Journal of Abnormal Psychology
2001, Vol. 110, No. 1, 40–48

0021-843X/01/$5.00 DOI: 10.1037//0021-843X.110.1.40

Misunderstanding Analysis of Covariance

Gregory A. Miller
University of Illinois, Champaign

Jean P. Chapman
University of Wisconsin—Madison

Despite numerous technical treatments in many venues, analysis of covariance (ANCOVA) remains a widely misused approach to dealing with substantive group differences on potential covariates, particularly in psychopathology research. Published articles reach unfounded conclusions, and some statistics texts neglect the issue. The problem with ANCOVA in such cases is reviewed. In many cases, there is no means of achieving the superficially appealing goal of "correcting" or "controlling for" real group differences on a potential covariate. In hopes of curtailing misuse of ANCOVA and promoting appropriate use, a nontechnical discussion is provided, emphasizing a substantive confound rarely articulated in textbooks and other general presentations, to complement the mathematical critiques already available. Some alternatives are discussed for contexts in which ANCOVA is inappropriate or questionable.

In research comparing groups of participants, classical experimental design (Campbell & Stanley, 1963) relies, whenever possible, on random assignment of participants to groups. Observed differences between such groups, prior to experimental treatments, are due to chance rather than being meaningfully related to the group variable. In contrast, when preexisting groups are studied, observed pretreatment differences may reflect some meaningful, substantive differences that are attributable to group membership. It has been noted that, "Even in the absence of a true treatment effect, the outcome scores in the treatment groups are likely to differ substantially because of initial selection differences. As a result, selection differences are a threat to validity. . . . " (Reichardt & Bormann, 1994, p. 442).

In this article, we discuss why attempts to control statistically for such differences are, in general, inappropriate. For example, consider a data set consisting of age as a potential covariate, grade in school as the grouping variable, and basketball performance as the dependent variable. An analysis of covariance (ANCOVA) might be run in hopes of asking whether 3rd and 4th graders would differ in performance were they not different in age. This might seem to be a reasonable question, in that one could ask whether some maturational change at that age makes a nonlinear contribution to basketball ability. However, in fact it makes no sense to ask how 3rd graders would do if they were 4th graders. They are, inherently, not 4th graders, and ANCOVA cannot "control for" that fact. Age is so intimately associated with grade in school that removal of variance in basketball ability associated with age would remove considerable (perhaps nearly all) variance in basketball ability associated with grade. The results of the ANCOVA would be meaningless. As a complement to this problem of the covariate removing too much of the independent variable of interest, a problem can arise when preexisting groups differ systematically on more than the covariate. The covariate will leave those differences intact, thus biasing the estimate of the treatment effect (Reichardt & Bormann, 1994), which has been called *specification error.*

Nonrandom Group Assignment

When group membership is determined nonrandomly, there is typically no thorough basis for determining whether a given pretreatment difference reflects random error or a true group difference. This uncertainty complicates interpretation of apparent treatment effects because it is impossible to distinguish a main effect of treatment from an interaction between the effects of the treatment and the pretreatment difference, and from meaningful overlap (variance shared) between the treatment and the pretreatment characteristic. The confound due to an undetected interaction is widely understood—the pooled regression is inappropriate as a basis for "correcting" for the covariate. However, the other confound, due to undetected or misinterpreted overlap of pretreatment difference and grouping factor, is commonly neglected, although it was noted in the first modern treatment of ANCOVA (Cochran, 1957) and occasionally more recently (e.g., Porter & Raudenbush, 1987).

The problem of preexisting group differences arises very commonly in psychopathology research because random assignment to group (diagnostic category in the typical experimental design in psychopathology but called "treatment" