categories. The closer the proportions are to 50/50 (that is, half successes and half failures), the less serious the problem, and it is not until the proportion in one category or the other gets above about 80% that problems resulting from heteroscedasticity are very serious. A similar lack of extreme skewness would undoubtedly be of value in the case where numbers of subsequents are used as the outcome measure. Second, the amount of bias is related somewhat to the power of the equation to explain outcome differences. Other things being equal, the better the equation is at predicting the outcome scores (the higher the R^2), the more serious are the problems attributable to heteroscedastic error terms. However, due to the limited variation in the outcome scores, there are probable limits, at least practically, to the degree of prediction accuracy to be expected from these equations. In criminal justice applications in particular, the general weakness of past attempts to predict criminal behavior or recidivism suggest that large R^2 values are unlikely. Finally, the bias in estimating standard errors resulting from heteroscedasticity have been found to be slight when sample sizes are fairly large ($n \geqslant 100$). Thus, the problems, which are somewhat minimized by the likelihood of low R^2 values in criminal justice evaluations, can be further minimized by using fairly large samples and keeping the amount of skewness in the outcome variable as low as possible (Goodman, 1976).

To summarize, the use of outcome variables common to delinquency and criminal justice research can lead to problems in estimating the statistical significance of the direct effects of treatment or of the covariates used to adjust for initial differences between groups. Where these statistical tests are used in interpreting the meaningfulness of the treatment effect estimate, the researcher may mistakenly attribute (or fail to attribute) a treatment effect to the program. Where they are used to determine whether or not a covariate is important enough to warrant its being maintained in the equation to adjust the treatment effect estimate, an important variable may be omitted from the equation or an irrelevant one included. In these instances the treatment effect estimate may actually be biased. Strict attention to the results of the statistical tests under these conditions, then, is unwise. At best they should be used as general guides to the importance of covariates and/or the confidence to be placed in the treatment effect estimate. By and large, however, the problems will probably not be very serious in most criminal justice applications, if attention is paid to sample size and skewness of the outcome variable.

When highly skewed outcome variables (other than dichotomies) are employed, it is possible to reduce that skewness to some degree by transforming the variable. Various kinds of transformations are possible, perhaps the most common being the log-transformation (Cohen and Cohen, 1975). The procedure simply involves using the logarithm of the variable in place of the raw scores in the ANCOVA analysis. The result is a change in the distributions of the skewed variables such that the difference in scores becomes greater for smaller values than for larger ones: the difference between the log of 1 and log of 2 is the same as the difference between the log of 5 and the log of 10. The skewness of the distribution is decreased

because the difference in transformed value for each unit change decreases as the values of the outcome variable increase.

One effect of using this transformation is that certain nonlinear relationships are linearized. As discussed earlier, the importance of having an additional subsequent arrest is probably much greater when the number is small than when the number is large, both conceptually and in terms of the relationship between criminality and other variables in the equation. Such nonlinear relationships with the covariates may be more linear when the dependent variable is log-transformed, resulting in fewer errors of prediction and a more homoscedastic error distribution. In addition, such a transformation has the general effect of reducing the influence of skewness on the variance of the outcome measure. It makes the variances more nearly equal at different values of the outcome variable and thereby facilitates the interpretation of differences between the groups. The heteroscedasticity assumption of the statistical tests is violated to a lesser extent, making it possible to place more confidence in that test with respect to the treatment effect estimate.

Problems of functional form. The simple linear, additive ANCOVA model may not adequately describe the relationships between the outcome variable and the covariates included in the analysis even when the outcome variable has been transformed. The effects of certain variables may be nonlinear in the sense that increasing values of the covariate may be associated with increasing (or decreasing) amounts of change in the dependent variable. The effects of other covariates may not simply be additive in the sense that the size of their effects depend on one another's values. These mutual dependencies are called "interactions": the two variables interact to produce an effect over and above the direct effects of each of the variables individually. If these relationships exist, their inclusion in the analysis is important for correctly estimating the effects of these variables on the outcome and, consequently, for estimating the treatment effect. ANCOVA is quite flexible, allowing for the inclusion of both nonlinear and interaction effects, but these effects may be difficult to detect, especially when measurement error and heteroscedastic error terms reduce the confidence in ANCOVA's statistical significance tests. Thus, although these functional form misspecifications can cause biases in the estimate of the treatment effect, their inclusion into the equation may be problematic.

The inclusion of nonlinear effects in the ANCOVA equation is fairly straightforward. Over the range of zero to eighteen, for example, the effect of age on delinquency may well be described best as a curvilinear one, with each year difference in age between, say, ages ten and eighteen being associated with steadily increasing amounts of delinquency change (the difference in delinquency between fourteen years old and fifteen years old would be greater than that between thirteen and fourteen years old). Such a relationship can be incorporated in the equation by including another variable in the model--in this case, age squared--in addition to the simple linear effect of age. With the effects of age and age squared both being positive, the overall effect of age would be described as involving a general increase in delinquency as age increases, with the increase being greater as age goes up. If the range of ages included in the model were increased another ten years, the relationship might appear different again, with each year after age eighteen associated with increasingly less amounts of delinquency (or criminality) change. In this case, a different kind of curved line must be fit, including, in addition to a general effect of age and an upward curve over the lower part of the range, a downward curve occurring

around age eighteen. Again, what is needed is the addition of another term, this time involving age cubed (age^3).

In general, any form of curve can be fit by including the appropriate power of the covariate plus all lower powers. The power function needed is one greater than the number of "bends" in the data (for one bend, a quadratic; for two bends, a cubic; etc.). As illustrated above, the nature of the nonlinear effect appropriate for describing a given relationship depends on the range of the variables involved. Even if a cubic relationship is called for over a wide range of ages, a simple linear effect may serve just as well if the range is over only a few years.

As mentioned previously, the problem of nonlinear relationships with respect to continuous dependent variables can be minimized in some instances through using the logarithm of the dependent variable rather than the variable itself. For dichotomous variables, an analogous change in the functional form may be made by fitting the equation to an S-shaped (logistic) curve using a logit model. When attempting to predict the probability of being in the "0" category (e.g., no subsequent arrests) or the "1" category (one or more subsequents) with ANCOVA, the distribution of predicted scores may contain values less than zero and greater than one. The logit model eliminates this possibility and also minimizes the problem of heteroscasdicity. The standard errors are more accurate, increasing the confidence that can be placed in the statistical significance tests. This method may also serve to better account for nonlinearity in the relationships between outcome and the independent variables. More will be said about this technique in the next chapter.

The inclusion of interaction effects in the ANCOVA equation is similarly straightforward. Instead of multiplying a variable by itself, however, to obtain quadratic or cubic forms, variables are multiplied by one another so that their effects are allowed to differ in direct relationship to each other. When added to the equation, the effect of this new variable ("interaction term") indicates the amount of change in the outcome variable associated with each variable as the other increases (or, if the effect is negative, as the other decreases). As an example, the effect of school problems on delinquency may be greater for youths with more family problems as well. By multiplying the measure of school problems by the measure of family problems and adding this interaction term to the model along with the two problem measures themselves, one allows for differences in the effects of these variables as the other changes. These interaction effects are often referred to as "multiplicative effects," to distinguish them from the simple "additive effects" of the two variables individually.

Interactions can involve dichotomous variables (such as treatment/control) as well. Not only may treatment and control groups differ from one another on important variables influencing outcome, but the actual effect of one or more of these variables on the outcome may differ between clients and controls. The effect of prior delinquency, for instance, may be different between the groups, indicating that increasing numbers of priors associated with more (or less) delinquency for clients than for comparisons. Since the converse is also indicated, such an interaction effect would suggest that the treatment served to mitigate (or exascerbate) the influence of prior delinquent involvement on subsequent delinquency. Including this kind of possible interaction effect in the equation simply involves adding a multiplicative term involving the treatment/control dichotomy and the measure of prior delinquency. The coefficient for this interaction term indicates the extent to which the effect of priors on outcome differs

between the treatment and control groups. However, because these effects are allowed to differ, it no longer makes sense to think of a single treatment effect: it differs depending on the amount of prior delinquency. Thus, the direct effect of treatment no longer refers to the estimated average difference between the groups but rather the estimated difference between the groups when the number of priors is zero. Obviously, this coefficient would be of little use as an indicator of an overall treatment effect. Although such a finding would still be meaningful, the interpretation of the treatment effect would be much different since it would be directly tied to levels of prior delinquency.

Whether or not one should include nonlinear or interaction effects in the ANCOVA model depends upon their importance for explaining the joint relationships of the covariates (including treatment) and outcome. The inclusion of irrelevant nonlinear or interaction effects may have the same effect as the inclusion of other irrelevant variables in the analysis: unnecessary adjustments may be made to the treatment effect, resulting in a biased estimate of that effect. The inclusion of irrelevant interaction terms involving treatment would further complicate the results since no single treatment effect estimate would be available from the analysis. Typically, the determination of the relevance of such effects rests on whether their inclusion results in a statistically significant reduction in the errors of prediction. The nonlinear or interaction terms are added to the equation after all of the simple, additive effects are included and those that add significantly to the predictive power of the equation are retained in the model.

In most criminal justice evaluation applications, however, reliance on the tests of statistical significance in ANCOVA poses certain problems, as we have seen. Measurement error in the covariates and the dependent variable, along with poorly distributed error terms possibly resulting from using limited, skewed outcome measures seriously reduce the confidence that can be placed in these tests. Strong theories concerning the relationships among variables in the model and thorough examination of how the measures used in the analysis interrelate can be used to guide the researcher in making these determinations. But without being able to rely on empirical evidence concerning the importance of these effects in a particular context, there remains the possibility of including irrelevant effects or of mistakenly omitting important ones. Either way, there is the potential for obtaining biased estimates of the treatment effect.

CHAPTER II

Alternatives to ANCOVA: General Description

Two different analytic strategies, each designed to overcome certain problems associated with ANCOVA, were applied to data sets from earlier evaluation studies involving delinquents.[1] The results of these analyses were compared to results obtained with ANCOVA in its basic form in an attempt to determine the extent to which the ANCOVA results are affected by these problems. The first analytic strategy involved an application of the causal modeling approach with multiple indicators of unmeasured variables, pioneered by Jöreskog and Sörbom (Jöreskog, 1973; Sörbom, 1978; Jöreskog and Sörbom, 1979). With the aid of the LISREL computer program (Jöreskog and Sörbom, 1981) the research "tests" alternative causal models hypothesized to explain variation in outcome scores. The nature and importance of treatment effects are assessed by comparing the results obtained when a treatment effect is included in the model to those obtained when no treatment effect is included. Variables hypothesized to account for selection differences between treatment and control groups can be included individually or as indicators of underlying "unmeasured" variables, which are related to selection and to outcome. The second strategy involved employing a number of analytic tools in combination in order to overcome the various problems found with ANCOVA. Although the strategy does not involve the "testing" of causal models per se, we used it to analyze models similar to those used in the LISREL analyses. Measurement error in the independent variables was handled through combining the fallible measures into factors on the basis of factor

[1] The data sets and the prior analyses will be described in chapters III and IV.

analysis. The determination of the appropriate functional form of the model was made through a loglinear analysis, which does not require any distributional assumptions regarding the data. Finally, the problems associated with the limited, skewed dependent variables were approached through using tobit models for nondichotomous variables and logit models for dichotomous outcome variables. With this combination of methods, we hoped to minimize the effect of each of the problems associated with the application of ANCOVA to these kinds of data.

Both analytic strategies are obviously more complex than ANCOVA. Although each should theoretically provide better estimates of treatment effects than found with ANCOVA, the earlier discussion should have made it clear that the extent of improvement to be expected from employing these more complex methods is not clear. Ordinary least squares estimation has been repeatedly shown to be quite "robust" with respect to violations of the assumptions underlying it. Even if technically misused, ANCOVA may provide treatment effect estimates that are adequate--for all practical purposes at least. In other words, even if the treatment effect estimate obtained by ANCOVA is biased, it may be close enough to allow for the determination of any substantively important effects of treatment. Similarly, the inefficiency of the statistical tests may not be serious enough that their use as guides would provide badly misleading interpretations of findings or of the importance of nonlinear or multiplicative effects in the equation. Comparisons of the results obtained through the various methods, then, should provide insights into the value of employing more complex and less easily interpretable analytic strategies in place of ANCOVA.

LISREL Approach

General Description

The LISREL analytic approach involves the estimation and testing of various "causal models" hypothesized to explain differences in outcome. The estimation and testing are performed with the aid of a particular computer program, LISREL V (Jöreskog and Sörbom, 1981) designed for use with these kinds of models. We will not attempt a full explication of the LISREL approach to data analysis here; rather, we will merely provide an intuitive description of the method as applied to the present research. The discussion is not intended, therefore, to provide the reader with an outline of how to use the LISREL program.[2] In describing the approach, we will avoid the use of equations as much as possible, but will present diagrams of models (called "path diagrams") and use the system of notation for elements of the models that has come to be associated with this class of analytic methods. We will first describe the main parts of the LISREL model. We will then describe the particular model used in the present research.

The LISREL method is quite flexible, allowing for the estimation and testing of a number of different kinds of causal models. The researcher starts with an idea concerning how the measures (observed variables) in the analysis are related to one another. These relationships are hypothesized to account for whatever correlations are observed among the measures. For example, high positive correlations between certain measures may be

[2] In recent years, several well-written descriptions of the method and how to use it have appeared in the research literature. The interested reader is directed to Long (1976), Kenny (1979), Murayama and Garvey (1980), and Rindskopf (1981). The most complete discussions are provided by the pioneers of the method and the authors of the computer program (Jöreskog and Sörbom, 1979, 1981--accompanying manual).

hypothetically explained in terms of their being joint indicators of a single underlying, unmeasured variable. Similarly, a researcher may hypothesize, based on theory, that whatever correlation exists between two particular variables can be accounted for by one variable's being the "cause" of the other (rather than the other way around or both being reciprocal causes of one another). The sum of these hypothesized relationships concerning how each variable in the analysis is measured (the "measurement model") and how the variables are causally related (the "structural model") comprise the overall LISREL model. This model, again, is hypothesized to account for the relationships among the observed variables (the input data), and it is in relation to its ability to account for these relationships that the model is "tested."

An important condition for the estimation and testing of the model is that the model be "identified." In general, a model is identified when the unknown parameters of the model are capable of being uniquely estimated from the information provided by the relationships among the observed variables in the analysis. We will not go into detail concerning the determination of the identification status of models[3] except to point out that the LISREL program will in most instances "warn" the user of possibly under-identified parameters. A parameter is under-identified when there is not enough information available from the data to obtain a unique estimate of that parameter. In particular instances, under-identification can be resolved by constraining certain of the parameters of the model (by "fixing" them at particular values, or constraining them to be equal to other parameters in the model). This reduces the number of parameters that must be estimated from the data.

[3] See references found in footnote 2.

When there is just enough information in the data to identify all of the parameters, the model is said to be "just-identified," and when more than enough information is available, the model is "over-identified." In just-identified models, the observed relationships among the measures allow for a single estimate of each of the unknown parameters in the model, and the model will therefore "fit" the data perfectly. This does not mean that the model is necessarily true, but rather that the estimates that are obtained will necessarily be consistent with the data. In over-identified models, on the other hand, there is more than enough information available, and some of the parameters of the model can be estimated in more than one way from the data at hand. It is under these conditions that it becomes possible to "test" the model.

Through a "maximum likelihood" estimation procedure, the LISREL program arrives at a "best" estimate of the over-identified parameters. It then uses all of the estimated parameters of the model to recreate the information matrix used to calculate them in the first place. A model is judged to "fit" the data from which it was estimated to the extent that the correlations (or covariances) among the observed variables that are implied by the estimates are close to those actually used to obtain them. Since the estimates of the over-identified parameters will enter into the recalculation of a number of original relationships, it is highly unlikely that there will be a perfect fit of the over-identified model to the data; the greater the extent of over-identification, the greater the expected discrepancy between the model and the data, even if the model is relatively good. However, the greater the consistency among the possible estimates of the over-identified parameters, the more likely that the single best estimate of them will enable the program to closely recalculate the original information matrix.

As an indication of how well a particular model fits the data, the program calculates a chi square (χ^2) goodness-of-fit statistic, based upon the discrepancy between the original information matrix and the matrix calculated from the parameter estimates. This statistic, with degrees of freedom equal to the number of over-identifying constraints in the model, is used to estimate the probability of obtaining such a discrepancy on the basis of chance. A small value of χ^2 relative to the degrees of freedom, indicates that the model fits the data, and is associated with a high probability level. Using predetermined probability levels as with other null hypothesis statistical tests (.05 or .10), the researcher accepts that the model adequately fits the data when the probability level is exceeded, indicating that the two matrices are reasonably similar. Of course, this test rests on certain distributional assumptions concerning the variables in the analysis,[4] just as do those used in ANCOVA. In fact, these assumptions are more restrictive than those underlying the ANCOVA statistical tests: the observed variables themselves are assumed to have multivariate normal distributions, whereas with ANCOVA only the errors are assumed to be normally distributed. Further, χ^2 is sensitive to sample size, such that larger samples lead to larger χ^2 values over and above what can be expected due to errors in specifying the model. Thus, a good fit to the data with any particular model and set of data may be difficult to determine using this method, and as with most statistical tests, then, tests based on this statistic should be interpreted somewhat cautiously.

[4] "...the χ^2 is a valid test statistic only if:
(1) all the observed variables have a multivariate normal distribution,
(2) the analysis is based on the sample covariance matrix $\underset{\sim}{S}$ (standardization is not permitted),
(3) the sample size is fairly large." (Jöreskog and Sörbom, 1981, pg. I.39).

However, these restrictions on the usefulness of the χ^2 test apply most strongly to "tests" of the adequacy of the overall model for describing the relationship in the population from which the sample and data were drawn. It is possible to use the statistic in a more limited and justifiable way to assess the importance of particular parameters. This use involves comparing the χ^2 values obtained when these parameters are constrained and when they are freely estimated. The model with the constrained values will usually result in a larger χ^2 value and will also have more degrees of freedom. A large drop in χ^2 relative to the difference in degrees of freedom when these parameters are freely estimated indicates that the model is improved in the process (i.e., that the freely-estimated parameters have some "significance" in the model). The probability of obtaining differences in χ^2 relative to differences in degree of freedom is similar to that of obtaining particular values of χ^2 with particular degrees of freedom, and standard χ^2 probability tables can be used for determining these probabilities. Through this hierarchical testing procedure, the model can be progressively improved and the importance of particular parameters (such as treatment effect estimates) can be assessed. The emphasis here is not to achieve some particular level of "fit" to the data, but rather to use the statistics to help arrive at a reasonable model of the causal process and obtain reliable and useful estimates of the parameters of interest.

In evaluation applications, LISREL, like ANCOVA, is used to estimate the predicted difference in outcome between the treatment and control groups, taking into account the effects of other variables. As suggested above, however, analyses using LISREL can go beyond ANCOVA in that the researcher can more clearly specify the nature of the effects of other variables and can specify that certain of the variables are not measured directly but rather are indicated by a number of separate measures. For example, the researcher could specify that certain measures are all indicators of a single underlying, unmeasured variable ("latent factor") which has a causal effect on outcome and which is related to treatment. He could also specify that other variables have only indirect effects through causing differences only in the underlying factor.

Another major advantage with LISREL is that it allows the researcher to analyze the treatment and comparison groups separately, but simultaneously. That is, the data for each group is read in separately, and the program estimates the model for each group at the same time. The model can be constrained to be identical for each group or certain parameters can be allowed to differ between them. Through this procedure, the researcher can determine the extent to which the same causal model of outcome is applicable to both groups and can also more fully take into account differences between them (such as differing levels of within-group homogeneity relative to certain variables) in estimating treatment effects. When the mean levels of the variables for each group are included in the analysis ("structured means analysis"), the program will estimate the mean differences between the groups for all variables; these differences will correspond to actually observed differences between the groups on the predictor (or independent) variables and to adjusted differences for dependent variables. As long as the direct effects of the variables in the analysis on outcome are constrained to be equal across groups, the program will estimate the predicted mean difference in outcome (the treatment effect) between the groups--this difference corresponds to the treatment effect estimate obtained with ANCOVA. The researcher can also test whether the direct effects of variables in question "interact" with treatment to cause differences in outcome. Under these

conditions, again, no single treatment effect estimate is possible, but a better understanding of the effects of the variables, including treatment, on outcome may be obtained.

In terms of the specific problems discussed earlier, the LISREL approach has some definite advantages over ANCOVA. First, although this approach cannot directly address the problem of the effects of omitted variables, it does require the user to specify the causal assumptions underlying his or her analysis. The treatment effect estimates obtained through this method are derived in the context of a particular hypothesized causal model, which is ordinarily presented in the form of a diagram. Such an explicit presentation allows for a direct examination of these assumptions both by the researcher and others and sets the stage for a better understanding of the meaning and the limitations of the analysis. Second, LISREL allows for the use of multiple indicators of variables in the model; the researcher thereby can directly address the issue of measurement error, both systematic and random. Third, when the groups are analyzed separately, certain (first-order) interaction effects involving treatment can easily be included in the analysis and tested for statistical significance, making it possible to incorporate both measurement error and interaction effects in the same model. Together, these advantages should lead to better estimates of treatment effects using criminal justice data.

There are also problems which this approach cannot address, placing some limits on its ability to fully overcome the limitations of ANCOVA. First, LISREL cannot completely solve the problem of using limited, skewed dependent variables, although the variables can be transformed in the same way as is possible with ANCOVA. Such transformations "robustify" the observed relationships used in the analysis, but do not completely remove the possibility of bias in estimates of standard errors and "goodness of fit" measures (χ^2). In fact, it has been argued that LISREL is even more sensitive to departures from normality in the variables than is ANCOVA, and caution is again called for in the use of statistical tests. Second, LISREL cannot easily incorporate either nonlinear effects of particular variables on outcome or higher-order interaction effects.[5] As discussed earlier, the omission of important effects of these kinds may lead to biased estimates of treatment effects. On the other hand, their omission should also result in a less-than-adequate "fit" of the misspecified model to the data; thus, the researcher can at least be aware that his or her model is misspecified and interpret the findings accordingly.

The measurement model. The measurement model specifies the hypothetical relations between the variables to be used in the analyses and the observed data. On the one hand, the researcher may specify that certain variables are equivalent to certain observed variables, the assumption being that the variable is "perfectly measured." One may, for example, specify that an official offense history variable (number of arrests) is perfectly measured. This would not mean that the measure is considered a perfect measure of criminality or delinquency, but rather that in the particular model, the variable of interest is precisely what was measured. In the present study, this assumption of perfect measurement was made with respect to all official offense and demographic variables. On the other hand, certain variables may be considered to be "unmeasured" in any direct sense but "indicated" by several observed variables included in the analysis. Using the present study as an example, we had at our disposal several questionnaire scales focusing on self-reported behavior of various kinds. We considered these measures to be best thought of as multiple fallible indicators of a single, delinquency-related "Behavioral Orientation." In the measurement model, then, we specified, among other things, that "outcome" had a single indicator and that the relationship between the variable and the observed measure was one of equivalence. The causal variable "Behavioral Orientation" was specified as having three indicators, each of which had a shared component (related to the underlying variable) and an "error component." Each of the variables in the analysis is similarly specified to be indicated by one or more observed variables. Together, these specifications make up the overall measurement model.

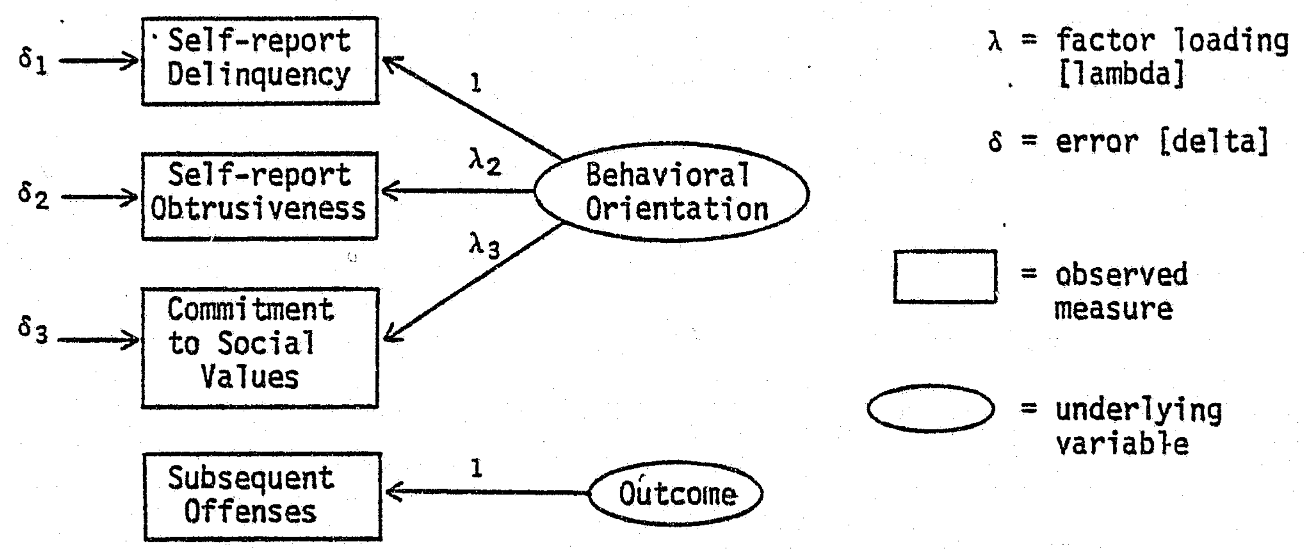
When there is more than one indicator, the measurement model for the variable is essentially a factor-analysis model, with the coefficients relating the variable to its indicators analogous to factor loadings. In this case, however, the factor structure is specified prior to the analysis, and the LISREL program merely estimates the loadings and "tests" the model in terms of its ability to account for the observed relationships among the observed variables. The entire measurement model, in fact, can be seen as a single, restricted factor analysis model, wherein the researcher specifies the number of factors involved and their structure (what measures indicate what variables and whether the coefficients are constrained or to be estimated by the program). In other words, the measurement model in LISREL may be seen as a "confirmatory" factor model; it is part of the overall model which is "tested" (or confirmed) in relation to how well it fits the observed data.

[5] A higher order interaction exists where the nature of a first-order interaction effect differs depending on the value of a third variable. For example, a first-order interaction may exist between prior delinquency and treatment (so that the effect of priors in outcome differs for treatment and control groups), but the nature of that effect may depend on whether the individual was male or female. This would be a second-order interaction (treatment by priors by sex).

Following common conventions, the measurement model for the two variables used as examples above can be diagrammed as in Figure 1.

FIGURE 1

Measurement Model for "Behavioral Orientation" and "Outcome"



Note that the arrows lead from the hypothesized variables to their respective observed indicators (measures), suggesting that the underlying variables are the "causes" of their indicators. Actually, what is meant is that the variation in the observed measures is defined as due, in part, to their being related to the underlying variable, so that variation in that variable will necessarily be accompanied by variation in the observed measure of it. The remainder of the variation in the observed measures is considered, in this context, to be simply due to "error" in measuring the underlying variable. For variables with only one indicator, of course, there is no error component, and the error term is omitted; the coefficient relating the variable to the measure is "fixed" at the value of one to make the variances equal. In the

remainder of this report, we will omit the observed measure in diagrams of these "perfectly indicated" variables and denote the variable by the name of the measure (e.g., SUBS).

Note that in the three-indicator model above, the path from the factor to Self-report Delinquency was similarly fixed at the value of "one." Without this constraint imposed, that part of the model would be underidentified in a single-sample analysis: There is not enough information available from the three correlations among the measures to estimate all of the possible parameters (the factor loadings, the error terms and the variance of the factor). The identification issue was resolved in this instance by fixing this path at "one," which merely sets the units of measurement of the Behavioral Orientation factor equal to those of Self-report Delinquency. Since the units of this factor are arbitrary anyway (one could just as easily set the variance of the factor equal to "one"), nothing is lost in the way of information, and the remaining parameters are identified. This identifying constraint would not be necessary in a two-sample analysis in which the factor loadings are constrained to be equal (the common procedure). In this case, there would be twice as many correlations among the measures available for estimation purposes and three constrained parameters. However, in order to allow for the inclusion of additional free parameters, as discussed shortly, this identifying constraint was used throughout this study. In all the models presented, one path from each underlying factor to a measure will be shown to have a fixed value of "one."

Assuming that all of the parameters are properly identified, the LISREL program estimates that nonfixed parameters and also assesses the fit of the model to the data. When two groups are analyzed simultaneously, as in the present research, the goodness-of-fit statistic relates to the model as a whole, including whatever differences or equalities are specified to exist across groups. For example, it is possible to test whether the measurement model for Behavioral Orientation is the same for treatment and control groups. This is done by specifying the same model for both groups, with the constraint that the factor loadings are invariant across groups (e.g., that λ_2 is the same for treatment and control groups). The error variances, the variance of the factor, and the mean level of the factor (using "structured means" analysis) would all be allowed to vary between groups. What would be tested here is simply whether the measures are related to the underlying variable in the same way in the different groups. Due to random influences, there is no reason to assume that the variances of the error terms or the variance of the factor itself should be equal across groups, although wide discrepancies may suggest that the groups differ more than one would expect in certain instances. One could also go on to test whether the means on this factor differ between groups by constraining the mean level to be equal and comparing the resultant χ^2 value to the one obtained when this parameter was free to differ. In short, one can test any degree of difference between the groups as well as test whether a particular model, given a certain degree of equality between groups, fits the data obtained from them together.

A poor fit at this level may indicate, among other things, that the model is not specified correctly. For example, it may be that the measures are not related in the same way between groups--that the measurement model does not apply equally between them. In extreme cases, it may be that the measures are actually correlated in opposite directions between the groups, suggesting that a measurement model which hypothesizes a similarly constituted

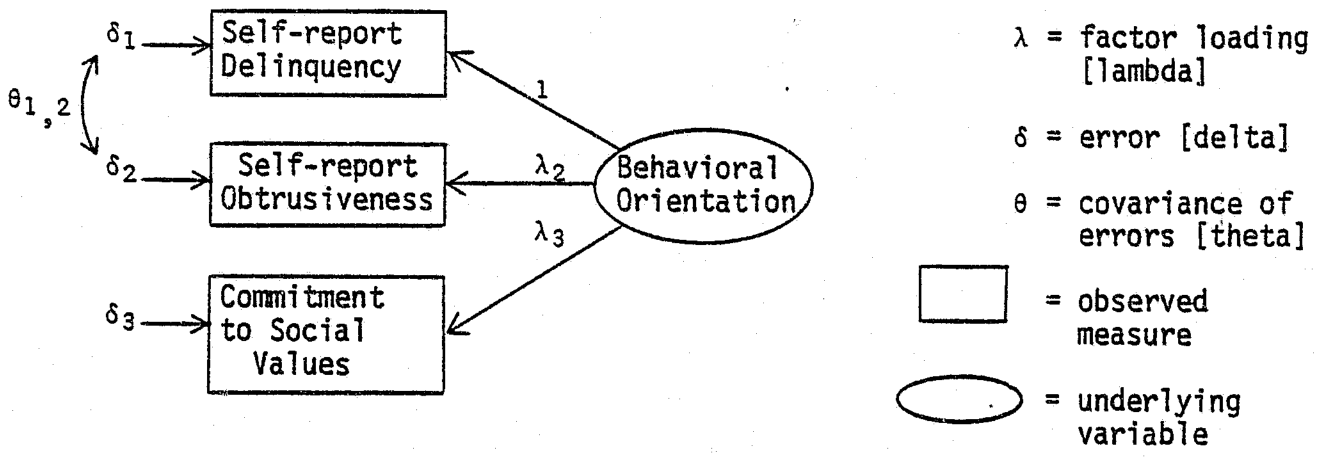
factor between the groups is fundamentally wrong. In such cases where the hypothesized model is basically inconsistent with the data, the poor fit will likely be accompanied by other indications of extreme misspecification. The program may simply be unable to arrive at a "best" estimate of the parameters after trying for a specified time or it may arrive at unreasonable values, such as negative error variances or extreme values of the factor loadings. In such instances, it is best to rethink the nature of the measurement model entirely.

A high χ^2 value, indicating a poor fit to the data, does not necessarily imply that the model is fundamentally wrong and should be abandoned, however. Oftentimes, simple modifications can bring the model more into line with the data. For example, it may be that certain measures are correlated over and above their being joint indicators of an underlying variable. In the above example, due to a similarity in focus and format of the questions included in the Self-report Delinquency and Self-report Obtrusiveness scales, these two measures shared a certain amount of "method variance" unrelated to the generalized Behavioral Orientation variable of interest in the study. The measurement model diagrammed in Figure 1, which implies that any relationship between the two occurs solely because of their being joint indicators of that underlying variable, would not allow for this extra relationship and would therefore not fit the actually observed data very well. This additional relationship between the measures can be "included" in the model by specifying that the "error variances" of these measures (in relation to Behavioral Orientation) are correlated,[6] as diagrammed below:

[6] Implied here is that there is an additional, underlying "method factor" involved, which causes variation in the two self-report measures but not in the commitment measure. Since we are not interested in this factor as part of our overall model, we can ignore it and simply include the extra correlation between the measures in the form of correlated "errors."

FIGURE 2

Measurement Model With Correlated Errors

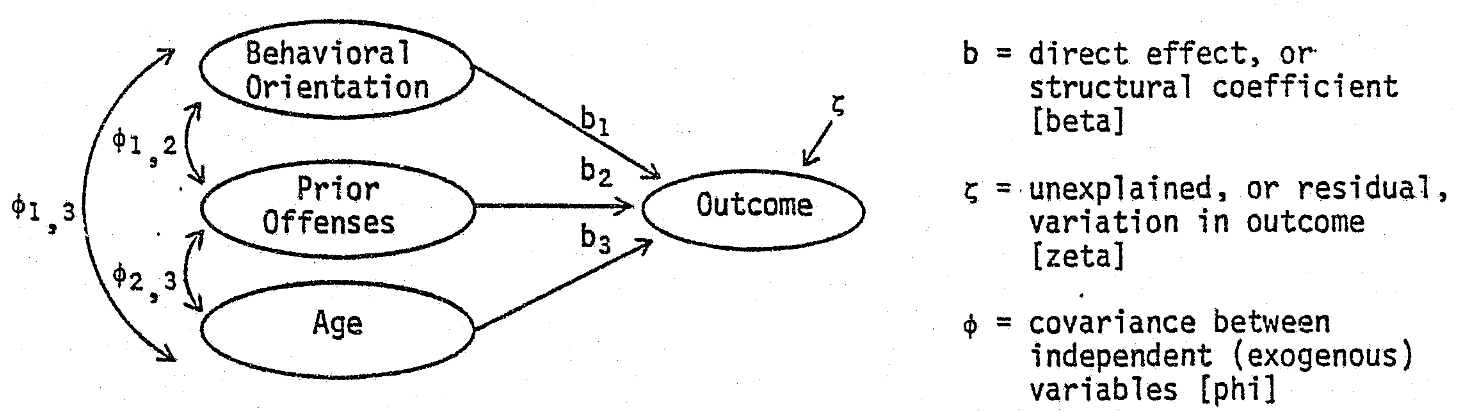


In making such adjustments to the model, the researcher must be aware of possible problems with the identification of the additional parameters. Fortunately, the LISREL program provides a mechanism for helping the researcher avoid adding parameters that are not identified. As part of the output from the program, LISREL provides "modification indices," which estimate the minimum change in χ^2 that can be expected by freeing particular constrained elements of the model. Only those parameters that would be identified, if freed, are provided with a modification index by the LISREL program. Thus, the user can study the modification indices to ensure that freeing a particular parameter would provide an improvement in the model's fit and would be identified. If the freeing of the parameter makes theoretical sense, it can be included. This manner of determining the identification status of the parameter is not foolproof, however, and the researcher should assess the identification status of all models used. In the above example, the additional parameter (the correlated error terms) would not, in a single sample analysis, be identified. It was identified in the present research because of the one fixed path and because we analyzed the two groups simultaneously.

The structural model. The structural model in LISREL describes the hypothesized interrelationships among the variables defined by the measurement model. Some of these interrelationships may be "causal," while others may merely be included as "unanalyzed correlations." Unanalyzed correlations imply that although the relationship between the variables is considered an important part of the overall model, the "reason" for that relationship is not of interest in the model. An example of a simple causal model is diagrammed below, with Behavioral Orientation, Prior Offenses, and Age all being simultaneous "causes" of outcome.

FIGURE 3

A Simple Causal Model of Outcome


Behavioral Orientation
Prior Offenses
Age
Outcome
b1
b2
b3
ζ
φ1,2
φ1,3
φ2,3
b = direct effect, or structural coefficient [beta]
ζ = unexplained, or residual, variation in outcome [zeta]
φ = covariance between independent (exogenous) variables [phi]

In this model, the three causal variables are hypothesized to be correlated with one another, but the reasons for these correlations are not of interest in the model. Multiple regression assumes this kind of a general causal structure among the variables in the analysis, and the model above could be estimated with a standard multiple regression program. If an additional, dichotomous variable indicating treatment/control were to be included as a fourth causal variable in this model, it would be an ANCOVA model. The coefficient for the dichotomous variable would indicate the difference between the groups not accounted for by the effects of the other variables.

LISREL could also be used to estimate this ANCOVA model, either by using the dichotomous variable as described above or by separately analyzing the treatment and control groups. When a two-groups analysis is used, as in the present research, the model diagrammed above (without the treatment/control dichotomy) would be specified for both groups, with the direct effects of the variables on outcome constrained to be equal across groups. The program would estimate the difference in the mean level of the outcome variable after adjusting (equally between groups) for the effects of the predictor variables. In order to obtain the same estimate of the treatment effect as would be obtained with ANCOVA, one would also constrain the variance of the error term (ζ) to be equal across groups, because with ANCOVA these error variances are assumed to be equal.

With these constraints, there is more than enough information available to the LISREL program to identify the unknown parameters in the model. In fact, there is enough information in the two covariance matrices to allow for the different direct effects of the covariates on outcome for the two groups (interactions) and even to allow for different error variances between groups. The ability to allow for different error variances is an important advantage of using LISREL, overcoming some of the restrictiveness of the homoscedasticity assumption in ANCOVA. With nonequivalent control group designs, one could well expect that the groups would differ with respect to variation in outcome scores (the treatment group members may, for example, be more similar to one another than the control group members), and such a difference in the variance of observed scores will likely result

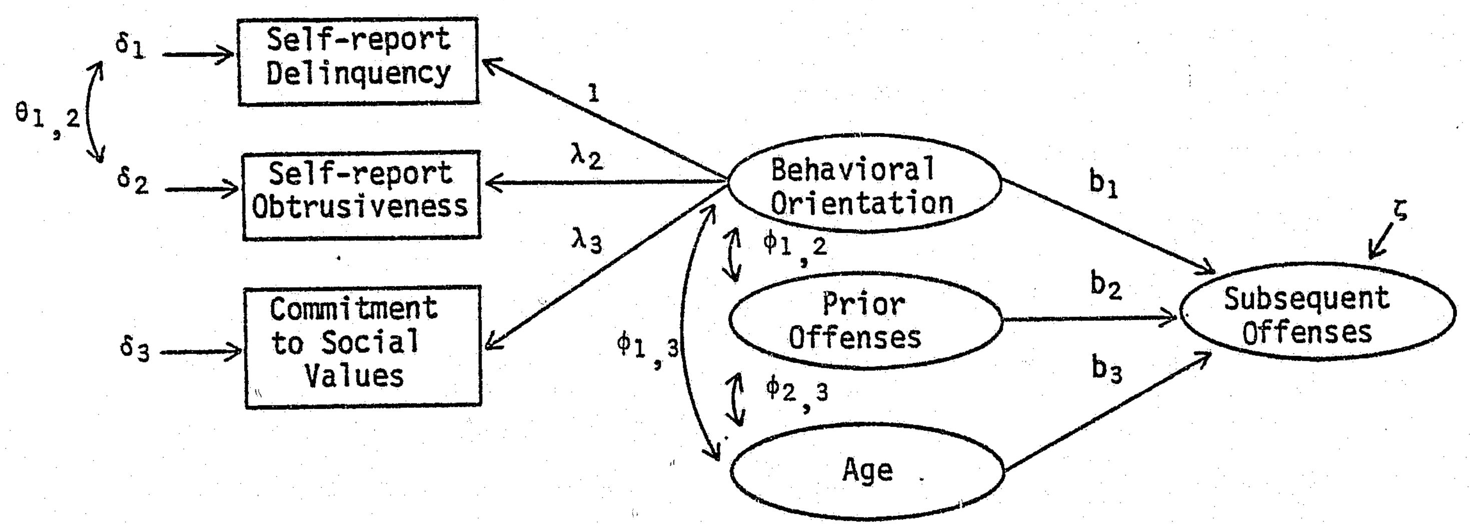
in a difference in the variance of errors in prediction as well. ANCOVA assumes that the error variances are equal across groups and assesses the statistical significance of the treatment effect in relation to a pooled estimate. When the error variances differ, this combined estimate is not a good basis for judging the significance of a particular difference. In LISREL, with the error variances allowed to vary between groups, the significance of the treatment effect is judged in relation to the error variance within the treatment group only, providing a better indication of the meaning of that difference in expected outcome. The assumption that errors are evenly distributed across all values of the other independent variables still holds, however.

Although conceptually distinct, the measurement model and the structural model are not independent. A factor hypothesized to underlie several measures is "defined" not only by those measures but also, to some extent, by its relations to other variables in the model (Burt, 1976). For example, the Behavioral Orientation factor discussed previously is something different when merely estimated in relation to its three indicators than when it is estimated as a variable with a causal influence on outcome. The factor loadings will differ in the two cases, since in estimating these loadings in the latter case the LISREL program takes into account not only the relationships among the three indicators, but also the relationship between each of the measures and outcome. The factor is, in the context of the model, treated as an error-free measure of a <u>particular</u> predispositional trait, indicated by the three measures. Similarly, this factor would be estimated somewhat differently if it were specified to be related (causally or otherwise) to other variables in the analysis. To specify that this factor were related to, say, prior offenses is to imply that the observed correlations between the measures used to indicate the factor and prior offenses can be accounted for solely by the relationship between the factor and priors. Thus, it is important to consider the implications of each specification on the entire model: a change in the structural model will result in changes (and may call for modifications) in the measurement model as well. Conversely, such an interdependence between the structural model and the measurement model also implies that the measurement model is best tested, and modified if necessary, in the context of the overall model.

The full model. When combined with the measurement model described earlier, the structural model above would be part of the full LISREL model diagrammed in Figure 4.

FIGURE 4

Full LISREL Model



When two groups are analyzed simultaneously, the measurement model is constrained to be invariant across groups (the same variables must be used and these must be defined the same way) and the effects of the three variables

on outcome are constrained to be equal as well, at least initially. If the results suggest that these effects may differ between groups, such an hypothesis may be tested by allowing them to vary and assessing whether the relaxation of these constraints results in a significant improvement in the model's fit with the data. Again, if these interaction effects are found to be important, the estimated group difference can no longer be used as an estimate of the treatment effect, which will vary depending on the value of the covariate that interacts with treatment.

The full LISREL model used in the evaluation applications is quite similar conceptually to the ANCOVA model, then, except for the inclusion of the unmeasured variable(s). When the groups are analyzed separately, LISREL estimates the average difference between them on outcome, taking into account the effects of the variables included as "causes" of outcome. Such a model, however, is theoretically superior to the common ANCOVA model because it allows for measurement error in the observed covariates and for differences in the error variances between groups, while still providing the flexibility needed to explore the possibility of simple (first-order) interaction effects.

Procedures

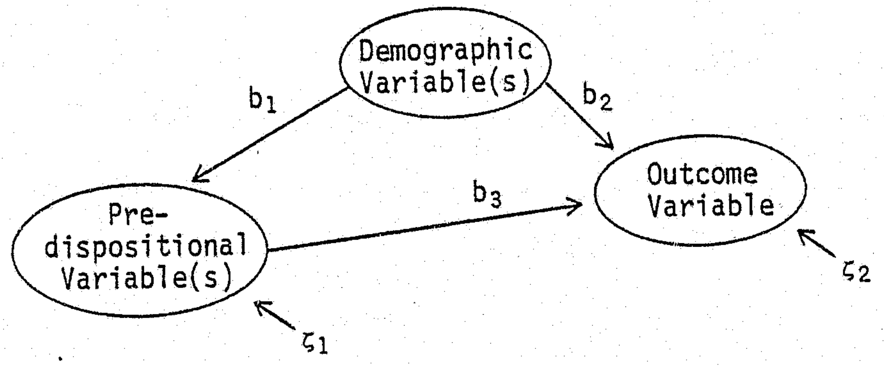
The LISREL model used in the present research is basically similar to the full LISREL model shown in Figure 4, except that it is somewhat more complex: it includes more variables and posits a causal structure among the variables used to adjust outcome scores. Although the particular variables used, and therefore the particular models specified, differ to some degree between the samples used in the study, they are similar in nature. Here, we will discuss the general models used and outline the general procedures used to estimate and test them. The specific models used for each sample will be described in chapters III and IV.

The data fall into three theoretically distinct categories: demographic variables, predispositional variables, and outcome variables. The demographic variables are considered causally prior to all other variables in the analyses. In the parlance of causal analysis, these variables are called "exogenous variables." The demographic variables included in the present study are, for the YSB sample, Sex and Age, and for the Preston sample, Age and Ethnicity (in the form of dummy variables referring to Hispanic and Black). These variables are hypothesized to have a causal effect on the predispositional variables, which comprise all other variables included as causes of outcome. Included here are prior offense variables and factors comprised of questionnaire scales. Since the offense variables tended to be skewed and limited in range, we used the natural logarithm of each so as to "robustify" the analysis.

The basic model linking these variables together is shown in Figure 5.

FIGURE 5

Basic Causal Model of Outcome



The difference between this model and the one shown in Figure 3 is that in the present case, there is an hypothesized causal relationship between the demographic variable(s) and the predispositional variable(s). This means that there are two kinds of dependent variables in the model (called "endogenous variables"), and that there is also a causal relationship among these dependent variables. Note that there is now an error term associated with the predispositional variables that was not included in the earlier causal models (figures 3 and 4); these terms refer to the variance of these variables not accounted for by demographics. Such a model does not necessarily imply that demographic characteristics themselves (e.g., "Sex" or "Ethnicity") are causes of either outcome or predispositional traits. Rather, these demographic variables refer to (or "stand for") all differences associated with the demographic characteristics which may have an influence on the variables in question. Since these demographically-related characteristics are assumed to exist prior to an individual's obtaining a prior record and certainly exist prior to his having answered the questionnaires for the original studies, they are hypothesized to have a causal influence on them. One could, of course, just as easily include the relationship between the demographics and the predispositional variables merely as "unanalyzed correlations," without affecting any of the parameters of primary interest; indeed, in the absence of any compelling reason to specify such a causal relationship, it may be best to avoid the potential misunderstanding that may result from using demographic variables as "stand-ins" for demographically-related "causal" differences. We include them here mainly to demonstrate the flexibility of the method for including such a causal structure among the covariates. Interpretation of any direct effects of these demographic variables, however, are difficult.

For each of the models used and for both samples, the procedures used to obtain an acceptable model to be used for estimating treatment effects were as follows:

1) Test the measurement model for the unmeasured variables to ensure that the same factor structure existed in both groups;
2) estimate and test the full model in its basic form (additive model);
3) modify the model as necessary to achieve an acceptable fit to the data;
4) test the statistical significance of the treatment effect estimate;
5) test for first-order interaction effects.

In order to investigate other methodological issues, we performed certain supplementary analyses with some models. To assess the effect of using log-transformed variables, we reestimated certain final models using the offense data in their original form. In other cases, we eliminated from the analysis those variables that had no direct effect on outcome, respecifying and reestimating these models so as to achieve a more parsimonious solution. These results were compared to those using the full model to assess whether the removal of these apparently irrelevant variables changed the results obtained.

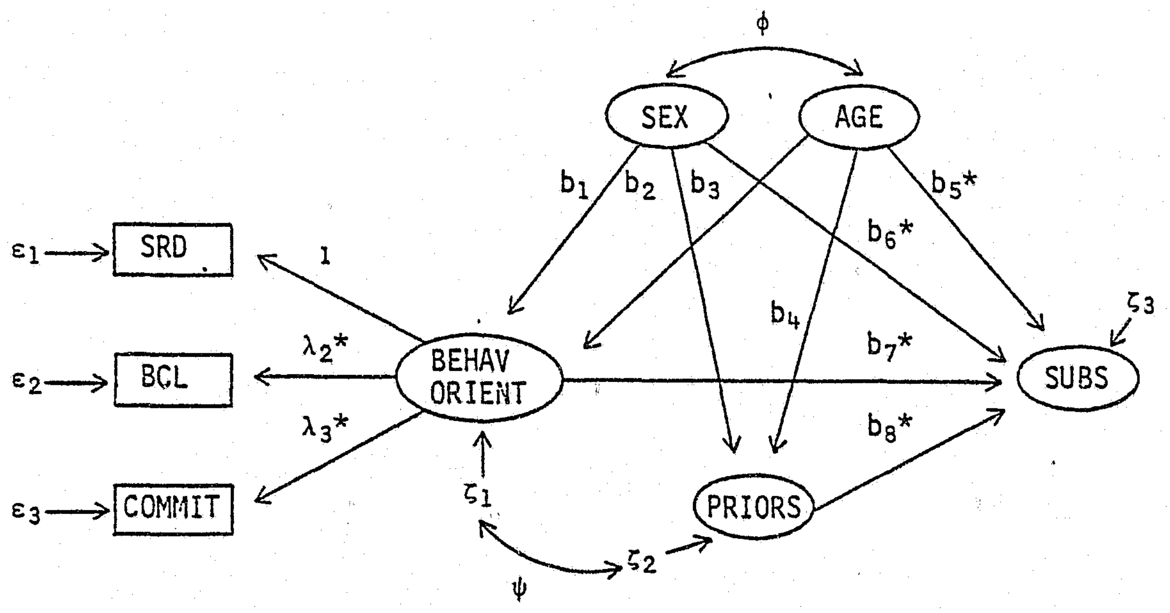
The testing of the measurement model basically involved specifying that for both groups, the same three measures were indicators of a single underlying factor and that the factor loadings for the two groups were equivalent. The measurement error terms and the variances (and covariances in the two-factor case) of the factor(s) were allowed to vary between groups.

The interest at this point was simply to determine whether such a model was consistent with the data; the actual parameter estimates were not of particular interest since they would be expected to change when the measurement model was estimated in the context of the full model. Residual correlations among error terms were added as appropriate if the model showed a poor fit. Unreasonable values of the parameters (e.g., negative error terms) indicated that the measurement model was completely unacceptable for the data (or sample) at hand.

Once it was determined that the measurement model for the unmeasured variables was appropriate, the full model was estimated and tested against the data. This model was initially estimated with all demographic variables and prior offense variables included. The full model was specified in such a way as to initially obtain the best fit to the data within the constraints of ANCOVA assumptions and with a minimum of residual correlations; this gave the best starting point for modifying the model and testing the significance of structural parameters. The measurement model was specified to be the same across groups, except that error terms were allowed to vary. The variances and covariances among the demographic variables and among the predispositonal variables were allowed to vary between groups, as were the direct effects (structural parameters) of the demographics on the predispositional variables. The direct effects of all the causal variables on outcome, however, were constrained to be equal across groups (i.e., no interaction effects were initially hypothesized). To illustrate the form of the initial specification of these models, one of the models for the YSB sample is diagrammed in Figure 6 (parameters constrained to be equal across groups are denoted by an asterisk).

FIGURE 6

Full LISREL Model With Demographics Included



Note that in this model the error terms for the indicators of the factor are now denoted by ε (epsilon) rather than δ (delta). This is simply in keeping with the conventional notation for measurement error in the indicators of endogenous, as opposed to exogenous, variables. The meaning of these error terms are equivalent for exogenous and endogenous variables; the difference is related to how they are specified for the LISREL computer program. We will continue to use the appropriate symbols for elements of the model wherever possible, so that interested readers can more easily compare our models and specifications to those in other research using this method.[7]

[7] Technically, the structural parameters from the demographic variables to other variables in the model are denoted in LISREL as γ (gamma) coefficients. We have chosen to use b (beta) to refer to all structural coefficients so as to avoid unnecessary complexity and confusion for those not thoroughly familiar with the LISREL notation.

This initial model rarely resulted in an acceptable fit to the data. The next step, therefore, was to modify the model so as to obtain a better fit, by freeing certain parameters that were fixed or constrained in this initial model. Since virtually all of the possible relationships among the variables in the structural model were allowed to be freely estimated and to vary across groups (with the exception of the direct effects on outcome, which we wished to constrain equal if possible), this modification process started with the measurement model. Using the modification indices, covariances among the error terms for the measures and between these and demographic variables were included if they were identified and made a significant contribution to the improvement of the model's fit.[8] A statistically significant covariance between the error term for the indicator of a factor and a demographic variable suggests that the relationship between that demographic characteristic and the indicator is not completely accounted for by the relationship between the demographic characteristic and the single factor. As an example, a positive covariance between AGE and BCL would simply suggest that BCL scores increase with age more than is implied by the relationship between AGE and Behavioral Orientation. The inclusion of such a covariance allows for a "purer" factor--one which is tailored to the various subgroups in the sample. The factor would still refer to a predispositional trait, but the nature of this trait would simply be understood to be constituted somewhat differently for different kinds of people. Such a process of testing, modification and retesting was continued

[8] In most LISREL applications it is not possible to include covariances between exogenous variables and the residuals for indicators of endogenous variables. However, in using "structured means" analysis, all variables in the model are specified to LISREL as endogenous variables, making it possible to include such covariances in the Theta Epsilon matrix.

until an acceptable fit was obtained within the confines of theoretical sense. As an example, it is sometimes possible to make even a badly misspecified model "fit" the data by including enough residual covariances. We included only those covariances that we felt were theoretically plausible and, in any case, stopped short of including so many that the model, for all practical purposes, had no meaning. The final "full" model was presented and the results compared to those obtained with ANCOVA using the same variables. To facilitate ease of interpretation and presentation of this model, we constrained all of the structural coefficients to be equal across groups. If such a constraint did not significantly reduce the model's fit to the data, we continued to specify them as equal.

Assuming no interactions, this "additive" model could be used as the basis for assessing the nature and significance of treatment effects. Ordinarily, before such an assessment can be made, the possibility of interaction effects must be investigated. However, we estimated this model as if no interactions existed in order to compare the results of the simple, additive LISREL model to those obtained with ANCOVA. In these models, the nature of the treatment effect was indicated by the direction and extent of the difference in the adjusted mean levels on outcome. This difference was estimated with LISREL using "structured-means analysis": the mean level of outcome is specified to be zero for controls, and the LISREL program estimates the mean of the outcome scores for treatment cases as a difference from zero. The t-value for this estimate is also provided by the program, and this statistic can serve as an indication of the probable statistical significance of this difference in means. The difference was also tested using the χ^2 difference method; the model was reestimated with the added constraint that the mean level of outcome for treatment cases was zero as well (that

the groups did not differ on outcome, taking into account the effects of the other variables). A statistically significant reduction in the model's fit with this additional equality constraint indicated that the model fit the observed data better when a treatment effect was included as part of the model.

Next, we investigated for each model the possibility of first-order interaction effects involving treatment. Such an investigation is quite simple with LISREL, since the modification indices will suggest which direct effects on outcome could be allowed to vary across groups to significantly improve the fit of the model to the data. Accordingly, we studied these modification indices to determine whether any of the direct effects appeared to differ between the groups. Possible interactions were tested through reestimating the model with these parameters freed to vary across groups; differences in χ^2 values were used to test the significance of these changes in the model. If interaction effects were found, we presented the intercept difference and calculated the interaction term coefficient.

The estimation and testing procedures outlined above served as the basis for comparing the results of this method to those obtained with ANCOVA. These procedures are those that would generally be followed if this method were used to investigate treatment effects in evaluation applications. However, certain additional methodological issues, not directly related to comparing the results of this approach to the ANCOVA approach, were of interest to us, and these were investigated through supplementary analyses. On the one hand, we were interested in the effects of "irrelevant" adjustments in the model. We investigated this issue by fixing at "zero" those direct effects with t-values, or t-ratios (estimate divided by its standard error) less than two, which indicates, under ideal circumstances, an effect that is not statistically significant. We used the χ^2 test to determine whether these coefficients should have been retained in the model, leaving out those that were not found to improve the fit of the model.[9] The result was a model in which all of the direct effect estimates had t-values over two for at least one of the groups or made significant contributions as determined by the χ^2 test. Demographic variables which were not significantly related to either the predispositional variables or outcome were removed from the model entirely. The results obtained with this "reduced model" were compared to those obtained with the "full model" and with ANCOVA.

On the other hand, we were also interested, as in the ANCOVA analyses, in the effect of having used log-transformed data in these analyses. Although the use of such transformations is methodologically justified, their effects on treatment effect estimates was unclear. In order to assess these effects, we reestimated certain models using the data in raw form and compared these results to what was obtained with the log-transformed data. We did not, at this point, go through the entire process of formulation, modification, and testing outlined above, but rather simply substituted the raw data into the final model obtained through the above procedures. Of interest was whether the use of the raw data resulted in a less acceptable fit to the data and whether treatment effect estimates differed in apparent statistical significance.

[9] We found that the t-values generally served as a good estimate of whether the parameter would be found significant by the χ^2 test, although both tests were used in all cases (since the data were not multivariate normal, we felt both tests were potentially biased, but not necessarily to the same extent, making them most useful in combination).

Combined Methods Approach

General Description

Several statistical problems of evaluation research were noted in Chapter I that are not solved by LISREL. One type of problem is interaction effects, in which the effects of one or more variables on the outcome are not homogeneous; that is, these effects are different for different people. With LISREL, as described above, direct effects on the outcome can be specified as different for treatment and comparison groups. Other forms of interactions, such as different direct effects for males and females, are more difficult so specify in these LISREL models. ANCOVA is more flexible, and virtually any type of interaction effect can be specified using multiplicative terms. With ANCOVA, however, it is cumbersome to determine what interactions should be included in the model. Loglinear models, described below, are convenient for searching for interaction effects.

The second type of problem is related to the use of limited dependent variables. Since people cannot have fewer than zero arrests, the assumption in LISREL that variables are distributed multivariate normally and the assumption in ANCOVA of homoscedastic error disturbances are likely to be violated. Violation of these assumptions results in biased estimates of statistical significance. The limitation of "never below zero" (called a "floor effect") also suggests that the assumption of a linear model is implausible, since linear relationships appropriate around mean values of the independent variables may predict negative outcomes (i.e., below zero) when extrapolated to extreme values of the independent variables (Hanushek and Jackson, 1979). Simple transformation of variables, such as the logarithm, may reduce these biases; but often nonlinear models explicitly designed for limited dependent variables will provide more satisfactory solutions. Two nonlinear models were used in this research: tobit models for continuous outcome variables and logit models for dichotomous outcomes. Their description will follow that of the loglinear models.

Loglinear Models

Loglinear models are described by Fienberg (1980), Knoke and Burke (1980), and in many other recent texts. The use of these models requires that all variables be categorical, necessitating the collapsing of continuous variables into a few categories. The objective is to account for observed cell frequencies in a cross-tabulation with the simplest possible set of assumptions about marginal distributions. This strategy is familiar in two-variable cross-tabulations, where one tests whether the cell frequencies can be accounted for by the simple marginal distributions of the two variables, using chi square as a measure of association. If the marginal distributions can account for the cell frequencies, the two variables are said to be independent; if not, the two variables are associated.

Three-variable cross-classifications are conceptually more complicated, and examples will be given using hypothetical data. Suppose we have three variables, A, B, and C, and that each has only two categories (which is convenient but not necessary). In Table II-1 below, the three variables are independent, because cell frequencies are accounted for simply by the three univariate distributions, ie., 40% of the sample is "No" on A, 50% is "Low" on B, and 33% is "Weak" on C.

In Table II-2 below, with the same univariate distributions as in Table II-1, variables A and C are associated with each other, but both are independent of variable B. The association between variables A and C can be seen as follows: people who are "No" on A are equally likely to be "Weak" or "Strong" on C (70 people in each category), whereas people who are "Yes" on A are much more likely to be "Strong" on C (140 people) than to be "Weak" (40 people). Finally, in Table II-3, variables A and C are associated (as in Table II-2); but variables B and C also are associated, while variables A and B are independent.

TABLE II-1

Hypothetical Data for Variables A, B, and C:
Model (A) (B) (C)

	Variable C					
	Weak			Strong		
Variable B:	Low	High	Total	Low	High	Total
Variable A						
No	20	20	40	40	40	80
Yes	30	30	60	60	60	120
Total	50	50	100	100	100	200

TABLE II-2

Hypothetical Data for Variables A, B, and C:
Model (AC) (B)

	Variable C					
	Weak			Strong		
Variable B:	Low	High	Total	Low	High	Total
Variable A:						
No	30	30	60	30	30	60
Yes	20	20	40	70	70	140
Total	50	50	100	100	100	200

TABLE II-3

Hypothetical Data for Variables A, B, and C:
Model (AC) (BC)

	Variable C					
	Weak			Strong		
Variable B:	Low	High	Total	Low	High	Total
Variable A:						
No	42	18	60	24	36	60
Yes	28	12	40	56	84	140
Total	70	30	100	80	120	200

In the notation used to designate loglinear models, the model depicted in Table II-1, where the three variables are independent, is (A) (B) (C). The model depicted in Table II-2, where A and C are associated and both are independent of B, is written as (AC) (B). The model depicted in Table II-3 is written as (AC) (BC). If A and B were associated also, the model would be written as (AB) (AC) (BC). Finally, if there was an interaction effect, such that the strength or the direction of the (AB) association was different among people "Weak" on C than among people "Strong" on C, then the model would be written as (ABC).

The model (AC) (B) "fits" the data in Table II-2 exactly. With real data exact fits are rare, and statistical tests must be used to determine which of the many possible models provides the best fit. Each model predicts a complete set of cell frequencies; the chi square distribution is used to compare the predicted and observed cell frequencies and to test whether that model adquately fits the data. As with LISREL, two models can be compared to determine which one fits the data better, if one model includes

all of the associations implied by the other, plus additional associations. In this case the difference in the two models' χ^2 values can be used to test whether the additional associations of the one model significantly improve the fit to the data. For example, if real data were available for variables A, B, and C, models (AC) (B) and (AC) (BC) could be compared but models (AC) (BC) and (AC) (AB) could not be compared in this fashion.

When more than three variables are cross-classified, the number and possible complexity of models increase sharply. For example if five variables, numbered 1 to 5, are cross-classified, some of the models that might be tested to determine the most parsimonious description of the data are listed below:

1. (12) (13) (14) (15) (23) (24) (25) (34) (35) (45)
2. (124) (235) (13) (15) (34) (45)
3. (124) (135) (245) (13) (23) (34) (35)
4. (2345) (123) (14) (15)

The first model specifies all possible pairwise associations among the five variables, but no higher-order interactions. This is analogous to the linear additive specification common in regression analysis. The second model specifies two second-order interactions: the (12) association is different in different categories of variable 4, and also the (23) association is different in different categories of variable 5. All pairwise associations are included in the second model, so it can be compared to the first model. The third model includes three second-order interactions, as well as all pairwise associations; it too can be compared to the first model, but it cannot be compared to the second model. Finally, the fourth model includes a third-order interaction, which specifies that the (234) interaction is different in different categories of variable 5; in addition, this model also includes a separate second-order interaction (123) and all pairwise associations.

Loglinear notation and computer software make it relatively easy and inexpensive to specify, estimate, and test loglinear models, so this is a useful procedure for investigating complex interactions. When a priori assumptions and restrictions on the data can validly be made, then loglinear models are exceptionally convenient. For example, if one of the variables clearly is dependent and the other variables are independent, as in the usual single-equation regression model, then associations among the independent variables are not of interest and they can be ignored (at the highest level of complexity). In our five-variable example, if variable 1 is dependent then the term (2345) can be specified in all models; that is, there is no need to "test" or simplify the interactions among the independent variables. This allows attention (and time and energy) to be focused on the dependent variable. In this case, the four models listed above would be specified as follows:

1. (2345) (12) (13) (14) (15)
2. (2345) (124) (13) (15)
3. (2345) (124) (135) (13)
4. (2345) (123) (14) (15)

Now the meaning of each model is clearer, as are differences between models.

Further simplifications may be possible. For example, if variable 2 is a pretest measure of the outcome variable and one is interested in searching for different rates of "gain" or maturation, then interactions involving both variables 1 and 2 would be of interest. Models 1, 2, and 4 above would be useful, but model 3 would be irrelevant because of the lack of interest in a possible (135) interaction.

Several strategies are available for searching among the large number of possible models for the "best" one. One strategy is forward stepwise inclusion of interaction terms. One starts with a model of no higher-order interactions (e.g., model 1 above), then compares models with a single second-order interaction (e.g., models 2 and 4 above); to the best of these, one adds a second second-order interaction (e.g., model 3 above), and so on until the addition of terms no longer significantly improves the model.

An alternative strategy is backwards stepwise deletion of interaction terms. One would start with a model with many higher-order interactions, such as one of the following:

5. (2345) (123) (124) (125) (134) (135) (145)
6. (2345) (1234) (1235) (1245) (1345)

If one only wanted to consider second-order interactions, model 5 would be the starting point; otherwise, model 6. Terms would be deleted in successive models, until deletion of a term made the fit to the data significantly worse. The forward and backward strategies do not necessarily yield the same "best" model, and it is recommended that both be tried, if possible.

The disadvantages of loglinear models stem from the need to work with cross-classifications of categorized variables. First, interval-level variables must be categorized. This always entails a loss of information. In addition, there is a possibility (usually quite unlikely) that the results of loglinear analysis depend on how the variables are categorized. Second, cross-tabulations require relatively large data sets. This is not a unique feature of loglinear analysis, however, but always is a requirement of models with interaction effects; it simply is uniquely explicit in loglinear analysis. In ANCOVA models with multiplicative interactions, for example, a large sample is needed to offset the problems resulting from often very high correlations between the variables and their interaction terms (multicollinearity). With cross-tabulations, the number of cells in a table is constrained by the sample size, so that with a given data set there is a tradeoff between the number of variables that can be used and the number of categories in each variable. In samples of fewer than 1,000 observations, it is impractical to analyze more than five or six variables at a time, or to have more than three categories per variable.

When loglinear models are being used as an intermediate stage of data analysis, in order to locate interaction effects to be specified in a subsequent stage (e.g., ANCOVA), these constraints are seldom serious. If there are more than six variables to be analyzed, it will be necessary to "partition" the task and investigate several six-variable cross-tabulations; for example, the same four variables may appear in every cross-tabulation, while the other two variables in each table are various combinations from the remaining variables. It is unlikely that an important interaction effect will go undetected by not controlling for <u>all</u> other variables at once. Spurious interactions are more likely, but these will be discovered in the subsequent stage of analysis. In like manner, if variables have too many categories (for the size of sample), it will be necessary to investigate several cross-tabulations, each representing an alternative way of collapsing categories.

Nonlinear Models

Nonlinear regression models are quite complicated mathematically and computationally, and they are only now becoming widely used (due mainly to advances in computer technology). Two models are appropriate for criminal justice research and will be described nonmathematically. The first is the logit model (Hanushek and Jackson, 1977), which is used when the dependent variable is categorical. Mathematically, logit and loglinear models are identical, except that with logit the independent variables may be interval-level measures. In practical terms, logit does not begin with cross-tabulation, and à large number of independent variables may be used in a single equation. Interaction effects may be specified, but, as with ANCOVA, it is not convenient to "look for" them with this method. Usually logit models are used when there are only two categories of the dependent variable, e.g., ex-offenders either commit a subsequent offense or not. In this case, the model predicts the (logarithm of the) odds of committing a subsequent offense, where "odds" is the ratio of the probability of committing a subsequent offense to the probability of not committing one. Coefficients for the independent variables indicate how the odds vary for different values of the independent variables. The nonlinear functional form, due to using odds rather than probabilities, guarantees that the underlying probability never is less than zero nor greater than one, even at extreme values of the independent variables.

Because of the nonlinearity, logit coefficients do not have the same ease of interpretation as ordinary least squares regression coefficients. At a quick glance, however, they present the same type of information. The sign of the logit coefficient indicates the direction of the effect (positive, or inverse), and the standard error can be used to test whether the effect is significantly different from zero. Magnitudes of logit coefficients are difficult to interpret. A procedure that may be useful is to convert the logit coefficient to a slope at the mean value of the independent variable; this slope is obtained by multiplying the logit coefficient by the quantity P (1-P), where P is the percentage of the sample in one category of the dependent variable and (1-P) is the percentage in the other category. Slopes calculated in this fashion are comparable to regression coefficients, e.g., they indicate (around the mean) the change in the probability of recidivism for a unit change in the independent variable.

The second nonlinear model, the tobit model (Green, 1981, 1982), is appropriate when the dependent variable has a lower limit of zero but no upper limit, e.g., number of arrests. The assumption behind this model is that the underlying concept has no lower limit, but that the operational measure of the concept is limited. For example, anti-social behavior has no conceptual limits; a person, in theory, can be infinitely pro-social or anti-social. Arrests, however, is a limited measure of anti-social behavior, focusing on the anti-social side and grouping individuals with varying degrees of pro-social behavior at the value of zero. The tobit model takes account of the possibility that the large cluster of people with zero arrests may demonstrate a wide range of values on the independent variables, consistent with a wide range of degrees of pro-social (but unmeasured) behavior.

Because of the nonlinearity, tobit coefficients are difficult to interpret, and there is no convenient transformation as with logit. Qualitatively, the usual information is available and is readily interpretable (i.e., the sign of the coefficient, and the estimated standard error). Operationally, the independent variables may be interval-level or dichotomous, and interaction effects may be specified multiplicatively, as with ANCOVA.

Procedures

The procedures used were essentially the same for all analyses, with the final estimation of parameters performed with tobit models for offense measures involving counts or rates of offenses and with logit models for dichotomized variables.

Loglinear analysis was performed by first collapsing into categories all of the variables that were not already categorical. The categorization divided the sample into roughly equal categories, the number of categories being dependent on the size of the sample and the nature of the variable.

The analysis itself was performed using either the FUNCAT procedure included in the SAS statistical software package or ECTA. The two procedures provide the same general information, but differ in terms of ease of use and computational cost. The SAS loglinear procedure allows the researcher to use raw data as input, to specify which of the variables is the dependent variable and to specify the effects to be considered in the form of a "model" statement that is conceptually similar to a regression equation (e.g., SUBS=AGE, SEX, PRIORS, AGE*SEX, AGE*PRIORS, SEX*PRIORS). All interactions among the independent variables are automatically taken into account in estimating the fit of the model to the data; effect estimates and the statistical significance of these effects are also provided. The disadvantage of this procedure is that when raw data are used, considerable computational time, and cost, may be involved in creating the necessary cross-tabulations. For our large data set, then, we used ECTA, which is not as convenient to use (the input data must be in the form of cell frequencies for the n-way cross-tabulation) nor as readily available as SAS, but which is much less expensive. Models for ECTA are specified in the form described earlier, with the variables identified by number and the effects specified by grouping variables together on a model card. For both procedures, the variables were partitioned where necessary to avoid having too many empty cells. Models were specified initially to include only the main effects of the independent variables on the outcome variable. Interactions were included in a foreward stepwise fashion until an acceptable fit was obtained.

Although the loglinear programs provide estimates of the effects of the variables, the loss of information resulting from collapsing the variables was potentially great. Therefore, treatment effects for the models found by loglinear to fit the data were estimated using tobit or logit models. Tobit estimates were obtained with the use of a computer program authored by Greene (1981, 1982). Logit estimates were obtained using the logistic regression procedure included in the BMDP software package. Logit procedures are also available in SAS.

CHAPTER III

Youth Service Bureau Evaluation Sample Analyses and Results

The primary data set used for comparing the results obtained with the various methods was established during a study of California's Youth Service Bureau (YSB) Program (Haapanen and Rudisill, 1979). In this chapter, we will describe this data set, the models and procedures used in the present study and the results obtained from the three analytical approaches: ANCOVA, LISREL, and the Combined Methods. To set the stage for the discussion, the YSB Evaluation Project will be briefly described, along with the present sample and variables. ANCOVA results using this sample and data will be presented first in order to facilitate ongoing comparison of the results obtained with the alternative methods, which will be presented in turn. Finally, a summary of the results obtained with the different methods will be presented and discussed. Results for continuous and dichotomous outcomes will be presented separately.

Sample and Data

The Youth Service Bureau Evaluation Project was a three-year study designed to assess the effectiveness of several youth service bureaus (YSBs) with respect to the goals of a) preventing or reducing delinquent behavior among clients, b) diverting young people from the juvenile justice system (JJS), and c) developing opportunities for youth to function as responsible members of their communities. Only those data collected in relation to the first goal were of interest for the present research. The YSB evaluation occurred between October 1976 and September 1979, and focused on clients seen by these YSBs during the 1977/78 fiscal year. Nine YSBs took part in this study. They were selected from among those considered by Youth Authority consultants to be the most effective, showed some interest in the research, and had available some data on clients and services provided. Results, therefore, are not generalizable to all YSBs in California.

The main thrust of the evaluation was a pre/post study of changes in delinquency involvement among YSB clients who received direct services as compared with changes showed by youths who did not receive these services. The study involved 5,954 youths (2,762 clients, 462 juvenile justice comparisons and 2,730 school comparisons). The clients comprised either the entire population or a representative subsample of clients seen by these YSBs for the fiscal year during which data were collected. Juvenile justice comparisons were chosen on the basis of rough similarity to YSB clients from police and probation departments not served by YSBs or from among youths not referred to YSBs due to unavailability of space or unwillingness to accept treatment. School comparisons were obtained from schools in three of the YSB service areas.

Official delinquency was measured by a) police contacts for delinquent behavior in which the youth was directly involved (based upon police reports), regardless of whether an arrest was made, and b) reports to police of runaway incidents. For a sizable proportion of the clients and comparisons, behavioral and attitudinal information was also obtained by means of a questionnaire. It included a self-report delinquency scale as well as numerous items designed to measure family relations, attitudes toward school, self-concept, and minor misbehavior. Pretests were obtained from 815 clients, 400 juvenile justice comparisons and 2,516 school comparisons. Finally, data on demographic characteristics were gathered on all cases.

Pre/post comparisons of police contacts showed that both clients and comparisons had slightly higher amounts of delinquent behavior during the six and twelve months following YSB intake or pretest than during a comparable time-period beforehand. This increase was evident whether we considered the number of incidents occurring during this period or the proportions of each group having police contacts. Further, the same increases were found when subgroups of clients and comparisons were analyzed separately. In general, there was no evidence of a decrease in the delinquent behavior of clients subsequent to YSB involvement. Further, rough comparisons of the rates of arrests for clients with those of the comparisons suggested that the slight increase in delinquent behavior shown by the clients was approximately the same as that observed for comparisons.

In an attempt to control as much as possible for differences between the groups, analysis of covariance was performed using a multiple regression procedure, with official delinquency (average number of police contacts per month subsequent to YSB intake or pretest) used as the dependent variable and background variables and prior police contacts used as independent variables. Background variables included age, ethnicity and sex. The model was specified initially as an additive and linear one, with interaction terms between treatment and prior delinquency included, if significant, in a subsequent step.

The ANCOVA results showed that after adjusting for the covariates, the clients and juvenile justice comparisons had virtually equal rates of subsequent police contacts. The adjusted rates for school comparisons, however, were significantly lower than that of clients, suggesting to us that ANCOVA could not completely adjust for preexisting differences. If the estimates of the treatment effects were unbiased, of course, such a finding would indicate a negative treatment effect attributable to YSBs, at least in relation to school comparisons. However, the ANCOVA adjustments were fairly small. Much of the difficulty was undoubtedly due to the extreme skewness of the outcome variable: the percentages of the clients, JJS comparisons and school comparisons with subsequent police contacts within six months were 18%, 22%, and 5% respectively. Because of the problems inherent in the method, we did not feel methodologically justified in placing primary emphasis on the results of these ANCOVA analyses. Due to time and resource constraints, attempts to more adequately control for differences among the groups or to apply weighting or other data-transformational procedures were not possible. The extent of the bias in the obtained ANCOVA results therefore remained unclear.

Present sample and data. The present research focused only on the clients and comparisons for whom pretest questionnaires were sought and obtained. A comparison of demographic and prior offense characteristics of these youths indicated that they did not constitute representative subsamples of the larger samples. Results obtained with these groups, then, cannot be generalized to the larger groups. However, the interest here is methodological, rather than substantive, and the results will not be used for obtaining generalizable findings regarding treatment effects of YSBs. To facilitate the investigation of methodological issues, we decided to restrict the sample in a number of ways. First, since we were interested in the usefulness of pretest questionnaire measures as indicators of group differences, we restricted the sample to those who completed pretest questionnaires. Second, we decided to use only official delinquency as our outcome measure (subsequent police contacts). As a consequence of this decision, it became apparent that the school comparison group was not,

as such, a very good group to use in the analysis. Within that group, both prior police contacts and subsequent police contacts were extremely skewed, with fewer than 10% of the sample having any contacts at all. Accordingly, we decided to combine the two comparison samples and restrict the analysis (for both clients and comparisons) to those who had a prior police contact within twelve months of the study. In effect, this restricted the analysis to a) juvenile justice comparisons; b) school comparisons with a police contact within twelve months prior; c) clients referred by juvenile justice agencies; and d) clients referred from other agencies who had a police contact within the prior twelve months. These two restrictions alone reduced the size of the sample to less than 450. Third, of this restricted sample, minority group members made up only a small part, and since they generally came from programs that used somewhat different selection criteria, they tended to have more serious prior and subsequent records. Due to their small numbers, we decided to omit them from the sample. Finally, in order to reduce the skewness of this variable slightly, we omitted from the analysis that handful of youths who had more than ten prior police contacts.

Since the analyses are made more interpretable by including only individuals with no missing data of any kind, a few more of the sample were omitted. However, scales used in the analysis were constructed in such a way as to minimize the attrition due to missing data. These scales were made up of items with responses arranged along a continuum of positive or negative attitudes, frequencies of behaviors, etc. (Likert scale items) and in most cases it was reasonable to assume that the midpoint of the possible range could act as a substitute for missing responses (an individual who failed to answer an attitude question for which possible responses ranged from one--very negative attitude--to five--very positive attitude--was given a score of three). Cases for whom a majority of the items of a particular scale were answered was given a score for that scale, using the midpoints for unanswered items. Such a procedure undoubtedly introduced an additional amount of random measurement error into these measures, but we felt that the increase in sample size compensated for the additional error.

As a consequence of placing the restrictions on the sample and using listwise deletion of missing data (no missing data allowed), the sample was reduced to 197 clients and 202 comparisons. These youths were White, had at least one prior police contact within the twelve months of the study, and had no missing data (except possibly on a few scale items). Although the size of this sample is very small when compared to the original sample used in the evaluation, it is probably similar to (or larger than) what might be expected from most evaluations of delinquency programs. Few evaluations have the resources to collect data on very large samples and few take as their focus a statewide program such as California's Youth Service Bureau Program. Thus, the restricted sample was felt to provide a realistic example for comparing the results of these analytic strategies.

Another difference between the present research and that carried out originally was in terms of the content of the questionnaire scales used as covariates. The original scales focused on particular substantive areas (family relations, school attitudes, peer relations, etc.) that had theoretical relevance for understanding YSB treatment and outcome. For the present purposes, however, we felt them to be too specific in their content. We felt that better measures of these characteristics would focus more squarely on the attitudes and behaviors of the youths themselves. Consequently, we constructed new scales from the questionnaire items. These new scales were designed to tap more "general" dimensions of the social

orientation of the youths. Drawing on the early work of Hirschi (1969) and the subsequent work of Wiatrowski, Griswold and Roberts (1981), we constructed scales which focused on feelings of Attachment to others, Commitment to social values (such as school), Belief in the legitimacy of the legal and moral order, and Positive Peer Association, which focused on the extent to which the respondent felt his friends were not inclined to engage in delinquent behavior. In addition to these scales we included the Self-report Delinquency scale and the Jesness Behavior Checklist (BCL) subscale measuring self-reported obtrusive behavior. This last scale was included to tap the minor end of the spectrum of misbehavior. The scales, and their constituent items, are described in Appendix A.

As discussed in Chapter II, the data were divided into three general categories: demographic variables, predispositional variables, and outcome. The demographic variables for this sample were AGE and SEX (in the form of a dummy variable referring to Male). The predispositional variables included the number of prior police contacts (PRIORS) and the aforementioned six scales made up of questionnaire items. For some analyses, these scales were grouped together into two more general factors, with factor scores calculated using standard scores for each of the scales:

(1) Behavioral Orientation (FBEHAV)--Self-report Delinquency (SRD) plus Self-report Obtrusiveness (BCL)[1] minus Commitment to Social Values (COMMIT); and

(2) Social Orientation (FSOCIAL)--Attachment to Other People (ATTACH) plus Belief in the Legitimacy of the Law (BELIEF), plus Positive Peer Association (PEERS).

[1] This scale is a subscale of the Jesness Behavior Checklist (Jesness, 1971a).

The outcome variables for this sample were the subsequent rate of police contacts per month of followup (SUBRATE)[2] and whether or not the individual had any police contacts over the followup period (IFSUBS), coded "1" for any subsequent contacts and zero otherwise. For most analyses, we used the natural logarithms of the continuous offense variables (LOGPRIOR and LOGSUBS). Treatment is indicated by a dummy variable referring to YSB clients (CLIENT).

The means, standard deviations and intercorrelations among these variables are shown in Table III-1. In Table III-2 are shown the means on the variables for clients and controls and the simple differences in means. Note that for this artificial subsample of the larger YSB Evaluation sample, the clients had a somewhat higher rate of subsequent police contacts and a somewhat lower number of prior offenses. Only one of these individual mean differences, however, was statistically significant: that for the Behavioral Orientation factor, where clients scored, on average, higher (i.e, more negatively). This lack of substantial difference between clients and controls is evidenced also by the relatively low correlations between CLIENT and all of the other variables (Table III-1). This degree of similarity was unexpected, given the nature of the research design, and under these conditions even the best of methods could not be expected to provide a substantial adjustment for group differences. Our interest, then, is primarily in the <u>relative</u> amount of adjustment made by the various methods, all of which can be expected to be small.

[2] This variable was extremely skewed due to its being a rate. Further, when it was originally constructed, some cases with relatively large numbers of police contacts or short followup periods were mistakenly retained in the sample. These errors created problems with the use of this variable as an appropriate outcome variable, but were not discovered until the present analyses were almost complete. The present results, then, are partially organized around how the various methods worked with flawed data. The distribution of this variable is shown in Appendix B.

TABLE III-1

YSB Total Sample
Means, Standard Deviations and Correlations of Variables

	CLIENT	FBEHAV	FSOCIAL	SRD	BCL	COMMIT	ATTACH	PEERS	MEAN	STD DEV
CLIENT	1.000								0.494	0.501
FBEHAV	0.099	1.000							0.019	2.325
FSOCIAL	-0.002	-0.537	1.000						0.009	2.085
SRD	0.088	0.827	-0.539	1.000					22.348	7.977
BCL	0.091	0.789	-0.326	0.537	1.000				19.669	7.049
COMMIT	-0.052	-0.724	0.395	-0.394	-0.308	1.000			17.960	3.291
ATTACH	0.048	-0.163	0.601	-0.140	-0.099	0.147	1.000		21.644	2.314
PEERS	-0.020	-0.470	0.742	-0.507	-0.304	0.287	0.123	1.000	10.697	2.780
BELIEF	-0.032	-0.494	0.746	-0.482	-0.280	0.395	0.129	0.431	14.679	2.334
AGE	-0.028	-0.019	-0.087	0.040	-0.147	-0.061	-0.002	-0.038	14.649	1.754
SEX	-0.052	0.052	-0.069	0.096	0.021	-0.008	-0.099	0.005	0.689	0.463
PRIORS	-0.068	0.149	-0.149	0.170	0.044	-0.139	-0.106	-0.128	1.915	1.568
LOGPRIOR	-0.056	0.159	-0.158	0.179	0.042	-0.154	-0.103	-0.127	0.971	0.408
SUBS	0.089	0.247	-0.215	0.220	0.143	-0.219	-0.185	-0.153	0.071	0.156
LOGSUBS	0.082	0.249	-0.225	0.231	0.137	-0.217	-0.189	-0.158	0.060	0.118
IFSUBS	0.009	0.168	-0.211	0.214	0.054	-0.127	-0.139	-0.162	0.346	0.476

	BELIEF	AGE	SEX	PRIORS	LOGPRIOR	SUBS	LOGSUBS	IFSUBS
BELIEF	1.000							
AGE	-0.139	1.000						
SEX	-0.046	-0.057	1.000					
PRIORS	-0.078	0.036	0.105	1.000				
LOGPRIOR	-0.102	0.047	0.102	0.970	1.000			
SUBS	-0.110	-0.033	0.046	0.400	0.387	1.000		
LOGSUBS	-0.123	-0.030	0.062	0.418	0.405	0.989	1.000	
IFSUBS	-0.142	-0.038	0.101	0.376	0.377	0.623	0.701	1.000

N of cases = 399

TABLE III-2

Mean Differences for YSB Sample

Variable	Mean		Difference
	Clients	Controls	
FBEHAV	.252	-.208	.460*
FSOCIAL	.005	.013	-.008
SRD	23.061	21.654	1.407
BCL	20.320	19.035	1.285
COMMIT	17.787	18.129	-.342
ATTACH	21.756	21.535	.221
PEERS	10.640	10.752	-.112
BELIEF	14.604	14.752	-.148
AGE	14.599	14.698	-.099
SEX (Male)	.665	.713	-.048
PRIORS	1.807	2.020	-.213
LOGPRIOR	.948	.994	-.046
SUBS	.084	.057	.027
LOGSUBS	.070	.051	.019
IFSUBS	.350	.342	.008
	N = 197	N = 202	

*p<.05

Results

As mentioned in the introduction to this chapter we performed separate comparative analyses for the continuous and dichotomous outcome variables. With the continuous variables, we also performed separate analyses using differing amounts of pretest information to assess the effect of omitted variables to some extent. ANCOVA and LISREL results will be presented for each of the two factors (or their constituent scales) and for both in combination. For the dichotomous variable, only the two-factor results will be presented.

The results will be discussed exclusively in terms of comparing the methods. Although significance tests were used, these are discussed relative to the conclusions that might be drawn from the various analyses using these data. Due to the artificial nature of the sample and potential problems of using the present outcome variables to evaluate YSB programs,[3] we do not intend to suggest that the statistical tests are valid for drawing conclusions about Youth Service Bureaus themselves.

Continuous Outcome Variable: Rates of Subsequent Police Contacts (SUBRATE)

With this type of outcome variable, we compared ANCOVA results to those obtained with LISREL and the Combined Methods (Loglinear and tobit, using precalculated factor scores). For ANCOVA and LISREL, we were also interested in the effect of using log-transformed data in place of the outcome data in its raw form, and consequently, these analyses were performed both ways.

To serve as a baseline for comparison, ANCOVA analyses were performed using:

(1) The individual scales, AGE, SEX, PRIORS and a dichotomous treatment variable (CLIENT), with SUBS as the outcome variable;

[3] See footnote 2 in this chapter.

CONTINUED

1 OF 2

(2) the individual scales, AGE, SEX, LOGPRIORS, and CLIENT, with LOGSUBS as the outcome variable; and

(3) the factor scores, AGE, SEX, LOGPRIORS, CLIENT, with LOGSUBS again as the outcome variable.

With these three variations, we could compare the results using raw versus logged variables and the results using individual scales versus factor scores. We first estimated an additive model by entering CLIENT and the demographic and prior offense variables in one step and then adding in the scale scores (or factor) in a second step. In this way, we could observe the effect of adding these variables on the treatment effect estimate. The coefficient for the treatment/control dichotomy (CLIENT) was the estimated treatment effect.

We next investigated the possibility of first-order interaction effects by adding the equation (stepwise) those interaction terms involving treatment. These were entered if their statistical significances reached the .05 level. If any interactions were evidenced by this procedure, we presented these results in terms of the new coefficient for CLIENT, which refers in this case to the expected difference between clients and controls when the interacting variable has a value of zero, and the coefficient for the interaction term, which refers to the difference in the effect of this variable for clients. The results for the ANCOVA analyses using the three different sets of variables were tabled together to facilitate comparison of the results.

The LISREL analyses and the Combined Methods analyses followed the general procedures outlined in Chapter II. The covariance matrices for the two groups can be found in Appendix C. For the loglinear analyses all continuous variables were dichotomized at the median, prior police contacts were

dichotomized as priors/no priors, and outcome was dichotomized as subsequents/no subsequents.

One-factor analyses (Behavioral Orientation). The results of the three ANCOVA analyses using the three measures included in this factor and the factor itself are presented in Table III-3. For the additive models, the most significant predictor in each equation was prior police contacts. Hierarchical tests showed that since clients had fewer prior contacts, the effect of adjusting for this variable was to increase the observed difference between the two groups. With the addition of the scales or the factor, this difference was reduced, but these additional adjustments could not make up for the effect of prior contacts. The treatment effect estimates shown in the table, therefore, indicate a difference in rates that is greater than the simple observed difference between the groups. When the logged variables were used, the estimated treatment effect was not statistically significant. The difference between the results for the logged variables and the raw variables suggest that the observed difference between the groups might be accounted for by differences in the numbers of each group with relatively high rates of subsequent police contacts; larger values are affected more by using logarithms than are lower values. Note that the estimated treatment effect for the analyses using the logged variables was exactly the same whether the scales themselves were used or the factor.

The tests for interaction effects in the three equations also produced consistent results, indicating an interaction between Commitment to Social Values and treatment for both the raw and logged variables. When the factor was used in place of the scales themselves, it too showed an interaction with treatment. The interaction term coefficients for the first two equations indicate that as Commitment to Social values increases, the scores on

TABLE III-3

YSB ANCOVA Estimates For One Factor Model
(Behavioral Orientation)

Variable	SUBRATE/Scales Coefficient	SUBRATE/Scales t-ratio	LOGSUBS/Scales Coefficient	LOGSUBS/Scales t-ratio	LOGSUBS/Factor Coefficient	LOGSUBS/Factor t-ratio
Scales in Behavioral Orientation Factor:						
Self-Report Delinquency......	.0017	1.54	.0016	1.86	-	-
Self-Report Obtrusiveness....	.0006	.47	.0002	.25	-	-
Commitment to Social Values..	-.0058[a]	2.47*	-.0040[a]	2.26*	-	-
Behavioral Orientation Factor..	-	-	-	-	.0091[a]	3.88**
Age............................	-.0045	1.10	-.0036	1.16	-.0028	.92
Sex (male).....................	-.0004	.02	.0034	.29	.0039	.34
No. of Prior Offenses..........	.0373	8.13**	-	-	-	-
Prior Offenses Logged..........	-	-	.1085	8.09**	.1108	8.33**
Treatment......................	.0300	2.12*	.0202	1.89	.0202	1.88
Constant.................	.1054		.0284		-.0194	
Multiple R^2	.210		.212		.208	

Interaction Model:						
Intercept Difference.........	.3203		.2240		.0206	
Interaction Term Coefficient.	-.0161		-.0112		.0140	

* p < .05
** p < .01

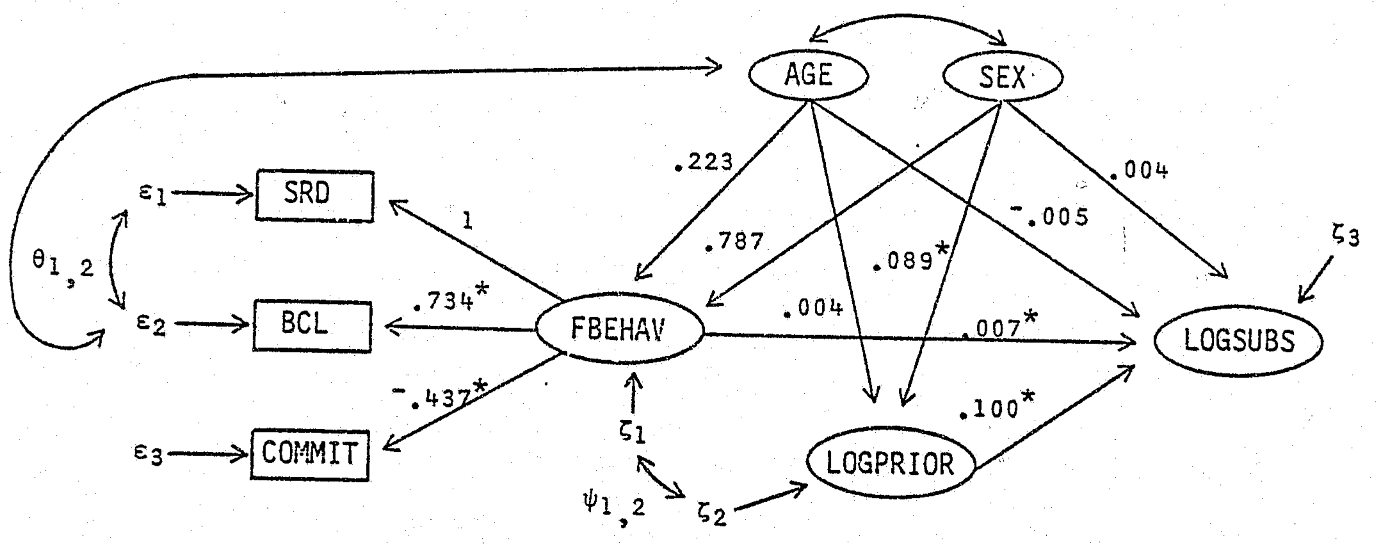
[a] Interacts with treatment in nonadditive model

the outcome measure decrease for clients more than for comparisons, with the estimated treatment effect being zero at a value of about 20 on the Commitment scale. This value is well within the range of the scale, suggesting that for some proportion of the client sample, outcome scores were predicted to be actually lower than for controls. Thus, although the additive model would indicate a negative treatment effect (taking into account the average effect of Commitment), the interaction model would indicate that the treatment effect varied according to the clients' level of Commitment to Social Values. A similar interpretation would be made with respect to the equation using the Behavioral Orientation factor. Here, however, the factor is scored in the opposite direction of the Commitment scale, so that clients are shown to do better at low values, and have higher (i.e., worse) outcome scores as their factor scores increase. Since the factor scores were established by adding or subtracting standardized scores on the individual scales, the scale for the factor was basically arbitrary, ranging from negative to positive and having an average of about zero. The intercept term for the factor in the interaction model is therefore also arbitrary, indicating the difference between clients and controls at a value slightly lower than the average for the sample. If this factor were rescaled to have a minimum value of zero. the intercept term at this point (the point of least negative self-reported behavior) would have been -.045, indicating less subsequent delinquency for clients at that lowest value.

The LISREL results for this set of variables (for the additive model) are shown in Figure 7 and Tables III-4A and III-4B. Included in the figure are the estimates that were constrained to be equal for the client and control groups: the factor loadings and direct effects. To facilitate comparing these results to those obtained with ANCOVA, the estimates of the direct

FIGURE 7

LISREL Model YSB-1:` Behavioral Orientation Factor



*t-ratio > 2.0

TABLE III-4A

Direct Effect Estimates and Test Statistics for LISREL Model YSB-1

Variable:	Coefficient	t-value
Behavioral Orientation..............	.007	3.44
Age.................................	-.005	1.83
Sex.................................	.004	0.36
Prior arrests (logged)..............	.100	7.77
Treatment (est. mean difference).....	.016	1.45
Interaction Model:		
Intercept difference................	.015	
Difference in effect of factor.......	.008 (clients higher)	

Test Statistics:

Overall goodness of fit: χ^2 = 33.66 (df=24) p = .09

Test treatment effect = 0: χ^2 = 35.68 (df=25)

χ^2 test = 2.02 (df=1) ns (rejected)

Test factor by treatment interaction: χ^2 = 28.06 (df=23) p = .21

χ^2 test = 5.60 (df=1) p < .05 (accepted)

TABLE III-4B

Other LISREL Estimates for Model YSB-1

Variance, covariance:	Clients (t-ratio)	Comparisons (t-ratio)
SRD................(θ_{11})	37.883 (4.54)	45.306 (7.84)
BCL................(θ_{22})	40.107 (6.53)	35.950 (8.51)
COMMIT.............(θ_{33})	5.452 (3.66)	7.991 (7.77)
SRD,BCL............(θ_{12})	15.206 (2.43)	13.185 (3.36)
AGE,BCL............	-1.829 (2.62)	-2.731 (3.40)
FBEHAV.............(ψ_{11})	32.071 (3.55)	9.545 (2.53)
LOGPRIOR...........(ψ_{22})	.144 (9.90)	.183 (10.03)
LOGSUBS............(ψ_{33})	.015 (9.46)	.006 (9.33)
FBEHAV,LOGPRIOR.....(ψ_{12})	.043 (2.18)	.479 (2.78)
Difference in means:		
AGE................	-.075 (.43)	0
SEX................	-.048 (1.03)	0
FBEHAV.............	1.202 (1.88)	0
LOGPRIOR...........	-.041 (1.02)	0

effects of the variables on outcome and the estimated difference between clients and comparisons (the treatment effect) are shown in Table III-4A. Also shown in this table are the χ^2 tests of the model, of the statistical significance of the treatment effect estimate and of the interaction effect for the factor (whether the effect of this factor on outcome differs between the groups). The remaining LISREL estimates are presented in Table III-4B. The variances of the demographic variables and the covariance between them was of no particular interest in this study, so these estimates have been omitted from the tables.

The χ^2 value of 33.66, with 24 degrees of freedom (Table III-4A), was not significant at the .05 level, indicating a relatively good fit to the data; that is, the model shown in Figure 7, with the factor loadings and all direct effects constrained to be equal across groups (other parameters were allowed to be different), did a reasonably good job of "explaining" the observed relationships among the variables for both groups. Comparing the direct effect estimates to those obtained with ANCOVA using logged data and the single factor, we see that they are similar. The estimated difference in outcome for clients and comparisons was .016, with a t-value of 1.45. As with ANCOVA, clients had a somewhat higher mean value on the estimated factor and a lower mean value for LOGPRIOR. The effects of these differences adjusted the treatment effect estimate in opposite directions, leading to only a slight overall adjustment for group differences. Although this estimate is close to that obtained with ANCOVA, it is slightly lower; it is, in fact, lower than the raw difference between the groups on the outcome measure (.019). Both the t-value and the chi square tests showed this effect to be nonsignificant. By and large, then, it appears that for these variables, controlling for measurement error in the pretests did slightly alter

the results. It is interesting that the residual variance on outcome for clients (Table III-4B) was estimated to be over twice that obtained for comparisons, suggesting major differences in the distributions of the raw outcome scores as well. This kind of difference may suggest problems with the data, as we will see when we discuss the findings for the Combined Methods approach.

An examination of the modification indices for this model suggested two noteworthy modifications that could be made to improve its fit to the data: allowing the effect of the factor on outcome to differ between the groups (a factor by treatment interaction effect) or a residual correlation between Commitment and outcome that differs for the groups (analogous to the Commitment by treatment interaction found with ANCOVA). The latter modification would be difficult to justify, or interpret, theoretically and would not be included in the model; it is mentioned to point out that with two-group, structured-means analysis, LISREL can locate these kinds of differences between groups even when the variables are hypothesized to comprise a single factor. An interaction between the factor itself and outcome is justifiable, however, and a test for this interaction yielded a χ^2 value of 5.6 with one degree of freedom, which would be significant at the .05 level. The nature of this treatment effect would be interpreted the same as for ANCOVA: as factor scores increased, clients did worse on followup relative to comparisons.

Since none of the effects of the demographic variables in the full model were statistically significant, the reduced form of the model excluded both AGE and SEX. The results for this model are shown in tables III-5A and III-5B. The estimates of all the parameters are very similar to those obtained with the full model, and the tests of the statistical significance

TABLE III-5A

Direct Effect Estimates and Test Statistics for
Reduced Form of LISREL Model YSB-1

Variable:	Coefficient	t-value
Behavioral Orientation.............	.006	3.31
Prior arrests (logged).............	.100	7.88
Treatment (est. mean difference)....	.017	1.51

Test statistics:

Overall goodness of fit: χ^2 = 20.04 (df=12) p = .07

Test treatment effect = 0: χ^2 = 22.25 (df=13)

χ^2 test = 2.21 (df=1) ns (accepted)

TABLE III-5B

Other LISREL Estimates for Reduced Form of Model YSB-1

Factor loadings:	Clients (t-ratio)	Comparisons (t-ratio)
SRD.................(λ_1)	1.000	constrained equal across groups
BCL.................(λ_2)	.719 (7.68)	
COMMIT..............(λ_3)	-.433 (4.25)	
Variances, covariances:		
SRD.................(θ_{11})	37.287 (4.28)	44.871 (7.61)
BCL.................(θ_{22})	40.916 (6.56)	35.611 (8.42)
COMMIT..............(θ_{33})	5.609 (3.61)	7.778 (7.50)
SRD,BCL.............(θ_{13})	15.775 (2.45)	12.622 (3.18)
FBEHAV..............(ϕ_{11})	32.641 (3.46)	10.582 (2.59)
LOGPRIOR............(ϕ_{22})	.145 (9.90)	.184 (10.02)
FBEHAV,LOGPRIOR.....(ϕ_{12})	.448 (2.24)	.499 (2.82)
LOGSUBS.............(ψ_{11})	.015 (9.52)	.006 (9.40)
Difference in means:		
FBEHAV..............	1.155 (1.79)	0
LOGPRIOR............	-.046 (1.12)	0

of the treatment effect estimate and of the interaction between the factor and treatment show the same results. The removal of irrelevant variables from this model, then, did not change the estimates of the remaining parameters.

Of some interest are the results obtained when the full LISREL model was estimated using raw, rather than logged, offense variables. These results (not tabled) showed that LISREL again adjusted the treatment effect estimate more for the pretests than did ANCOVA using the same variables. The additive model, with the same number of degrees of freedom as the full LISREL model using logged variables (24) had a chi square value of 35.69, only slightly higher than for the "log" model. The treatment effect was estimated to be .023, which was .007 lower than the ANCOVA estimate with these variables and .005 lower than the simple difference between the group means. This adjusted difference, in contrast to the ANCOVA results, was not statistically significant (t=1.61, χ^2=2.44 with 1 degree of freedom). Again, it appears that with these data, LISREL made a somewhat greater adjustment for the effects of the pretest scales than did ANCOVA. Whether logged variables or raw variables were used, the estimated treatment effect, though still negative, was smaller than the actual observed difference between the groups, even though the clients had fewer prior police contacts (and should therefore have had a higher adjusted level of subsequent police contacts). The interaction model showed the same pattern as the other LISREL models: an interaction between Behavioral Orientation and treatment, with outcomes being worse for clients as factor scores increased.

One-factor analyses (Social Orientation). The same pattern of results was obtained when the three scales making up the Social Orientation factor were employed in place of the earlier three scales. However, although all three scales were negatively related to the outcome variables, clients scored higher on the one with the largest correlation with outcome (ATTACH). The ANCOVA results, shown in Table III-6, indicate that the result is an additional adjustment in favor of comparisons--with these scales in the equation, clients had a slightly larger estimated mean difference from controls than if only PRIORS were included. Again, the results from the various ANCOVA analyses were essentially equivalent with respect to the additive model, except that when the three scales are combined into a single factor, the differences in means are almost completely cancelled out. As a consequence, when the factor is used in the equation, almost no adjustment is made, and the estimated treatment effect of .024 is the same as if LOGPRIORS alone was used as the covariate. This combination of scales also appears to dilute the effects of Attachment somewhat, so that while this scale was found to interact with treatment in the equations using the scale scores, no interaction was found for the factor.

The results obtained with LISREL were again consistent with those found for the earlier set of scales. With the full model (Figure 8 and tables III-7A and III-7B), LISREL adjusted slightly more for the Social Orientation factor, bringing the treatment effect estimate below that obtained with ANCOVA. Nevertheless, the t-value and χ^2 tests both showed the difference in adjusted means on outcome to be statistically significant, suggesting a negative treatment effect. Thus, although the estimate itself differed slightly, the conclusions that would be drawn from the analysis would be the same. As with the ANCOVA analysis using the factor scores, no interaction effect was found for these variables. However, an examination of the modification indices for this model suggested that the effect of Attachment on LOGSUBS for clients was not completely accounted for in the

TABLE III-6

YSB ANCOVA Estimates For One Factor Model
(Social Orientation)

Variable	SUBRATE/Scales		LOGSUBS/Scales		LOGSUBS/Factors	
	Coefficient	t-ratio	Coefficient	t-ratio	Coefficient	t-ratio
Scales in Social Orientation Factor:						
Attachment to Others.........	-.0092[a]	2.98**	-.0071[a]	3.04**	-	-
Peer Association............	-.0040	1.43	-.0032	1.52	-	-
Belief in Legitimacy of Law..	-.0023	-.68	-.0018	-.70	-	-
Social Orientation Factor......	-	-	-	-	-.0096	3.68**
Age..........................	-.0046	1.13	-.0036	1.16	-.0040	1.31
Sex (male)...................	-.0020	-.13	.0025	.21	.0032	.27
No. of Prior Offenses..........	.0382	8.35**	-	-	-	-
Prior Offenses Logged..........	-	-	.1116	8.56**	.1119	8.41**
Treatment.....................	.0364	2.58*	.0251	2.36*	.0241	2.26*
Constant.................	.3248		.2058		-.0040	
Multiple R^2..................	.206		.210		.205	

Interaction Model:

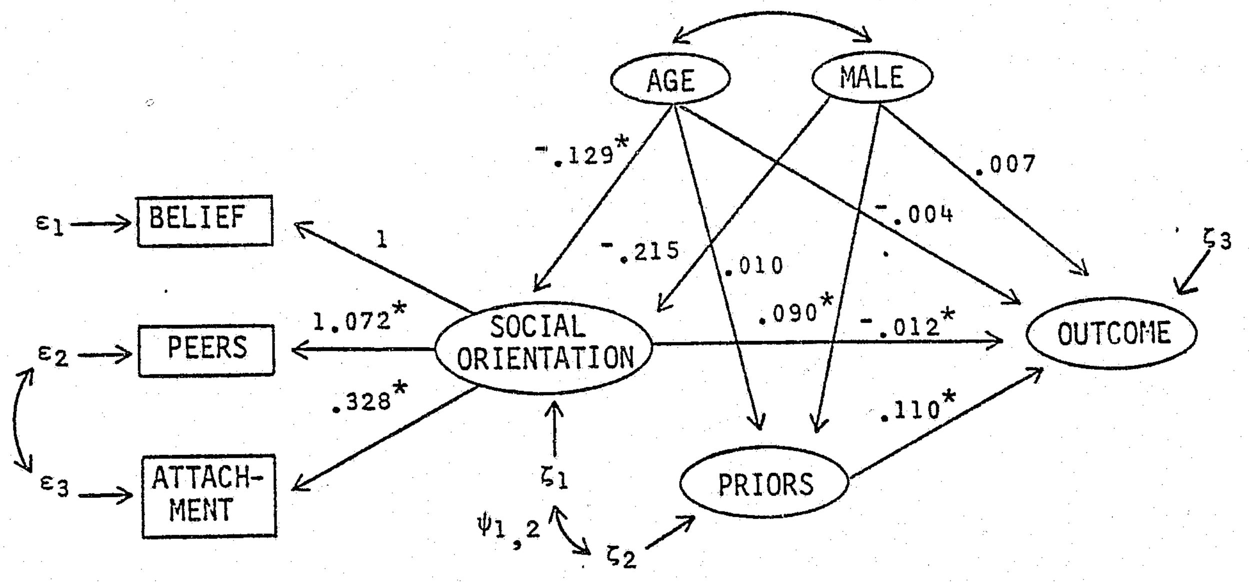
Interaction Model:	SUBRATE/Scales	LOGSUBS/Scales
Intercept Difference.........	.3537	.2672
Interaction Term Coefficient.	-.0147	-.0112

* p < .05
** p < .01

[a] Interacts with treatment in nonadditive model

FIGURE 8

LISREL Model YSB-2: Social Orientation Factor



*t-ratio > 2.0

TABLE III-7A

Direct Effect-Estimates and Test Statistics for LISREL Model YSB-2

Variable	Coefficient	t-ratio
Social Orientation factor.......	-.012	
Age............................	-.004	
Sex............................	.007	
Logpriors......................	.110	
Treatment (est. mean difference)	.023	2.141

Test Statistics:

Overall goodness of fit: χ^2 = 29.17 (df=26) p = .30

Test treatment effect = 0: χ^2 = 33.72 (df=27)

χ^2 test = 4.55 (df=1) p < .05 (rejected)

TABLE III-7B

Other LISREL Estimates for Model YSB-2

Variance, covariance:	Clients (t-ratio)	Comparisons (t-ratio)
BELIEF.............(θ_{11})	2.764 (9.54)	3.011 (9.43)
PEERS..............(θ_{22})	5.198 (5.17)	4.324 (4.55)
ATTACH.............(θ_{33})	5.872 (3.54)	4.309 (3.80)
PEERS,ATTACH........(θ_{23})	-.848 (1.63)	.648 (1.44)
FSOCIAL............(ψ_{11})	2.418 (3.06)	2.545 (3.08)
LOGPRIOR...........(ψ_{22})	.144 (9.90)	.181 (10.02)
LOGSUBS............(ψ_{33})	.017 (9.81)	.006 (9.69)
FSOCIAL,LOGPRIOR....(ψ_{12})	-.092 (1.64)	-.125 (1.95)
Difference in means:		
AGE................	-.099 (.56)	0
SEX................	-.048 (1.03)	0
FSOCIAL............	-.124 (.61)	0
LOGPRIOR...........	-.040 (1.00)	0

model. Although this interaction effect cannot be included in the model, the researcher could be aware of it and interpret the results accordingly.

When we estimated a reduced model from these data, we were interested in the effect of removing from the model a variable (AGE) that had a significant effect on the factor but not on outcome. The results of the reduced model are shown in tables III-8A and III-8B. Note that the factor loadings for FOSCIAL differ from those obtained with the full model and the difference between the group means on the factor is smaller, as is its estimated variance.

Parallel results to those obtained earlier were also found with respect to the use of raw offense measures in this LISREL model. Although the adjustments were not great enough to make the treatment effect nonsignificant, or even to bring it down to the level of the observed difference in means between the groups, a larger adjustment for the pretest scales was found: whereas a very small, but negative, adjustment for the scales was made by ANCOVA (making the clients look even worse), the LISREL adjustment was positive, overcoming rather than adding to the adjustment for PRIORS.

Two-factor analyses. The ANCOVA results obtained when all pretest scales were included are shown in Table III-9. As might be expected, the estimated treatment effects in these analyses were between those obtained with each set of pretests taken individually; in each analysis, the effect was marginally significant. The test for interactions showed that either Commitment or Attachment would interact with treatment in the equation; the effect of the Commitment by treatment interaction, however, was larger, and once it entered the equation, the effect of the Attachment by treatment interaction was no longer significant. The interpretation of this interaction effect would be the same as for the single-factor analyses involving the Commitment scale.

TABLE III-8A

Direct Effect Estimates and Test Statistics for
Reduced Form of LISREL Model YSB-2

Variable:	Coefficient	t-ratio
Social Orientation................	-.014	2.82
Prior arrests (logged)............	.110	9.40
Treatment (est. mean difference)...	.024	2.20

Test statistics:

Overall goodness of fit: χ^2 = 14.83 (df=12) p = .25

Test treatment effect = 0: χ^2 = 19.62 (df=13)

χ^2 test = 4.79 (df=1) p < .05 (rejected)

TABLE III-8B

Other LISREL Estimates for Reduced Form of Model YSB-2

Factor loadings:	Clients (t-ratio)	Comparisons (t-ratio)
BELIEF..............(λ_1)	1	constrained equal across groups
PEERS...............(λ_2)	1.586 (2.34)	
ATTACH..............(λ_3)	.509 (3.02)	
Variance, covariance:		
BELIEF..............(θ_{11})	3.573 (8.88)	3.866 (8.46)
PEERS...............(θ_{22})	3.754 (2.41)	3.021 (2.05)
ATTACH..............(θ_{33})	5.701 (5.15)	4.141 (5.60)
PEERS,ATTACH........(θ_{23})	-1.353 (-2.01)	.179 (.29)
FSOCIAL.............(ϕ_{11})	1.719 (2.62)	1.700 (2.58)
LOGPRIOR............(ϕ_{22})	.145 (9.90)	.184 (10.02)
FSOCIAL,LOGPRIOR....(ϕ_{12})	-.089 (-1.79)	-.104 (-1.86)
LOGSUBS.............(ψ_{11})	.016 (9.82)	.006 (9.74)
Difference in means:		
FSOCIAL............	-.052 (.323)	0
LOGPRIORS...........	-.046 (1.124)	0

TABLE III-9

YSB ANCOVA Estimates For Two Factor Model

Variable	SUBRATE/Scales		LOGSUBS/Scales		LOGSUBS/Factors	
	Coefficient	t-ratio	Coefficient	t-ratio	Coefficient	t-ratio
Scales in the Behavior Orientation Factor:						
Self-Report Delinquency......	.0016	1.27	.0014	1.51		
Self-Report Obtrusiveness....	.0005	.41	.0002	.17		
Commitment to Social Values..	-.0055[a]	2.25*	-.0037[a]	2.02*		
Scales in the Social Orientation Factor:						
Attachment to Others.........	-.0083	2.69**	-.0064	2.75**		
Peer Association.............	-.0014	.46	-.0012	-.52		
Belief in Legitimacy of Law..	.0021	.58	.0014	-.51		
Behavioral Orientation Factor..	-	-	-	-	.0063[a]	2.28*
Social Orientation Factor......	-	-	-	-	-.0059	-1.92
Age...........................	-.0042	1.02	-.0034	1.09	-.0035	-1.13
Sex (male)....................	-.0031	-.20	.0013	.11	.0029	.25
No. of Prior Offenses..........	.0363	7.93**	-	-	-	-
Prior Offenses Logged..........	-	-	.1059	7.92**	.1089	8.19**
Treatment.....................	.0321	2.28*	.0220	2.06*	.0212	1.98*
Constant.................	.2649		.1589		-.0076	
Multiple R^2....................	.225		.228		.215	
Interaction Model:						
Intercept Difference.........	.3601		.2548		.0221	
Interaction Term Coefficient.	-.0182		-.0129		.0161	

* $p < .05$

** $p < .01$

[a] Interacts with treatment in nonadditive model

The LISREL analysis with the two factor was not very successful, but it was instructive for understanding the limitations of the model and for understanding some of the difficulties of using LISREL with complex models such as these. A test of the measurement model showed that a two-factor structure for the six scales did fit the data; however, the calculated correlation for these factors was very high (.79 for clients and .98 for comparisons).[4] The reason for having obtained this high correlation is unclear, considering that the correlations among the various scales were only moderate. These correlations differed somewhat between groups, however, suggesting that the problem was in locating factors that were similarly constituted between the groups. Given these high correlations between the factors, continuing with the two-factor analysis would ordinarily be considered inappropriate. Nevertheless, we proceeded to estimate a full model using this hypothesized factor structure to determine whether these factors would be estimated differently in the context of that model and provide different adjustments to the treatment effect estimate.

The χ^2 value for this model was 67.1 (df=57), indicating an acceptable fit to the data. The parameter estimates for the factor loadings, the effects of the demographic variables on the factors and the effects of the demographic and prior offense variables on SUBS were all very similar to those obtained with the one-factor models. The major difference between this model and those estimated previously was in the estimated effects of the factors on SUBS, with the effect of FSOCIAL being estimated to be in a

[4] These correlations are not calculated by LISREL when covariance or moment matrices are used as input. They can, however, be calculated by hand using the LISREL estimates for the variances and covariances of the factors. It is calculated as the covariance divided by the square-root of the product of the variances.

direction opposite to that obtained with the one-factor model. The resultant treatment effect estimate was virtually zero, a profound difference from the results obtained with ANCOVA. The difference in the estimated effects of the factors, in itself, is not surprising, suggesting a high degree of collinearity between the factors (a high correlation between them). However, we again calculated the correlation between these factors, and found that, for the control group, the correlation was about 1.06. Unreasonable estimates of this kind do suggest fundamental problems with the model: The hypothesized two-factor model was apparently inconsistent with the observed relationships among the variables.

It should be noted that this correlation greater than one, which had to be calculated by hand from the LISREL estimates, was the only indication that the model was seriously flawed. Although LISREL does provide warning messages indicating problems of this kind, when structured-means analysis is used, these messages routinely appear in every solution. Had the model been estimated without attempting to estimate treatment effects--simply to test the validity of the model as a description of the causal process within the two groups--these messages would have alerted us to the problem. The researcher, therefore, must be careful when using LISREL with complicated models to estimate treatment effects. It may not be easy to notice problems such as these, since the correlation in question was only slightly larger than "1" and may not be noticeable on the basis of a casual examination of the variances and covariances estimated by the program. Each parameter in these complex models should be checked for reasonableness.

The very high correlations between the factors appears to indicate that, as estimated by LISREL, they are essentially equivalent measures. Accordingly, we reestimated the model using a single factor, General

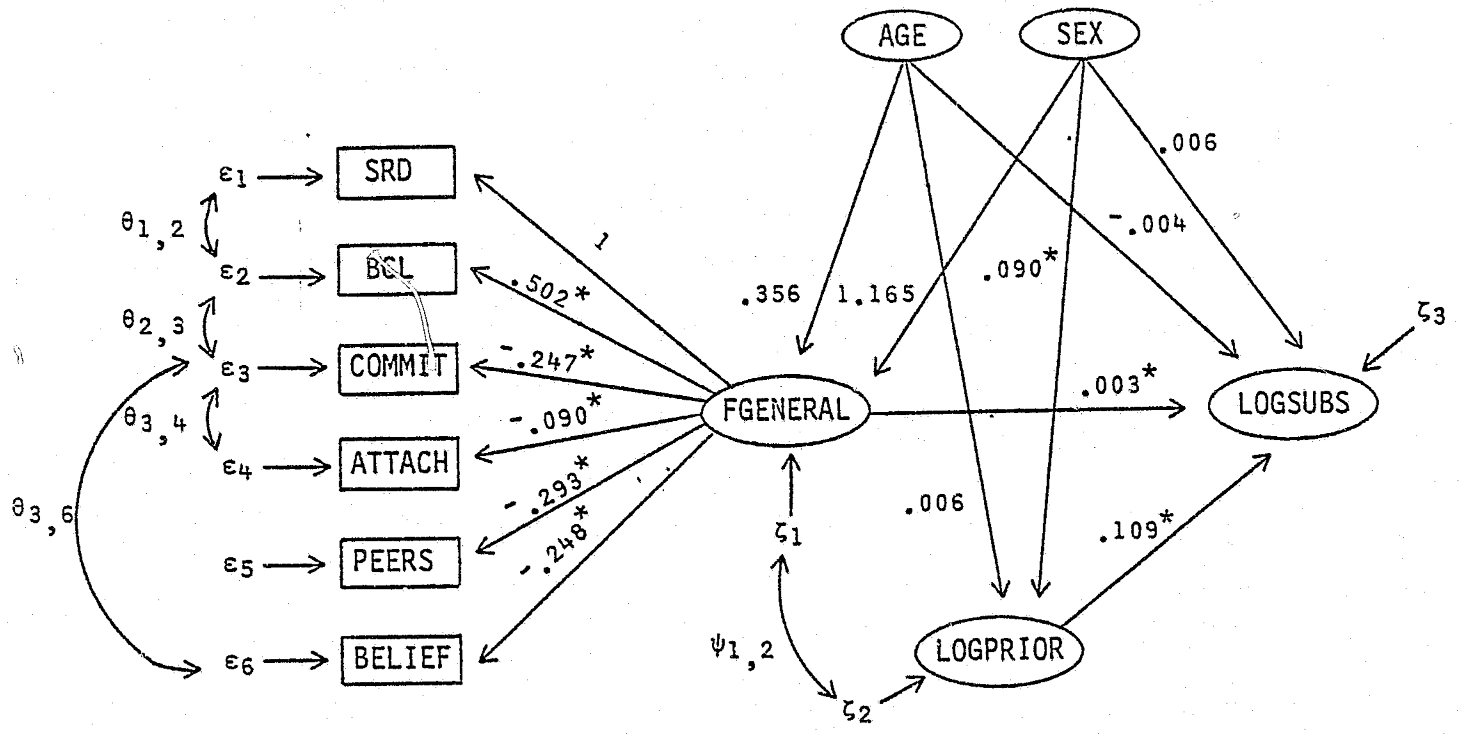
Orientation (FGENERAL), indicated by all six of the scales. The results of this analysis are shown in Figure 9 and tables III-10A and III-10B. These results show that the model does not fit the data very well (χ^2 = 92.2, df=66, p = .02), but given non-normality of the distributions of the variables and the complexity of the model, we considered it adequate for the purposes at hand. Note that the treatment effect estimate is virtually identical to that obtained with ANCOVA using all of the scales individually and logged offense data. The corresponding t-ratios for this estimate are also nearly identical. Thus, it appears that using these various scales to estimate a single predispositional factor does not change the estimate of the treatment effect obtained when the scales are used individually. The reason for this lack of increased adjustment lies in the fact that the estimated factor, although significantly related to outcome, does not differentiate between clients and controls. The estimated mean difference on the factor, although higher for clients, was not significantly higher, and its adjustment of outcome scores is not large enough to compensate for the effect of priors.

Once again, the modification indices suggested a possible interaction between the factor and treatment, which was confirmed by the chi square test. Even with this interaction included, however, the overall fit of the model is only slightly improved, with the differential relationships between Commitment, Attachment and outcome for the groups not completely accounted for.

The results obtained with the Combined Methods approach did differ in important ways from those obtained with ANCOVA and LISREL. Using various combinations of variables, loglinear analysis found no interaction effects involving subsequents. In general, models with only main effects of the

FIGURE 9

LISREL Model YSB-3: General Orientation Factor



*t-ratio > 2.0

TABLE III-10A

Direct Effect Estimates and Test Statistics for LISREL Model YSB-3

Variable:	Coefficient	t-ratio
General Orientation................	.003	3.03
Age................................	-.004	1.72
Sex................................	.006	.60
Prior Arrests (logged).............	.109	9.28
Treatment (est. mean difference)....	.022	2.02
Interaction Model:		
Intercept difference...............	.018	
Difference in effect of factor......	.004 (clients higher)	

Test Statistics:

Overall goodness of fit: χ^2 = 92.2 (df=66) p = .02

Test treatment effect = 0: χ^2 = 96.23 (df=67)

χ^2 test = 4.05 (df=1) p <.05 (rejected)

Test factor by treatment interaction: χ^2 = 87.37 (df=65) p = .03

χ^2 test = 4.81 (df=1) p <.05 (accepted)

TABLE III-10B

Other LISREL Estimates for Model YSB-3

Variance, covariance:	Clients (t-ratio)	Comparisons (t-ratio)
SRD................(θ_{11})	30.013 (6.00)	21.104 (5.11)
BCL................(θ_{22})	45.595 (9.14)	32.234 (8.75)
COMMIT.............(θ_{33})	8.306 (8.84)	8.344 (9.24)
ATTACH.............(θ_{44})	6.021 (9.75)	4.221 (9.84)
PEERS..............(θ_{55})	5.066 (8.01)	3.965 (7.79)
BELIEF.............(θ_{66})	3.092 (7.37)	3.284 (8.07)
SRD,BCL............(θ_{12})	16.059 (4.19)	3.469 (1.19)
BCL,COMMIT.........(θ_{23})	-4.067 (-3.01)	.658 (.60)
COMMIT,ATTACH......(θ_{34})	-.570 (-1.10)	1.266 (2.97)
COMMIT,BELIEF......(θ_{36})	-.018 (-.04)	1.322 (2.98)
AGE,BCL............	-1.797 (-2.58)	-2.690 (-3.49)
FGENERAL...........(ψ_{11})	36.793 (5.84)	36.346 (6.22)
LOGPRIOR...........(ψ_{22})	.144 (2.17)	.182 (10.02)
LOGSUBS............(ψ_{33})	.016 (9.85)	.006 (9.89)
FGENERAL,LOGPRIOR...(ψ_{12})	.420 (2.17)	.601 (2.84)
Difference in means:		
AGE................	-.060 (.35)	0
SEX................	-.048 (1.03)	0
FGENERAL...........	1.023 (1.46)	0
LOGPRIOR...........	-.041 (1.01)	0

predictors on subsequents fit the data very well, and no interactions involving treatment came close to being significant. Further, the main effect of treatment was found to be nonsignificant as well: clients were no more likely than comparisons to have subsequent police contacts.

The results of the loglinear analysis came as a surprise, considering that both ANCOVA and LISREL clearly indicated an interaction between the Behavioral Orientation factor and treatment. The failure to find such an interaction with loglinear models suggested that either a) too much information was lost to the analysis when the variables were dichotomized, or b) the apparent interaction effects were actually spurious, caused by outliers in the sample (individuals with extreme scores on outcome). Recall that LISREL results pointed to major differences in the variances of outcome scores between clients and comparisons, which could also have been the result of outliers, who fell disproportionately into the client sample. These cases could have been responsible for both the apparent interaction effects and for the apparent treatment effects obtained.

Upon investigation, it was determined that outliers were indeed a problem. When the fourteen cases with the highest outcome scores were removed from the analysis, the mean difference between the groups fell to .004 (from .027). ANCOVA found no significant interactions with this subsample; the apparent treatment effect found earlier also disappeared for this group. These results point clearly to the vulnerability of both ANCOVA and LISREL to extreme skewness in the dependent variable, even when it has been transformed to compensate somewhat for that skewness.[5]

[5] These cases were not responsible for the failure of the two-factor measurement model, but they did cause some of the earlier problems with the one-factor models. For example, the model using all six scales was reestimated, fit the data much better (χ^2=77.4, df=66) and no interaction effects were indicated.

Under ordinary circumstances, the researcher confronted with this result would attempt to determine whether these cases had valid scores and would probably reevaluate his or her research design. It may be necessary to modify the sample (removing the outliers) or the outcome variable.[7] Since we were interested in methodological issues, rather than in actually determining the effectiveness of YSBs, however, we continued the analysis with the present data. It allowed us the opportunity to examine the effect of the extreme skewness of the estimates obtained by the various methods. For comparison, we have included in the remaining table the direct effect estimates obtained with ANCOVA and LISREL using the reduced sample.

Having found no interactions involving subsequents, we proceeded to estimate, with tobit and ANCOVA, the two-factor model including only the main effects of the variables. Since results of a two-factor ANCOVA model using raw offense data (as was used with the tobit models), were not presented earlier, these are presented along with the tobit results in Table III-11. Due to the nature of the tobit estimates, the coefficients in the table cannot be directly compared between tobit and ANCOVA; however, the direction and estimated statistical significance of the coefficients can be compared.

The only differences between the two sets of results using the full sample, are the direction of the effect of Sex (which is essentially zero

[6] We found, for example, that extreme scores resulted both from some cases having relatively large numbers of subsequent police contacts (which may be considered unrepresentative of the populations of minor delinquents being studied) and in other cases from relatively short followup periods for certain clients (spuriously inflating their rates of contacts per month). These problem cases were not discovered in the original study. Their presence underscores the danger of using rates as outcome variables; problems with the original data may not be as apparent when they are combined into a rate nor as easily discovered during analysis.

TABLE III-11

YSB Tobit and ANCOVA Estimates For Two Factor Model:
Rate of Subsequent Police Contacts

	Full Sample		Reduced Sample
Variable	Tobit (t-ratio)	ANCOVA (t-ratio)	ANCOVA (t-ratio)
FBEHAV	.015 (1.70)	.009 (2.43)	.001 (.481)
FSOCIAL	-.026 (2.54)	-.007 (1.72)	-.005 (2.274)
AGE	-.014 (1.42)	-.004 (1.08)	-.002 (.94)
SEX	.020 (.51)	-.001 (.09)	.010 (1.18)
PRIORS	.077 (7.43)	.037 (8.20)	.021 (7.47)
CLIENT	.064 (1.78)	.031 (2.19)	.009 (1.09)
	n=399	n=399	n=385
		R^2=.213	R^2=.169

in both cases) and the estimated significances of some of the coefficients. In both analyses, clients were found to have slightly higher (i.e., worse) outcomes, with this estimate (barely) reaching the .05 level of significance with ANCOVA and only the .10 level with tobit. Similarly, the relative significances for factor coefficients differ between the two solutions. The tobit results and those obtained with the reduced sample were essentially the same, suggesting that the tobit models compensated somewhat better for the effects of the extreme cases than did ANCOVA. Both methods were affected by these cases, providing inflated estimates of the negative treatment effect, relative to those obtained with the reduced sample. The tobit results, however, would have suggested that less confidence be placed in that estimated difference.

For the continuous outcome variable, then, we found that all of the methods applied here would have led to similar conclusions for the additive model using the full sample: a slightly higher rate of subsequent police contacts for clients. Using all of the available information (factor scores for ANCOVA and tobit and a single, inclusive factor for LISREL), only the tobit model would have suggested that the adjusted difference was not statistically significant. In this regard, ANCOVA and LISREL appear to be more sensitive to the effects of outliers in the sample distributions. Still, for all practical purposes, the results did not differ. As we pointed out in the introduction to this report, a researcher can seldom be certain that important variables that could account for an observed difference between groups have not been omitted from the analysis. Further, the present analyses clearly demonstrate the problems that can be encountered by the existence of a few extreme cases in the sample. The possibility of these problems occurring with any particular data set would make one skeptical about the importance of any difference as small as that found in these analyses. Thus, we can conclude that for these flawed data, the use of methods other than ANCOVA would not have led to different conclusions regarding the programs' general effectiveness.

In relation to possible interaction effects in the data, on the other hand, the methods would have led to quite different conclusions. ANCOVA and LISREL, being more sensitive to the few cases in the sample at the extreme on outcome, would have indicated that the effect of treatment differed depending on the nature of the clients' responses to the questionnaire: their self-reported behavioral or lifestyle patterns. Since these characteristics may suggest certain policies for YSB programs (say, differential intervention strategies), the differences in results could be considered important. Short of calling for the use of loglinear methods to confirm any apparent interactions found in a particular data set, we would have to simply suggest caution in the interpretation of results obtained with ANCOVA or LISREL. A thorough, skeptically-oriented investigation of possible reasons for the results obtained with these methods could have led to the identification of the effect of the extreme cases without resort to loglinear analysis. This kind of "null-hypothesis" approach, wherein all results are considered spurious until reasonably demonstrated to be otherwise, is fundamental to all research.

Beyond the level of general conclusions, a few observations concerning the abilities of the various procedures to compensate for the problems outlined in Chapter I can be made. First, it was apparent from the ANCOVA analyses that problems resulting from skewed outcome distributions were not corrected by the addition of more pretest data, although the interaction effects may have suggested the possibility of anomolies in the data. Further, the pooling of the questionnaire data into factors was found to actually hide these interaction effects, at least with respect to the Social Orientation factor. Had these three scales been the only pretest questionnaire information available, the interaction between Attachment and treatment would have been unnoticed, perhaps leading to greater confidence in the data than was actually warranted. Thus, in situations wherein the samples are not large and the outcome variable is highly skewed, the inclusion of additional pretest information, in itself, may not be enough to overcome these potential problems of estimation.

Second, although LISREL was found also to be sensitive to these sample problems, the method was found to have certain important advantages over ANCOVA: taking pretests into greater account, providing important diagnostic information not available with ANCOVA, and forcing the researcher to confront certain problems with the data. The LISREL estimates of treatment effects were consistently lower than those found with ANCOVA, indicating that differences on the pretest questionnaire measures were taken into account to a greater extent. Part of this difference was undoubtedly due to the manner in which LISREL estimates the factors, using as much information as possible to estimate the joint effect of these measures on outcome for the two groups. Part of the difference may also have been due to having been able to specify that the error variances for the outcome measure were different for the clients and comparisons. Taking these differences in variances into account, LISREL may have been better able to compensate for the differential effects of some of the variables between groups. These variance estimates, by the way, could also serve as diagnostic tools, since differences such as those found for this sample could alert the researcher to the kinds of problems we found. More important in most cases, however, are the modification indices, which not only provide the user with information regarding how the model might be modified to better fit the data but also point to possible flaws in the overall model and unaccounted-for correlations. For example, recall that in the one-factor model involving Behavioral Orientation, the modification indices suggested the factor by treatment interaction and also indicated that the source of this interaction lay primarily in the fact that the model did not adequately account for the relationship between Commitment and outcome, which differed between the groups. This information would be lost if the scales were simply combined into a factor and submitted to ANCOVA analysis. Finally, LISREL forces the user to confront certain problems with the model, such as those found with the two-factor solutions. In general, although the method is difficult to learn and somewhat complicated to use, once mastered, it can provide a good deal more information than can be obtained with ANCOVA. With these data, it was not able to provide much of an improvement in the estimates of treatment effects, however, suggesting that the problems associated with badly skewed data are more serious and fundamental than can be remedied by a method such as this.

The Combined Methods approach, and in particular the loglinear analyses, pointed to the value of being aware of the distributional properties of the data. It is often too easy to simply ignore these issues, relying on the robustness of ANCOVA to compensate for poor data. With the growing availability and use of loglinear methods, it may be wise to go through the exercise of applying them, under the assumption that "important" effects (as opposed to merely statistically significant ones) will be apparent even when the data are violently reduced through gross categorization. Differences obtained with regression techniques and loglinear ones would serve as a starting point for understanding the sources of the effects found with ANCOVA.

The tobit results suggest that it is possible to compensate for poorly distributed data to some degree and that ANCOVA results could lead to mistaken conclusions if taken too seriously. In contrast to loglinear, however, tobit programs are not readily available to most researchers. Since the differences between the results were not very large, it would seem advisable again simply to call for caution in the use and interpretation of ANCOVA results.

Dichotomous Outcome Variable

For these analyses, all scales (or both factors) were included in the equations predicting whether or not the individual had any subsequent police contacts. The proportions of the client and comparison groups who fell into this category were .350 and .342, respectively, for a difference of .008 in favor of the comparison group. This difference is not statistically significant. As discussed previously, loglinear analysis with this sample showed no significant interaction effects, so here we will only present findings for the additive logit and ANCOVA models.

In Chapter II we argued, following Goodman (1976), that when the predictability of the outcome variable is not high (indicated by a low R^2 figure) and the proportion of the sample falling into either category of the dichotomous dependent variable is not below .2 (or above .8), the problem of heteroscedastic error terms should not be major. ANCOVA results, then, should provide fairly good estimates of the treatment effect and its statistical significance. The results shown in Table III-12 bear out this contention. Comparing the ANCOVA model using the factor scores with the logit model, we see that the coefficients all have the same sign and the ratios of the coefficients to their standard errors are almost identical.

TABLE III-12

YSB Logit and ANCOVA Estimates For Two-Factor Model: Any Subsequent Police Contacts

	ANCOVA/Scales		ANCOVA/Factor		LOGIT/Factors	
Variable	Coefficient	t-ratio	Coefficient	t-ratio	Coefficient	t-ratio
Scales in Behavioral Orientation Factor:						
Self-Report Delinquency......	.008	2.16				
Self-Report Obtrusiveness....	-.005	1.58				
Commitment to Social Values..	-.002	.26				
Scales in Social Orientation Factor:						
Attachment to Others.........	-.016	1.63				
Peer Association.............	-.008	.80				
Belief in Legitimacy of Law..	-.009	.83				
Behavioral Orientation Factor..	-		.006	.58	.034	.59
Social Orientation Factor......	-		-.032	2.60	-.170	2.52
Age...........................	-.020	1.56	-.015	1.26	-.082	1.22
Sex (male)....................	.043	.89	.053	1.10	.129	1.00
No. of Prior Offenses..........	.102	7.13	.106	7.43	.546	5.97
Treatment.....................	.027	.61	.029	.65	.066	.56
Constant.................	.928		.325		.576	
Multiple R^2..................	.185		.174			

Both analyses show clients to be slightly more likely to have subsequent police contacts, but this difference is far from statistically significant.

It is interesting to note that in this analysis the coefficient for BCL is of the opposite sign to that found when predicting the rate of police contacts per month. Since its effect works opposite to that of SRD, their effects are cancelled somewhat when combined into a single factor.

These results, then, substantiate the Goodman's argument that when the dichotomous outcome variable is not extremely skewed, and when the R^2 for the equation is relatively small, ANCOVA (using ordinary least squares regression) will provide reasonably good estimates of the effects of the variables in the equation and of their statistical significances.

CHAPTER IV

Preston Sample: Analyses and Results

We had originally proposed to use as a second data set one that was developed during a study of alternative treatment methods for first-time juvenile probationers. In analyzing the YSB data, however, it was determined that the relatively small size of the data set created certain problems for estimation, particularly where some correlations were essentially zero. Testing for interactions using loglinear models requires fairly large samples if more than a few factors, measures, and background characteristics variables are included in the analyses. Since the probation data set was also small, we decided that the full analysis of this data set would probably provide little additional methodological information. Therefore, we decided to use, instead, a larger data set originally developed in the course of evaluating an experimental program in a Youth Authority institution (The Preston Typology Study). We felt this data set might provide a better basis for comparing the results obtained by the different methods.

Sample and Data

The Preston sample consisted of 1,622 male youths who were committed to the Preston School of Industry during a 13-month period from February 1966 to March 1967.[1] Preston is a large California Youth Authority institution which at that time housed approximately 900 wards in 16 living units. The youths sent to Preston ranged in age from 16 to 20 (median 17.6) and remained

[1] A detailed description of this project (The Preston Typology Study) can be found in the project report (Jesness, 1969) and a summary description in a subsequent article (Jesness, 1971b).

in the institution for an average of 8.4 months. Most youths sent to Preston had more lengthy and serious records than those referred to other facilities--57% had previously been committed to a Youth Authority institution.

Five of the 16 units at Preston housed wards meeting special criteria in that they had been cleared for work outside the confines of the institution or had been assigned to one of two psychiatric treatment units. All subjects who were not preselected for special placement in one of these units were placed in a pool of eligibles who were then assigned by random methods to either an experimental or control group.[2] Experimental subjects were subsequently placed in one of six living units according to their I-level subtype classification (Jesness, 1974). The present study included only those youths assigned to experimental (n = 458) or control (n = 636) groups.

In this study, extensive demographic, psychological and behavioral data were collected and used to develop a typology scheme and test a differential treatment approach based on Interpersonal Maturity Level (I-level) subtype.

This sample was subsequently included in the Early Identification of the Chronic Offender Project, undertaken to explore the extent to which chronic adult criminal (and violent) offenders could be identified early in their careers (Haapanen and Jesness, 1982). Followup arrest data covering the early adult years of peak criminal activity (from approximately 18 to 26 years of age) were obtained, primarily from official arrest records of the California Bureau of Criminal Investigation and Identification (CII). Supplementary data were obtained from the Federal Bureau of Investigation

[2] In a small number of cases this procedure was circumvented in order to maintain racial balance in the various institutional living units. As a result, some additional minority members with higher I-levels were placed in the experimental groups.

(FBI) and the California Bureau of Vital Statistics to ensure that individuals with no records--or only minor records--of arrests in CII files did not have records in other states and/or were not deceased. The median followup period for this sample was 11.7 years, at which time the median age of the sample was 29.

The coding and summarization of the followup offense data focused on arrest incidents. The most serious charge for each arrest was recorded and subsequently classified as being a violent-aggressive (murder, manslaughter, assault, rape), violent-economic (robbery, kidnapping, extortion), property, or minor offense.

The followup data showed that a high percentage of the juvenile offenders engaged in serious criminal activity as adults. Most (66%) were arrested for one or more violent offenses (murder, rape, assault, robbery), and over 80% were arrested for at least one felony offense. Consequently, most of the subjects (86%) were classified as chronic offenders. During the approximately 10 years following their incarceration as juveniles, the 1,622 offenders in the sample were arrested a total of 17,059 times, for an average of 10.52 offenses per subject. Of these arrests, 2,997 were for violent offenses (violent-aggressive plus violent-economic). These arrest data taken from rap sheets undoubtedly understated the total number of offenses that occurred; they did not reflect, for example, the number of undetected crimes committed or the number for which no arrests were made. The amount of hidden crime involved can be estimated from data presented by Peterson and Braiker (1980). These authors administered extensive questionnaires to a large sample of California Department of Corrections inmates. Among those inmates who had serious juvenile records, the official rap sheets showed an arrest for only one out of every six self-reported robberies and one out of every 20 self-reported burglaries. If the number of arrests for robbery and burglary were multiplied by these figures, it would be clear that the offenders in these samples were responsible for a very large number of crimes. The tendency of these official data to underestimate offenses we felt would more than compensate for any over-estimation occurring as a result of using "arrests" rather than "convictions" as an indicator of criminal behavior.

Prior analyses. In the original study, institutional adjustment and parole outcome were compared between wards assigned to living units based on I-level subtype and those assigned on the basis of traditional criteria. Some differences in institutional adjustment were found, but although the treatment group had fewer parole revocations during 24 months of parole followup, these differences did not reach statistical significance.

No attempt was made during the Chronic Offender study to assess the possible effectiveness of the differential treatment program in relation to the development of long-term criminal or violent careers. Although it might seem unlikely that the institutional experiences of these very serious delinquents during one period of incarceration would have a marked effect on their overall criminal careers (especially given the extensiveness of these careers), we felt that the issue was worth investigating, both substantively and methodologically. On the one hand, the sheer numbers of offenses committed by these wards subsequent to their stay at Preston makes important differences possible, even from a rather "small" treatment effect. A 10% reduction in violent offenses, for example, would mean 300 fewer violent crimes for the whole sample. Any evidence that differential treatment may have an effect on criminal careers, then, might be of considerable importance.

Methodologically, this data set allows not only for a fuller application of some of the analytical techniques (because the sample size is large) but also allows for investigating the effects of sampling error in true experimental designs. This study used simple random assignment of subjects to treatment and control conditions. Although the assignment process was circumvented in a small number of cases to ensure racial balance within the various living units, any major preexisting differences between the groups should be the result of simple sampling error. These differences should not be large, but better treatment effect estimates should still result from taking those that are substantial into account.

Present sample and data. For the present study, we focused primarily on violent arrests (murder, manslaughter, assault, rape, robbery, extortion) occurring during the followup period. The distribution on this variable was skewed for both treatment and control groups, with 32% of the former and 28% of the latter having no subsequent violent offenses (this was the modal category in both cases). In addition, this outcome was the most predictive for this sample as a whole, making it more likely that any preexisting differences could have an effect on predicted outcome differences for the groups.

The sample and data were selected using similar procedures to those employed with the YSB sample. Because the sample was much larger and included only males, we included all ethnic groups. Cases were otherwise excluded if they had fewer than half of the items comprising various scales. As part of the pretest battery, a 136-item questionnaire was administered that tapped the subjects' perceptions of parents and family, and their opinions about school, prior offenses, and home and community environment. An additional 20 items covered a priori dimensions of Self-Concept, Fate

Control, Neutralization, and Alienation. From these items, scales as similar in focus and content to the self-report scales constructed for the YSB sample were constructed for this sample as well.[3] The remaining scales could not be duplicated. This similarity allowed us to explore the comparative usefulness of these kinds of behavioral variables for minor and serious delinquents and also facilitated the ease of presenting the results for the two data sets. Again, missing items were coded at the midpoint of the range of the item and scale scores were calculated only for those with at least half of the items. The final sample comprised 410 experimentals and 552 controls.

The variables used in the present analysis are not the same as those used in the YSB analysis. For simplicity, they have been given similar names. The independent variables are:

1) Self-report Delinquency (SRD), which for this sample excludes the more minor delinquent acts;
2) Observed Obtrusive behavior (BCL);
3) Commitment to Social Values (COMMIT);
4) Behavioral Orientation (FBEHAV), created by adding standard scores for SRD and BCL and subtracting the standard score for COMMIT;
5) Race (BLACK), a dummy variable referring to Black ethnicity;
6) Race (HISPANIC), a dummy variable referring to Hispanic ethnicity;
7) AGE, which ranges from fifteen to eighteen;
8) Number of prior violent arrests (PREVIOL) or its logarithm (LOGPVIOL);

[3] The items making up these scales are presented in Appendix D.

9) Number of prior nonviolent arrests (PRENVIOL), or the log (LOGPNVIO); and,
10) TREATMENT, a dummy variable coded "1" for the treatment group.

Outcome variables included:

1) Number of subsequent violent arrests (TOTVIOL), including murder, assault, rape, robbery, extortion and kidnapping [see Appendix E for the distribution] or its log (LOGVIOL);
2) NOVIOL, a dummy variable referring to no subsequent violent arrests;
3) FEWVIOL, a dummy variable for one or two subsequent violent arrests; and,
4) MANYVIOL, a dummy variable for more than two subsequent violent arrests.

The means, standard deviations and intercorrelations among these variables for the total sample are shown in Table IV-1. Note that the correlations among these variables are, at best, moderate or small. As with the YSB sample, these small correlations present a challenge for even the best analytic methods. Nevertheless, they are probably typical of the kinds of data with which a reseacher in the field of criminal justice would likely be faced, making them appropriate for comparing analytic strategies of interest to criminal justice researchers. Interesting is the fact that although the SRD, BCL, and Commitment scales (as well as the Combined factor) are all correlated in the expected direction with the number of subsequent violent offenses (TOTVIOL), these scales are virtually uncorrelated with the prior offense variables. Similarly, these scales are not correlated with the ethnicity variables in a manner consistent with the positive correlations

TABLE IV-1

Preston Total Sample
Means, Standard Deviations and Correlations of Variables

	SRD	BCL	COMMIT	FBEHAV	BLACK	HISPANIC	AGE	MEAN	STD DEV
SRD	1.000							2.406	2.472
BCL	0.070	1.000						14.817	3.477
COMMIT	-0.162	0.021	1.000					9.780	2.681
FBEHAV	0.665	0.567	-0.617	1.000				0.002	1.849
BLACK	0.029	0.182	0.027	0.099	1.000			0.320	0.467
HISPANIC	0.159	-0.128	-0.063	0.051	-0.348	1.000		0.205	0.404
AGE	-0.139	-0.041	0.053	-0.126	-0.122	-0.056	1.000	16.940	0.794
PREVIOL	0.083	-0.006	0.010	0.036	0.286	0.012	0.032	0.440	0.713
LOGPVIOL	0.086	-0.017	0.009	0.032	0.284	0.020	0.017	0.273	0.402
PRENVIOL	-0.010	0.016	0.019	-0.007	-0.024	-0.021	0.295	1.945	1.759
LOGPNVIO	0.000	0.043	-0.003	0.025	-0.066	-0.019	0.285	0.934	0.535
TOTVIOL	0.156	0.111	-0.055	0.173	0.277	0.033	-0.046	1.925	1.999
LOGVIOL	0.153	0.128	-0.061	0.183	0.264	0.053	-0.063	0.855	0.667
NOVIOL	-0.120	-0.134	0.068	-0.172	-0.177	-0.053	0.064	0.296	0.457
FEWVIOL	-0.013	0.057	-0.016	0.033	-0.056	-0.014	-0.007	0.384	0.487
MANYVIOL	0.131	0.071	-0.050	0.135	0.231	0.066	-0.055	0.320	0.467
TREATMNT	0.044	0.026	-0.083	0.083	0.035	-0.021	-0.062	0.426	0.495

	PREVIOL	LOGPVIOL	PRENVIOL	LOGPNVIO	TOTVIOL	LOGVIOL	NOVIOL	FEWVIOL	MANYVIOL
PREVIOL	1.000								
LOGPVIOL	0.977	1.000							
PRENVIOL	-0.156	-0.187	1.000						
LOGPNVIO	-0.264	-0.315	0.928	1.000					
TOTVIOL	0.222	0.214	0.086	0.079	1.000				
LOGVIOL	0.221	0.214	0.096	0.092	0.935	1.000			
NOVIOL	-0.148	-0.145	-0.075	-0.083	-0.625	-0.832	1.000		
FEWVIOL	-0.067	-0.060	0.005	0.024	-0.190	0.022	-0.512	1.000	
MANYVIOL	0.214	0.204	0.068	0.056	0.810	0.792	-0.445	-0.541	1.000
TREATMNT	0.061	0.070	-0.122	-0.121	-0.015	-0.028	0.048	-0.062	0.017

N of cases = 962

between ethnicity and subsequent offenses. These low correlations and consequent inconsistencies underscore the difficulties of understanding variations in criminal behavior within populations of relatively serious delinquents.

The means for the total sample indicate that these individuals averaged two subsequent violent offenses each. Only 30% of the sample had no subsequent violent arrests, and of those that did (70%), almost half had more than two subsequent arrests for violent offenses. Numbers of violent offenses ranged from 0-16, with an overall average of 1.93. Only 13 cases (1.4%) had more than seven subsequent violent arrests, but the size of the sample precluded their having an adverse effect on the analyses.

The means of these variables for each group and the simple differences between these means are shown in Table IV-2. It is apparent from these figures that these two groups differed, at least statistically, more than did the patently nonequivalent groups used for the YSB analysis. Although the size of the sample undoubtedly made small differences statistically significant, these differences point to the potential for even well-executed random assignment designs to create groups that differ in important ways. Three of the eight major background characteristics (Behavioral Orientation and Commitment are redundant) were found to differ significantly between the two groups. There was only a slight difference between them in terms of the number of subsequent violent arrests (TOTVIOL), but the direction of the differences between these groups on the Behavioral Orientation scales and the number of prior violent arrests would suggest that the treatment group be predicted to have more subsequent arrests than the controls. Adjusting for these differences might therefore be expected to increase the predicted difference in outcome, leading to a positive treatment effect.

TABLE IV-2

Means and Mean Differences for Preston Sample

	Treatment	Control	Difference
SRD	2.532	2.313	.219
BCL	14.921	14.740	.181
COMMIT	9.522	9.971	-.449**
FBEHAV	.179	-.131	.310**
AGE	16.883	16.982	-.099†
RACE (Black)	.339	.306	.033
RACE (Hispanic)	.195	.212	-.017
VPRIOR	.490	.402	.088†
LOGVPRI	.305	.248	.057*
NVPRIOR	1.695	2.130	-.435**
LOGNVPRI	.858	.990	-.132*
TOTVIOL	1.890	1.951	-.061
LOGVIOL	.833	.871	-.038
NOVIOL	.322	.277	.045
FEWVIOL (1 or 2)	.349	.409	-.060
MANYVIOL (over 2)	.329	.313	.016
	n=410	n=552	

†p <.10

*p <.05

**p <.01

The lower number of nonviolent priors observed for the treatment group would mitigate this adjustment to some degree, but the number of priors of this kind are less predictive of outcome than are the variables on which treatment cases had higher mean values. Glancing at the proportions of each group with no, few, or many violent subsequents, we see that the treatment cases are slightly overrepresented among those with no arrests and slightly underrepresented among those with only one or two arrests. Simple chi square tests of these differences showed neither was statistically significant, but again we might expect the difference to be more meaningful after correcting for preexisting differences between the groups.

Results

The analyses for this sample were carried out in the same way as for the YSB sample, with the exception that only one factor was available. For the continuous outcome variable (total subsequent violent arrests), we performed ANCOVA using the same procedures as before. Results are presented for analyses using a) the individual scales and raw scores on offense variables, b) the individual scales and the logged form of the offense variables, and c) the factor scores and logged offense variables. These differences provide for assessing the effects of using raw vs. logged data and scales vs. factor scores. LISREL analysis employed only the logged data;[4] and the tobit models were estimated using raw data, since the method is designed to compensate for the skewness of the variable. Due to the large size of tne sample, loglinear analyses were performed with the outcome variable and the scores on the Behavioral Orientation factor collapsed into three, rather

[4] The covariance matrices, with the means and standard deviation for each group, are presented in Appendix F.

than two, categories. The categories of the outcome variable corresponded to the categories indicated by the dummy variables mentioned earlier (no subsequents, one or two subsequents, and three or more subsequents), each category containing roughly a third of the sample. Behavioral Orientation was also collapsed in such a way as to obtain roughly equal numbers in each category. Ethnicity was entered as a three-category variable as well. Offense history variables were dichotomized: violent priors (0, 1+), non-violent priors (0 or 1, 2+). In separate analyses, AGE (15-16, 17, 18) was used in place of ethnicity in the loglinear models. Logit models were estimated separately for predicting no violent subsequents and three or more violent subsequents.

Continuous Outcome Variable: Total Violent Subsequent Arrests (TOTVIOL, LOGVIOL)

The ANCOVA results for this sample are shown in Table IV-3. As expected, the adjustment for preexisting differences increased the predicted difference between the groups in every case. None of the adjustments, however, resulted in a treatment effect estimate that was statistically significant. When logged variables were used in place of the actual numbers of arrests, the overall solution was relatively unchanged. The slightly higher t-ratio for the estimated effect of Age suggests that the relationship between this variable and outcome may not be linear; when the outcome variable was logged, the effect of age differences on lower values of outcome was given more importance than its effect on higher values. The increased predictive power of this variable, on which treatment cases had a lower mean, was also probably responsible for the slight increase in the t-value for the treatment effect estimate. In the equation using logged variables, we also found a marginal interaction effect between scores on the Commitment scale and treatment.

TABLE IV-3

ANCOVA Estimates For Preston Sample

Variable	TOTVIOL/Scales Coefficient	TOTVIOL/Scales t-ratio	LOGVIOL/Scales Coefficient	LOGVIOL/Scales t-ratio	LOGVIOL/Factor Coefficient	LOGVIOL/Factor t-ratio
Scales in Behavioral Orientation Factor:						
Self-Report Delinquency......	.083	3.30**	.023	2.81**	-	-
Self-Report Obtrusiveness....	.041	2.32**	.017	2.94**	-	-
Commitment to Social Values..	-.033	1.46	-.012[a]	1.59	-	-
Behavioral Orientation Factor..	-	-	-	-	.049	4.53**
Prior Violent Offenses.........	.459	5.07**	-	-	-	-
Prior Violent Offenses Logged..	-	-	.332	5.97**	.333	6.02**
Prior Nonviolent Offenses......	.145	3.98**	-	-	-	-
Prior Nonviolent Offenses Logged......................	-	-	.229	5.53**	.230	5.57**
Age...........................	-.086	1.06	-.056	2.09*	-0.57	2.13*
Race (Black)..................	1.084	7.33**	.342	6.99**	.347	7.15**
Race (Hispanic)...............	.542	3.33**	.208	3.85**	.207	3.89**
Treatment.....................	-.114	.93	-.054	1.32	-.055	1.37
Constant.................	2.005		1.180		1.386	
Multiple R^2...................	.149		.163		.161	
Interaction Model:						
Intercept Difference.........			-.341			
Interaction Term Coefficient.			.030			

* p < .05
** p < .01
[a] Interacts with treatment in interaction model

However, the effect was small, the negative intercept term and positive coefficient indicating a slight tendency for treatment to work better for those with low values on the Commitment scale. The same overall adjustment in the treatment effect estimate was obtained when factor scores were used in place of the scales themselves. This was to be expected, since each of the scales were predictive of violent offenses in the expected direction and since treatment cases scored, on average, higher on SRD and BCL and lower on COMMIT. Thus, nothing was gained through combining the scales into a factor; we lost, however, a certain amount of information concerning the effects of Commitment. Given the marginality of the apparent interaction effect, however, its importance is probably minimal anyway.

The LISREL results using the set of variables were quite interesting in a number of ways. The model used was similar in nature to that used for the YSB sample, as shown in Figure 10. The direct estimates of the variables on outcome and the estimated difference between the two groups are shown in Table IV-4A, along with the statistical tests for the model. In general, the results are consistent with those found with ANCOVA, with three major exceptions: the effect of the factor is shown to be considerably larger and the direct effects of Age and Hispanic ethnicity are no longer statistically significant. The reduction in the relative predictive powers of these two demographic characteristics was apparently the result of how the factor was estimated. In order to obtain an acceptable fit of the model to the data, we had to take into account the fact that both Age (for treatment cases) and Hispanic ethnicity (for both groups) were correlated with BCL in a direction opposite to that which is implied by the positive correlation between each of these variables and outcome. These inconsistencies were handled through adding to the model correlations between the

FIGURE 10

LISREL Model for Preston Sample

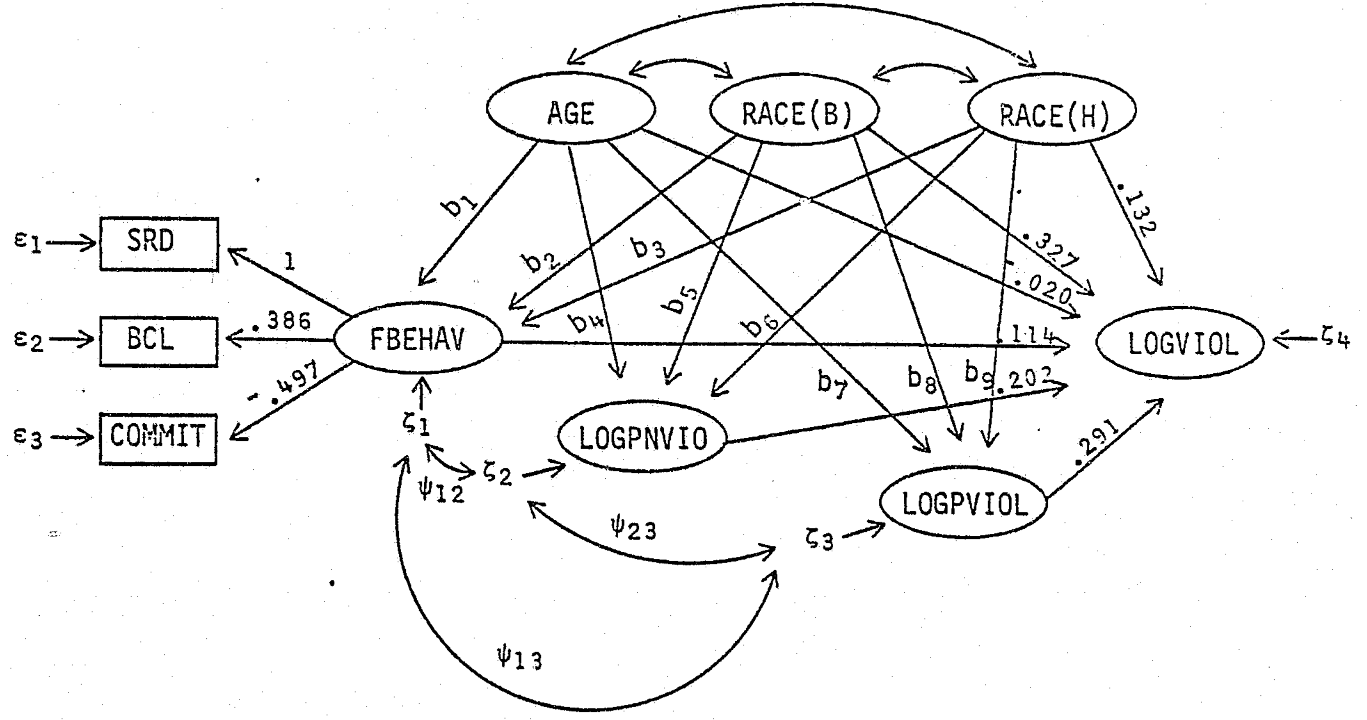


TABLE IV-4A

Direct Effect Estimates and Test Statistics:
LISREL Preston Model

Variable:	Coefficient	t-ratio
Behavioral Orientation............	.114	2.73
Nonviolent priors (logged)........	.202	4.41
Violent priors (logged)...........	.291	4.81
Age...............................	-.020	0.61
Race (Black)......................	.327	6.45
Race (Hispanic)...................	.132	1.92
Treatment (est. mean difference)...	-.074	1.67

Test statistics:

Overall goodness of fit: χ^2 = 36.89 (df=28) p = .12

Test treatment effect = 0: χ^2 = 39.62 (df=29)

χ^2 test = 2.73 (df=1) ns (accepted) error term for BCL and both Age and Hispanic ethnicity, as shown in Table IV-4B (these correlations were omitted from the figure for simplicity). Once they were taken into account, the correlations calculated from the LISREL estimates showed that the bivariate relationships between these demographics and the estimated factor were considerably higher than when the factor was constructed a priori. For example, the correlations between the factor created through adding standard scores and Hispanic was .06 for treatment cases and .04 for controls (for the total sample--Table IV-1--the correlation was .05). The correlations calculated from the LISREL estimates showed these correlations to be ≈ .23 for the treatment group and ≈ .30 for the control group. The factor estimated by LISREL, in other words, was better able to account for the relationships between these two demographic variables and outcome. Consistent with the increased effect of the factor on outcome and the decreased effect of Hispanic ethnicity (the treatment group scored higher on the factor and had fewer Hispanics), the estimated treatment effect in this model was higher than that found for ANCOVA.

As discussed in Chapter II, the inclusion of correlations between the residuals of the factor indicators and the demographic variables serves to "tailor" the factor to take into account differences in the subgroups found in the population.[5] Thus, the negative covariances (and their implied

[5] These correlations, in the form of covariances, were not constrained to be equal across groups, even though they are technically part of the measurement model for the factor. They were allowed to vary in order to compensate for differences between the groups in the variances of the variables themselves. The covariances shown in Table IV-4B are very close in value between the two groups, however, suggesting that these additional equality constraints could have been made without reducing the fit of the model to the data.

TABLE IV-4B

Other LISREL Estimates for Preston Model

	Treatment (t-ratio)	Control (t-ratio)
Variances, covariances:		
SRD..........................(θ_{11})	4.353 (6.34)	4.060 (6.73)
BCL..........................(θ_{22})	10.714 (13.89)	12.591 (16.18)
COMMIT.......................(θ_{33})	7.269 (13.48)	6.292 (15.39)
AGE,BCL......................	-.309 (2.36)	.179 (1.55)
RACE(Black),BCL..............	.325 (4.43)	.223 (3.27)
RACE(Hispanic),BCL...........	-.189 (2.83)	-.268 (4.13)
FBEHAV.......................(ψ_{11})	1.651 (2.59)	1.631 (2.83)
LOGPNVIO.....................(ψ_{22})	.234 (14.34)	.271 (16.60)
LOGPVIOL.....................(ψ_{33})	.141 (14.33)	.144 (16.61)
LOGVIOL......................(ψ_{44})	.389 (13.14)	.341 (14.82)
FBEHAV,LOGPNVIO..............(ψ_{12})	.013 (0.24)	.117 (2.35)
FBEHAV,LOGPVIOL..............(ψ_{13})	.093 (2.20)	.005 (0.14)
LOGPNVIO,LOGPVIOL............(ψ_{23})	-.068 (7.01)	-.059 (6.70)
Direct effects:		
AGE on FBEHAV................(b_1)	-.385 (2.75)	-.321 (2.58)
AGE on LOGPNVIO..............(b_4)	.168 (5.68)	.194 (6.70)
AGE on LOGPVIOL..............(b_7)	.057 (2.47)	.235 (1.12)
RACE(Blk) on FBEHAV..........(b_2)	.432 (1.67)	.293 (1.33)
RACE(Blk) on LOGPNVIO........(b_5)	-.192 (3.52)	.062 (1.20)
RACE(Blk) on LOGPVIOL........(b_8)	.383 (6.45)	.232 (6.18)
RACE(Hisp) on FBEHAV.........(b_3)	.995 (3.25)	1.057 (4.22)
RACE(Hisp) on LOGPNVIO.......(b_6)	-.127 (1.97)	.060 (1.04)
RACE(Hisp) on LOGPVIOL.......(b_9)	.214 (4.25)	.083 (1.98)
Differences in means:		
FBEHAV.......................	.280 (1.87)	0
LOGPNVIO.....................	-.111 (3.34)	0
LOGPVIOL.....................	.054 (2.15)	0
AGE..........................	-.098 (1.88)	0
RACE(Black)..................	.031 (1.04)	0
RACE(Hispanic)...............	-.016 (0.61)	0

correlations) between BCL and Hispanic (Table IV-4B) suggest that the Behavioral Orientation factor that best explains variations in outcome among Hispanics is different in both groups than that which best explains this variation for Whites: for Hispanics, less emphasis is placed on BCL scores. A similar, but opposite, interpretation can be made of the positive correlation between this scale and Black ethnicity: BCL scores are more indicative of this general, causally-relevant orientation for Blacks than for Whites. The correlations found for Age are not so easily interpreted, since they differ in sign for the two groups. However, these relationships are not strong, being marginally significant in one group and nonsignificant in the other; since one relationship in twenty can be expected to be significant simply due to random variation (at the .05 level), we may venture that these covariances are spurious; their inclusion in the model merely takes this random "noise" into account.

With the effects of the demographic variables in the model on the predispositional variables allowed to differ between groups, the LISREL results provide us with information regarding differences between the groups that would otherwise be apparent only by comparing the correlation matrices. For example, in Table IV-4B we see that the Black and Hispanic members of the treatment group had fewer prior nonviolent offenses than Whites, while this was not true in the control group. These kinds of differences may suggest ways in which one's sampling design was inadequate (in quasi-experimental design situations) or may suggest that a breakdown occurred in the random assignment procedure. In the present case, these relationships may be the result of having placed slightly more minorities with high I-levels (maturity levels) in the experimental group than was called for by the random assignment procedure. As mentioned in Chapter II, the inclusion

of the relationships between the demographic variables and the predispositional ones as "causal" does not necessarily imply that there is, in fact, a causal relationship between these characteristics and crime. The relationships could just as easily (and perhaps more justifiably) have been included in the form of unanalyzed correlations. The remaining parameters in the model would have been estimated the same either way. We chose this manner of including them in order to demonstrate the flexibility of LISREL.

The reduced form of this model simply excluded the direct effect of Hispanic ethnicity and Age on outcome. All other parameters in the model were the same as for the full model. The results, shown in Tables IV-5A and IV-5B, show an interesting effect of this change. Although the t-values for the direct effects of Age and Hispanic ethnicity were not negligible in the full model, the removal of these direct effects from the model (by specifying that these parameters were zero) resulted in an increase in χ^2 of only 2.3, with two degrees of freedom. This difference is far from significant, indicating that the removal of these direct effects did not significantly reduce the fit of the model to the data. As a result of these changes, the factor was estimated somewhat differently, (the factor loading for BCL increased), and the factor was found to have a stronger effect on outcome. This newly-constituted factor differentiated more clearly between clients and controls as well, with a net result of an increase in the treatment effect estimate. As shown in Table IV-5A, the estimated treatment effect for this model was -.102, which was almost twice that obtained with ANCOVA using the same logged offense data. The t-ratio and χ^2 test for this estimate both showed it to be statistically significant at the .05 level. The confidence to be placed in this result is open to question, due

TABLE IV-5A

Direct Effect Estimates and Test Statistics:
Reduced Form of LISREL Preston Model

Variable:	Coefficient	t-ratio
Behavioral Orientation............	.193	5.20
Nonviolent priors (logged).........	.176	3.92
Violent priors (logged)...........	.274	4.35
Race (Black)......................	.289	5.82
Treatment (est. mean difference)...	-.102	2.15

Test statistics:

Overall goodness of fit: χ^2 = 39.19 (df=30) p = .12

Test treatment effect = 0: χ^2 = 43.86 (df=31)

χ^2 test = 4.67 (df=1) p < .05 (rejected)

TABLE IV-5B

Other LISREL Estimates for Reduced Form of Preston Model

Factor loadings:	Treatment	(t-ratio)	Control	(t-ratio)
SRD.........................(λ_1)	1.000		constrained equal across groups	
BCL.........................(λ_2)	.553	(3.07)		
COMMIT......................(λ_3)	-.536	(4.29)		
Variances, covariances:				
SRD.........................(θ_{11})	4.892	(9.80)	4.604	(11.11)
BCL.........................(θ_{22})	10.565	(13.61)	12.462	(15.89)
COMMIT......................(θ_{33})	7.340	(13.65)	6.373	(15.67)
AGE,BCL....................	-.278	(2.08)	.214	(1.82)
RACE(Black),BCL............	.332	(4.44)	.226	(3.27)
RACE(Hispanic),BCL.........	-2.07	(3.01)	-.291	(4.36)
FBEHAV......................(ψ_{11})	1.107	(2.95)	1.086	(3.32)
LOGPNVIO....................(ψ_{22})	.234	(14.33)	.271	(16.60)
LOGPVIOL....................(ψ_{33})	.141	(14.33)	.144	(16.61)
LOGVIOL.....................(ψ_{44})	.372	(12.30)	.325	(13.85)
FBEHAV,LOGPNVIO.............(ψ_{12})	.022	(0.45)	.113	(2.44)
FBEHAV,LOGPVIOL.............(ψ_{13})	.079	(2.06)	.000	(0.01)
LOGPNVIO,LOGPVIOL...........(ψ_{23})	-.068	(7.08)	-.059	(6.70)
Direct effects:				
AGE on FBEHAV...............(b_1)	-.342	(2.72)	-.312	(2.81)
AGE on LOGPNVIO.............(b_4)	.168	(5.66)	.194	(6.71)
AGE on LOGPVIOL.............(b_7)	.056	(2.45)	.239	(1.14)
RACE(Blk) on FBEHAV.........(b_2)	.401	(1.68)	.342	(1.65)
RACE(Blk) on LOGPNVIO.......(b_5)	-.192	(3.51)	.062	(1.21)
RACE(Blk) on LOGPVIOL.......(b_8)	.384	(9.02)	.232	(6.18)
RACE(Hisp) on FBEHAV........(b_3)	1.046	(3.78)	1.104	(4.83)
RACE(Hisp) on LOGPNVIO......(b_6)	-.126	(1.95)	.059	(1.03)
RACE(Hisp) on LOGPVIOL......(b_9)	.213	(4.24)	.082	(1.95)
Differences in means:				
FBEHAV.....................	.292	(2.03)	0	
LOGPNVIO...................	-.110	(3.33)	0	
LOGPVIOL...................	.053	(2.14)	0	
AGE........................	-.099	(1.89)	0	
RACE(Black)................	.033	(1.09)	0	
RACE(Hispanic).............	-.017	(0.65)	0	

to the sensitivity of LISREL to departures from normality in the distributions of the variables. The results do suggest, however, that a) with the factor estimated in this way and b) with this model hypothesized to account for all important differences between the groups related to outcome scores, the program was shown to be more effective than it was shown to be using ANCOVA.

In general, the LISREL results for this sample showed this method to result in a greater adjustment for differences on the three scales used to indicate Behavioral Orientation. This result is consistent with that found earlier for the YSB sample. In this case, the increase in predictive power for the factor was largely the result of its being allowed to differ in constitution for different subgroups of the larger sample. With these variations allowed, pretest information for a sample containing subgroups differing with respect to values and culture can, theoretically, be taken more fully into account in adjusting for differences between groups. As an alternative to including the kinds of residual correlations used here, one may wish to consider analyzing the different ethnic groups separately, but this would ordinarily require very large samples. The extent to which the factor adjusted the treatment effect estimate depended on whether it was modeled as accounting for the apparent differences in outcome related to Hispanic ethnicity and Age or not. The differences in χ^2 values for the two models suggested that, with some modification, the factor could account for these relationships.

The Combined Methods approach to these data mainly involved estimating the additive equation using the tobit model. Loglinear analysis of these data failed to find any significant interaction effects involving outcome. The model containing only the main effects on outcome produced a value of

99.0 for G^2 (analogous in use and interpretation to chi square), with 128 degrees of freedom; the probability value was above .50, indicating a good fit to the data. Consequently, the tobit analysis included only the main effects of the variables (raw offense measures were used). The tobit estimates, along with the ANCOVA estimates for the same set of variables, are shown in Table IV-6. These estimates, once again, are very similar with respect to direction of effect and t-ratios. Although the t-ratio for the treatment effect estimate was slightly higher for the tobit model, it was actually slightly lower than that obtained with ANCOVA using logged offense data. For these data, then, which are probably typical of the kinds of data that might be used in criminal justice evaluation, the use of tobit models does not result in estimates that are noticeably different in terms of their estimated significances than ANCOVA.

In general, the conclusions to be drawn from these comparative analyses are the same as for the YSB analyses. ANCOVA, LISREL (full model), and the tobit model all provided virtually the same estimate of the significance of the adjusted difference in outcome for the treatment and control cases. In all cases, the treatment group was found to have somewhat lower numbers of subsequent violent offenses, controlling for preexisting differences, but none of these analyses showed the difference to be statistically significant. The various attempts to correct for skewness in the offense variables tended to result in somewhat lower standard errors for the treatment effect estimate, relative to the estimate itself, but none showed results very different from ANCOVA using simply the raw data. Correcting for measurement error through the use of LISREL, on the other hand, did result in a greater adjustment for differences in pretest scores. Given the admitted sensitivity of LISREL methods to departures from multivariate

TABLE IV-6

Preston ANCOVA and Tobit Estimates: Violent Subsequents
(t-ratios in parentheses)

Variable	ANCOVA (t-ratio)	Tobit (t-ratio)
Behavioral Orientation	.146 (4.43)	.222 (4.95)
Prior Violent Offenses	.466 (5.18)	.603 (5.00)
Prior Nonviolent Offenses	.146 (4.02)	.203 (4.14)
Age	-.093 (1.15)	-.159 (1.44)
Race (Black)	1.091 (7.43)	1.439 (7.21)
Race (Hispanic)	.560 (3.48)	.819 (3.73)
Treatment	-.119 (.98)	-.213 (1.27)
Constant	2.596	2.936

normality in the distributions of the variables in the analysis, the apparent statistical significance of the estimated treatment effect, however, is probably suspect. Without being able to control simultaneously for departures from normality and measurement error in the pretests, firm conclusions regarding the confidence to be placed in the LISREL results cannot be made.

Categorical Outcome Variables

Due to the relatively large sample size and the fact that over two-thirds of the total sample had at least one subsequent arrest for a violent offense, it was possible to employ three categories of the outcome variable in these analyses. As discussed earlier, loglinear analyses failed to find any significant interactions involving subsequent violent offenses. Logit analyses, then, were performed using a simple, additive model to predict two categories of outcome: the (log) odds of being in the "no violent subsequents" category (as opposed to the other two) and the (log) odds of being in the "three or more violent subsequents" category. The treatment effects estimated for these two analyses have implications for treatment effects relative to the category "one or two violent subsequents." These results were compared to ANCOVA results predicting the same outcomes. The treatment effect estimated with ANCOVA refers to the relative probability of being in these categories, in the form of estimated differences in the proportions of treatment and control groups falling into that category. The proportions of the total sample falling into the NOVIOL, FEWVIOL and MANYVIOL categories were .30, .38, and .32, respectively. Again, since these proportions all fall between .20 and .80 and since the squared multiple correlation for these equations can be expected to be small, the ANCOVA estimates of the treatment effects relative to these outcomes are likely to be pretty good.

The results for the ANCOVA and logit analyses for the two separate outcome variables are shown in Table IV-7. As expected, the ANCOVA estimates appear to be very good, with the ratios of estimates to their standard errors (t-ratios) being almost identical in most cases. In particular, the t-ratios for the respective estimates of the treatment effect are within .02 of one another.[6] Both methods show the treatment cases to be somewhat more likely than controls to have no subsequent violent arrests but equally as likely to have three or more such arrests. In both analyses, the adjustment for preexisting differences was small, but in a direction favorable to the treatment group. With ANCOVA, the differences in proportions in the NOVIOL group increased from a simple difference of .045 (more treatment cases) to an estimated difference of .056. Similarly, the difference in the proportions of treatment cases to that of controls in the MANYVIOL category fell from .016 to .006. These differences are not large, but large differences could not really be expected where the groups are fairly similar and the predictability of the outcome variable is low. No such simple interpretation of the treatment-effect coefficients for the logit analyses are possible.

These results suggest that the treatment provided during the original study had some effect on those who would otherwise have committed a small number of violent acts (enough to be arrested one or two times), dropping them into the "no violent subsequents" category. Without knowing the nature of acts that would have been committed (our definition of violence included

[6] Separate analyses, using more than three subsequent violent arrests (19% of the sample) and more than four subsequent violent arrests (10% of the sample) as the dependent variable showed a similar consistency between ANCOVA and logit results. The differences between the results were slightly larger, but the methods agreed as to the significances of the effects of the variables in the analyses.

TABLE IV-7

Preston ANCOVA and Logit Estimates
(t-values in parentheses)

	No Violent Subsequents		Over 2 Violent Subsequents	
Variable	ANCOVA (t-ratio)	LOGIT (t-ratio)	ANCOVA (t-ratio)	LOGIT (t-ratio)
Behavioral Orientation	-.038 (4.86)	-.203 (4.70)	.025 (3.24)	.129 (3.23)
Prior Violent Offenses	-.071 (3.36)	-.491 (3.54)	.105 (4.94)	.484 (4.53)
Prior Non-Violent Offenses	-.024 (2.94)	-.166 (3.72)	.027 (3.33)	.138 (3.29)
Race (Black)	-.167 (4.89)	-.428 (4.55)	.277 (6.59)	.561 (6.42)
Race (Hispanic)	-.118 (3.12)	-.276 (2.80)	.162 (4.26)	.428 (4.39)
Treatment	.056 (1.93)	.144 (1.91)	.006 (.21)	.014 (.19)
Constant	.428		.113	
	R^2=.088		R^2=.116	

misdemeanor assault along with the more serious violent offenses), the importance of such a difference is unclear. The fact that the estimated difference almost reached statistical significance would suggest that these issues might bear further investigation, perhaps through comparing the numbers with specific kinds of subsequent violent arrests.

CHAPTER V

Discussion

In this study, we investigated the nature and importance of some of the problems theoretically associated with the use of analysis of covariance (ANCOVA), as applied to criminal justice evaluation data. The research was not designed to test the validity of the criticisms of ANCOVA, but rather to assess their importance through comparing ANCOVA results for criminal justice evaluation data to results obtained with analytic methods designed to overcome its problems. The two data sets used for this purpose were considered typical of the kinds that are commonly used in evaluations of criminal justice programs. One data set was artificial, constructed for the present purposes from a larger set of data generated during an evaluation of California Youth Service Bureaus (YSB sample). The second data set came from an experimental study of an institutional program within the California Department of the Youth Authority (Preston sample). Outcome data for the second sample came from a recent long-range followup of these cases. For both data sets, outcome was a measure of subsequent official delinquency (YSB sample) or crime (Preston sample) and background variables included measures of prior police contacts or arrests, demographic information, and scales measuring pre-existing behavioral traits.

The importance of measurement error in these pretest scales was assessed by comparing the results obtained with ANCOVA using the scales with those obtained using a factor score in place of the scales, and with analyses using LISREL, which constructs a factor from these scales in the context of an overall causal model predicting outcome. Other analyses focused on the importance of violations of the distributional assumptions of ANCOVA, which arise as a result of using outcome variables with limited, skewed distributions.

For these analyses, interaction effects were investigated through loglinear models, which are free of distributional assumptions. The importance of the problems inherent in using these outcome measures in the estimation of treatment effects was examined by comparing ANCOVA results using raw outcome variables with ANCOVA using logged outcome variables and to analyses using logit models (for dichotomous outcome variables) and tobit models (for continuous outcome variables). The logit and tobit models were designed specifically for use in predicting variables with distributions like those found for offense-type measures.

The results of the various analyses with the two data sets showed that for these kinds of data, the use of alternative analytic techniques provided only minimal improvements in the estimation of overall treatment effects. The loglinear analysis of the YSB data set did help to identify certain problems in the data that went undetected by ANCOVA, LISREL and the tobit analysis; these problems increased the skewness of the outcome variable and led to the appearance of a negative treatment effect. Still, with these data, the three analytic techniques would have arrived at the same general conclusion regarding the efficacy of YSBs. For the Preston sample, all of the analyses produced roughly the same results: a slightly lower number of subsequent violent crimes among those in the experimental program. By comparing outcomes defined both as a continuous variable (number of violent crimes) and as dichotomous variables (no violent subsequents and many violent subsequents), we learned that the treatment effect for the Preston sample was primarily in terms of a reduction in the numbers who subsequently committed only a few violent crimes. However, the ANCOVA results for these two kinds of outcome variables were essentially the same as those obtained with the other analytic techniques. Thus, we found that

for these data sets, the results obtained with ANCOVA were adequate for determining the effectiveness of the two programs.

This general conclusion does not mean that the alternative strategies provided no improvement over the results obtained with ANCOVA, but only that the improvement was not great enough to suggest that the ANCOVA results would have led to misleading conclusions concerning the two programs. In general, the methods applied here did lead to slightly different estimates of treatment effects and/or of the statistical significance of those estimates. These differences, moreover, were in the directions that would be expected from methods which overcome, to some extent, the problems with ANCOVA. The LISREL analyses, for example, showed that by controlling for measurement error in the pretest scales, the effects of the variable hypothesized to underlie these scales is taken into account to a greater extent. In each case, the LISREL adjustment for these pre-existing differences was greater than the adjustment made by ANCOVA. The fact that these increased adjustments made only a slight difference in the overall estimate of the treatment effect simply suggests that when predicting something as complex (and difficult to predict) as crime and delinquency, even the best adjustment for these kinds of variables may not make a great deal of difference. Similarly, the results of the tobit analyses for the YSB sample disagreed slightly with the ANCOVA results; tobit found that the treatment effect was not statistically significant, a result which was expected given the extreme skewness of the outcome variable. Still, the ANCOVA results showed only a marginally significant effect, which under the circumstances could not have led to any firm conclusions about YSBs anyway. In short, the results suggests that there are problems with the use of ANCOVA with these data and that they can be corrected to some degree, but that these problems may not be as serious as some might claim (especially if the researcher is aware of them and interprets his or her findings accordingly).

As discussed throughout the report, there were some advantages to employing the alternative analytic strategies that were unrelated to the specific estimation of treatment effects. For example, the use of loglinear analysis called into question the findings of ANCOVA and LISREL regarding apparent interaction effects in the YSB sample. These effects, as it turned out were spurious, resulting from the extreme skewness of the outcome variable. Although in this case the skewness was caused partly by errors in the data, it is not inconceivable that an outcome variable such as ours could be skewed to that degree and still be correct. ANCOVA, LISREL, and to some extent tobit, were all affected by those few cases at the extreme, attributing effects to the entire sample that were mostly true only for those cases. Loglinear models, in which all variables were collapsed into categories, was not affected by these extreme cases and found neither interactions nor negative treatment effects. By employing a method such as loglinear analysis in conjunction with ANCOVA, the researcher may avoid the pitfalls of relying too heavily on the results of ANCOVA in these situations. Loglinear analysis, however, generally requires relatively large data sets and/or only a few categories of each variable. For the purposes mentioned above, simple bivariate cross-tabulations of categorized outcome scores and categorized predictors might serve just as well to identify these spurious effects. Alternatively, one may simply remove the cases with the highest outcome scores and use ANCOVA on the smaller sample to determine if the same effects are found. These kinds of alternatives would not provide the same amount of information as the loglinear analysis, which assesses effects while taking other effects into account, but may still be helpful in discovering potential problems in the data.

The LISREL analysis also provided information that was unavailable from the results of the other methods. The major advantage to using LISREL, in this regard, is that potentially important or interesting insights may be gained about the causes of outcome. LISREL is, after all, a method designed for testing particular causal hypotheses and overall causal models. In our analyses of the Preston data, as an example, we found that the kind of behavioral pattern we referred to as Behavioral Orientation differed in its indicators among ethnic groups. We also found that differences along this behavioral dimension could explain the differences in outcome between Whites and Hispanics. Such a finding leaves open the issue of the cause of the differences between the ethnic groups on this factor, but it does suggest that the differences in outcome between these two groups may be explained in terms of differences that exist at the onset of adulthood (that differences in later experiences are not very important). The higher numbers of arrests for Blacks could not be explained in this way, and this fact suggests either that measures of the important differences between Blacks and Whites were not included in the analysis or, perhaps, that the experiences and opportunities of these groups during later years differed enough to result in differential crime rates. These kinds of possibilities present themselves readily from LISREL analyses. The use of LISREL may also bring to light certain problems in the data, as we mentioned in the discussion of the two-factor YSB model, or help to identify differences between the groups used in a particular study, as we found for Preston.

Along with the advantages of using LISREL (and partially because of them), there are some important disadvantages to its use. First, the causal-modeling approach to analyzing data is conceptually complex, with its own nomenclature and set of concerns. The mastery of the method (and the computer program) may require more time and effort than it is worth for those whose main concern is determining whether a particular treatment program is effective. Our use of the LISREL program (using two separate groups and including the means of the variables in the analysis) was the most complicated, but even the most basic uses of the program require a familiarity with the method and the terms. A large portion of the time spent on the present research was spent in learning and mastering the technique, even though we were familiar with the concepts and issues involved.

Second, the LISREL computer program itself, and its use, are fairly expensive. Because it employs an iterative process for obtaining estimates of the parameters in the model, it may require a large amount of time to estimate complex models. This problem is compounded if the user misspecifies a model in certain ways (so that estimates are difficult to obtain) or if there are inconsistencies in the data. Large inconsistencies (such as strong negative correlations between indicators of hypothetically positively-correlated factors) may cause the program to cancel the job, but small inconsistencies may merely cause it to spend a large amount of time trying to find estimates that will fit the inconsistent data. For example, it is not unlikely that a demographic variable may be unrelated to a factor being estimated in one's model, this unrelatedness being suggested by essentially zero correlations between the demographic variable and each of the indicators. These zero correlations will not

actually be zero, however, but will vary around zero (some small and positive, others small and negative). Even if the correlation between the factor and the demographic variable is specified to be zero in the model, the program may have difficulty finding an estimate of the factor that is consistent with the fact that the indicators are correlated differentially with another variable in the analysis. The problem is compounded even further when more than one group is being analyzed or when the sample size is small; under these conditions, it is more likely that a few of these ("zero") correlations will not be small, suggesting to the program that the model is seriously flawed. In other words, chance correlations in the data may make it difficult, expensive, or even impossible to estimate a model that may well be true for the population.

Thus, the LISREL method may not be particularly appropriate as a tool for analyzing criminal justice evaluation data, even though it can provide a richness and flexibility not found in ANCOVA. Our results show that the benefits of its use were not sufficient to call into question the ability of ANCOVA to adequately estimate treatment effects with these kinds of data. Those who would take advantage of its potential for providing a greater understanding of the data would be advised to be careful of small samples and low correlations among some of the variables. Features of the latest version of the program will enable the user to minimize the cost associated with problems that may arise as a result of the data (time limits can be set or only initial estimates obtained), but the model-building and model-testing processes may still lead to a good deal of frustration.

The logit and tobit models are not so difficult practically or conceptually, and individuals with an understanding of ANCOVA should find no particular problem in using them. Logit programs are readily available in such commonly used statistical software packages as BMDP and Statistical Analysis System (SAS). These models appeared to provide no substantial improvement over what can be obtained with ANCOVA, however, and the results are not as easily interpreted. For most purposes, then, their use is probably unnecessary. The tobit model did seem to provide some improvement in the assessment of the significance of treatment effects, but the difference was not great. The main problem with this method is availability. The tobit model is not widely used and has not, as far as we know, been incorporated in the common statistical packages. Since SAS is constantly including new algorithms in its supplementary library, however, it may become available in the future. Its use bears more investigation with a wider variety of data conditions.

References

Blalock, H. (ed.) Measurement in the social sciences: theories and strategies. Chicago: Aldine, 1974.

Blumstein, A. & Cohen, J. Control of selection effects in the evaluation of social problems. Evaluation Quarterly, November 1979, 3(4), 583-607.

Burt, R. Interpretational confounding of unobserved variables in structural equation models. Sociological Methods and Research, August 1976, 5(1) 3-53.

Campbell, D. & Stanley, J. Experimental and quasi-experimental designs for research. Chicago: Rand McNally, 1963.

Carmines, E. & McIver, J. Analyzing models with unobserved variables: analysis of covariance structures, in G. Bohrnstedt & E. Borgatta (eds.) Social Measurement: current issues. Beverly Hills, CA: Sage Publications, 1981.

Cohen, J., & Cohen, P. Applied multiple regression/correlational analysis for the behavioral sciences. Hillsdale N.J.: Lawrence Erlbaum Associates, 1975.

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

Elliott, D., Ageton, S., Hunter, M. & Knowles, B. Research handbook for community planning and feedback instruments. Vol. I. Boulder, Colorado: Behavioral Research and Evaluation Corporation, 1976.

Fienberg, S. The analysis of cross-classified categorical data. Cambridge, Mass.: The MIT Press, 1978.

Goldberger, A. Econometric Theory. New York: Wiley, 1964.

Goodman, J., Jr. Working paper: is ordinary least squares estimation with a dichotomous dependent variable really that bad? Washington, D.C.: The Urban Institute, 1976.

Green, W. On the asymptotic bias of the ordinary least squares estimator of the tobit model. Econometrica, 1981, 49, 505-514.

Green, W. Estimation of limited dependent variable models by ordinary least squares and the method of moments. Unpublished ms. Cornell University, 1982.

Haapanen, R., & Jesness, C. Early identification of the chronic offender. Sacramento: California Youth Authority, October, 1981.

Haapanen, R. & Rudisill, D. Youth service bureaus: an evaluation of nine California youth service bureaus. Sacramento: California Youth Authority, 1980.

Hanushek, E. & Jackson, J. Statistical methods for social scientists. New York: Academic Press, 1977.

Hirschi, T. Causes of delinquency. Berkeley: University of California Press, 1969.

Jesness, C. The Preston typology study: final report. Sacramento: California Youth Authority, 1969.

Jesness, C. The Jesness behavior checklist. Palo Alto: Consulting Psychologists Press, Inc., 1971a.

Jesness, C. The Preston typology study: an experiment with differential treatment in an institution. Journal of Research in Crime and Delinquency. 1971b, 8, 38-52.

Jesness, C. Classifying juvenile offenders: the sequential I-level classification manual. Palo Alto: Consulting Psychologists Press, 1974.

Jesness, C., Allison, T., McCormick, P., Wedge, R. & Young, M. Cooperative behavior demonstration project, Sacramento: California Youth Authority, 1975.

Jöreskog, K. A general method for estimating a linear structural equation system, in A. S. Goldberger and O. D. Duncan (eds.) Structural equation models in the social sciences. New York: Seminar Press, 1973.

Jöreskog, K. and Sörbom, D. Advances in factor analysis and structural equation models. Cambridge, Mass.: Abt Books, 1979.

Jöreskog, K. and Sörbom, D. LISREL V: analysis of linear structural relationships by maximum likelihood and least squares methods. Chicago: International Educational Services, 1981.

Kenny, D. Correlation and causality. New York: Wiley, 1979.

Knoke, D. & Burke, P. Log-linear models. Beverly Hills, Sage Publications, 1980.

Linn, R. & Werts, C. Analysis implications of the choice of a structural model in the nonequivalent control group design. Psychological Bulletin, 1977, 84(2), 229-234.

Lipton, D., Martinson, R., & Wilks, J. The effectiveness of correctional treatment: a survey of treatment evaluation studies. New York: Praeger, 1975.

Long, J. Estimation and hypothesis testing in linear models containing measurement error: a review of Jöreskog's model for the analysis of covariance structures. Sociological Methods and Research. November 1976, 5(2), 157-206.

Maruyama, G. & McGarvey, B. Evaluating causal models: an application of maximum-likelihood analysis of structural equations. Psychological Bulletin, 1980, 87(3), 502-512.

Overall, J. & Woodward, J. Nonrandom assignment and the analysis of covariance. Psychological Bulletin, 1977, 84(3), 588-594.

Palmer, J. & Carlson, P. Problems with the use of regression analysis in prediction studies. Journal of Research in Crime and Delinquency, January 1976, 13, 64-81.

Palmer, T. Correctional intervention and research. Lexington, Mass.: D. C. Heath & Company, 1978.

Palmer, T., Bohnstedt, M., & Lewis, R. The evaluation of juvenile diversion programs. Sacramento: California Youth Authority, 1978.

Reichardt, C. The statistical analysis of data from nonequivalent group designs, in T. Cook and D. Campbell (eds.) Quasi-experimentation: design and analysis issues for field settings. Chicago: Rand McNally, 1979.

Rindskopf, D. Structural equation models in analysis of nonexperimental data, in R. Boruch, P. Wortman, D. Cordray, & associates (eds.) Reanalyzing program evaluations. San Francisco: Jossey-Bass, 1981.

Romig, D. Justice for our children: an examination of juvenile delinquency rehabilitation programs. Lexington, Mass.: D. C. Heath & Company, 1978.

Sörbom, D. An alternative to the methodology for analysis of covariance. Psychometrika, September 1978, 43(3), 381-396.

Wiatrowski, M., Griswold, D., & Roberts, M. Social control theory and delinquency. American Sociological Review, October 1981, 46, 525-541.

APPENDIX A - YSB Scales

I. Self-Report Delinquency (SRD)

(1 = never, 2 = one time, 3 = 2 to 3 times, 4 = 4 to 5 times, 5 = more than 5 times)

During the past year:

1. I took part in a fight where our group fought a different group.

2. Not counting fights you may have had with a brother or sister, have you beaten up anyone?

3. I damaged or messed up something in a school or some other building.

4. Have you gotten something by telling a person something bad would happen to them if you did not get what you wanted?

5. I have taken some part of a car or some gasoline.

6. Have you taken something not belonging to you worth between $2 and $50?

7. Have you taken something not belonging to you worth less than $2?

8. I have taken a car for a ride without the owner's permission (even if returned).

9. How many times have you had some beer, wine, or liquor without your parents' permission?

10. I have used marijuana.

11. I have used drugs other than marijuana.

12. Have you run away from home?

13. How many times did you skip school without a real excuse?

14. Have you taken things of large value (over $50)?

15. I have bought or gotten something that was stolen by someone else.

II. Self-Report Obtrusiveness (BCL)

(1 = almost never, 2 = not often, 3 = sometimes, 4 = fairly often, 5 = very often)

1. I interrupt others when they are talking or bother others who are busy.

2. I clown around, horseplay, or act up when I know I'm not supposed to.

3. I try to get others in trouble by getting them into fights or arguments or by talking about them.

4. I agitate or bother others by teasing, laughing, or making fun of them.

5. I get angry or upset when I am frustrated or don't get my way.

6. I pick on, push around, threaten, or bully others.

7. I like to tell others about things I've gotten away with, even some that were against the law.

8. I feel upset if I can't have what I want or do what I want right away.

9. I get loud and noisy at times or places when I probably shouldn't.

10. I tend to resist authority; I argue or don't go along with what people tell me to do.

III. Commitment to Social Values (COMMIT)

1. Most of the time I do not want to go to school.

   1 - agree
   2 - disagree

2. I am very happy when I am in school.

   1 - disagree
   2 - agree

3. I like school very much.

   1 - disagree
   2 - agree

4. I enjoy the work I do in class.

   1 - never
   2 - seldom
   3 - sometimes
   4 - often
   5 - always

5. How much schooling do you actually expect to get eventually?

   1 - some high school
   2 - high school graduation
   3 - on-the-job apprenticeship
   4 - trade or business school
   5 - some college or junior college
   6 - college graduation (4 years)

IV. Attachment to Other People (ATTACH)

1. How much do you care what your teachers think of you?

   1 - it doesn't matter to me at all
   2 - it matters very little
   3 - I care somewhat what they think
   4 - I care very much

2. Would you like to be the kind of person your best friends are?

   1 - not at all
   2 - in a few ways
   3 - in most ways

3. Do you respect your best friends' opinions about the important things in life?

   1 - not at all
   2 - a little
   3 - pretty much
   4 - completely

4. Do you share your thoughts and feelings with your parents?

   1 - never
   2 - sometimes
   3 - often

5. How likely are you to talk over problems with your parents?

   1 - not likely at all
   2 - somewhat likely
   3 - very likely

6. Are you interested in what your father thinks of you?

   1 - not at all
   2 - not much
   3 - somewhat
   4 - quite a lot

7. Are you interested in what your mother thinks of you?

   1 - not at all
   2 - not much
   3 - somewhat
   4 - quite a lot

8. I always like to hang around with the same bunch of friends.

   1 - false
   2 - true

9. I often feel lonesome and sad.

   1 - true
   2 - false

10. I would usually prefer to be alone with others.

    1 - true
    2 - false

11. I feel alone even when there are other people around me.

    1 - true
    2 - false

V. Positive Peer Association (PEERS)

1. Most of my friends do well in school.

   1 - false
   2 - true

2. Do your friends get in trouble in school?

   1 - yes, quite often
   2 - sometimes a few do
   3 - almost never
   4 - no, never

3. The kids in my group would think less of a person if he/she were to get in trouble with the law.

   1 - disagree
   2 - don't know
   3 - agree

4. The kids in my group sometimes like to have a little fun even if it means breaking the law.

   1 - agree
   2 - don't know
   3 - disagree

5. Have any of your close friends ever been picked up by the police?

   1 - four or more friends have
   2 - three friends have
   3 - two friends have
   4 - one friend has
   5 - no

VI. Belief in the Legitimacy of the Law (BELIEF)

1. Most police will try to help you.

   1 - false
   2 - true

2. If the police don't like you, they will try to get you for anything.

   1 - true
   2 - false

3. Police stick their noses into a lot of things that are none of their business.

   1 - true
   2 - false

4. If someone in your family gets into trouble it's better for you to stick together than to tell the police.

   1 - true
   2 - false

5. I think that boys fourteen years old are old enough to smoke.

   1 - true
   2 - false

6. Police usually treat you dirty.

   1 - true
   2 - false

7. It's fund to give the police a bad time.

   1 - true
   2 - false

8. I don't mind lying if I'm in trouble.

   1 - true
   2 - false

9. Stealing isn't so bad if it's from a rich person.

   1 - true
   2 - false

APPENDIX B

Distribution of Outcome Variable (SUBRATE) for YSB Sample

	No.	Percent	Cumulative Percent		No.	Percent	Cumulative Percent
0.0	261	65.4	65.4	.250	7	1.8	92.0
.038	1	.3	65.7	.267	1	.3	92.2
.056	4	1.0	66.7	.273	1	.3	92.5
.059	3	.8	67.4	.278	1	.3	92.7
.063	2	.5	67.9	.286	1	.3	93.0
.067	9	2.3	70.2	.294	1	.3	93.2
.071	6	1.5	71.7	.300	1	.3	93.5
.077	12	3.0	74.7	.313	1	.3	93.7
.083	11	2.8	77.4	.333	5	1.3	95.0
.091	9	2.3	79.7	.364	1	.3	95.2
.100	4	1.0	80.7	.375	4	1.0	96.2
.111	5	1.3	82.0	.400	1	.3	96.5
.125	2	.5	82.5	.412	1	.3	96.7
.133	1	.3	82.7	.417	1	.3	97.0
.143	6	1.5	84.2	.429	1	.3	97.2
.154	4	1.0	85.2	.455	1	.3	97.5
.167	3	.8	86.0	.467	1	.3	97.7
.176	1	.3	86.2	.500	2	.5	98.2
.182	3	.8	87.0	.545	1	.3	98.5
.188	2	.5	87.5	.600	1	.3	98.7
.200	5	1.3	88.7	.667	1	.3	99.0
.214	1	.3	89.0	.800	1	.3	99.2
.222	3	.8	89.7	1.000	1	.3	99.5
.231	2	.5	90.2	1.333	2	.5	100.0

APPENDIX C

Variance, Covariance Matrices: YSB Sample

CLIENTS	SRD	BCL	COMMIT	ATTACH	PEERS	BELIEF	AGE	SEX	LOGPRIOR	LOGSUBS	MEAN	STANDARD DEVIATION
SRD	67.7310										23.061	8.230
BCL	39.4651	59.4125									20.320	7.708
COMMIT	-13.1094	-11.1407	11.7808								17.780	3.432
ATTACH	-1.8473	-1.2830	0.2029	6.1036							21.756	2.470
PEERS	-10.9830	-7.2413	3.2034	-0.0219	8.2419						10.640	2.871
BELIEF	-8.3023	-4.9901	2.7825	0.5255	2.7698	5.1996					14.604	2.280
AGE	-0.3581	-1.6823	-0.6727	0.0038	-0.0789	-0.4249	2.6394				14.599	1.625
SEX	0.5766	0.1179	0.0098	-0.1586	-0.0805	-0.1027	-0.0534	0.2239			.665	.473
LOGPRIOR	0.4307	0.2534	-0.2060	-0.0425	-0.1562	-0.0632	-0.0292	0.0200	0.1462		.948	.382
LOGSUBS	0.3034	0.2163	-0.1503	-0.0786	-0.0631	-0.0268	-0.0118	0.0028	0.0180	0.0193	.070	.139

(n=197)

COMPARISONS	SRD	BCL	COMMIT	ATTACH	PEERS	BELIEF	AGE	SEX	LOGPRIOR	LOGSUBS	MEAN	STANDARD DEVIATION
SRD	58.9539										21.654	7.678
BCL	20.3603	39.6456									19.035	6.296
COMMIT	-7.4427	-3.0542	9.8938								18.129	3.145
ATTACH	-3.4556	-2.0783	2.0552	4.6281							21.535	2.151
PEERS	-11.4544	-4.6580	2.0519	1.6007	7.2618						10.752	2.695
BELIEF	-9.5738	-4.1755	3.2658	0.8842	2.8240	5.7096					14.752	2.390
AGE	1.5167	-1.8900	-0.0604	-0.0119	-0.2940	-0.7219	3.5153				14.698	1.875
SEX	0.1736	0.0498	-0.0425	-0.0497	0.0878	-0.0018	-0.0424	0.2057			.713	.453
LOGPRIOR	0.7640	0.0234	-0.2160	-0.1464	-0.1343	-0.1334	0.0923	0.0177	0.1854		.994	.430
LOGSUBS	0.1219	0.0025	-0.0174	-0.0278	-0.0403	-0.0396	0.0001	0.0044	0.0216	0.0087	.051	.093

(n=202)

APPENDIX D

Scales for Preston Sample

I. Self-Report Delinquency (SRD)

(1 = never, 2 = a few times, 3 = several times)

1. I took part in a fight where knives or other weapons were used.
2. I snatched someone's purse or wallet from them, but didn't hurt them.
3. I took part in a crime where weapons were used.
4. I threatened somebody with a weapon.
5. I took part in a planned robbery or burglary.

II. BCL Obtrusiveness: Observer Rating (BCL)

(1 = almost, 2 = not often, 3 = sometimes, 4 = fairly often, 5 = very often)

1. Interrupts others when they are talking, or bothers others who are busy.
2. Clowns around, horseplays, or acts up at the wrong time or place.
3. Tries to get others in trouble, by getting them into fights or arguments or by saying things about them.
4. Agitates or bothers others by teasing, laughing or making fun of them.
5. Gets angry and upset when he is frustrated.
6. Picks on, pushes around, threatens, or bullies others.
7. Likes to tell others about things he's gotten away with, even some things that were against the law.
8. Seems to be upset if he can't have what he wants or do what he wants right away.
9. Is loud and noisy at times or places when he shouldn't.
10. Resists authority; argues with or won't go along with what people tell him to do.

III. Commitment to Social Values

1. How do you fell most of the time when you are in school?

   1 - in low spirits
   2 - not very happy
   3 - pretty good
   4 - in very good spirits

2. Of all the teachers you have known, how many have you liked?

   1 - none
   2 - a few
   3 - about half
   4 - most
   5 - all

3. If you could be remembered at school for one of the six things below, which one would you like it to be.

   0 - (any of 5 responses)
   3 - honor student

4. If it were completely up to you, how far in school would you like to go?

   1 - get out as soon as possible
   2 - finish junior high
   3 - some high school
   4 - finish high school
   5 - business or tech school
   6 - four year college

APPENDIX E

Distribution of Outcome Variable (TOTVIOL) for Preston Sample

	No.	Percent	Cumulative Percent
0	285	29.6	29.6
1	205	21.3	50.9
2	164	17.0	68.0
3	128	13.3	81.3
4	84	8.7	90.0
5	44	4.6	94.6
6	22	2.3	96.9
7	17	1.8	98.7
8	4	.4	99.1
9	4	.4	99.5
10	1	.1	99.6
11	2	.2	99.8
12	1	.1	99.9
13	0	-	
14	0	-	
15	0	-	
16	1	.1	100.0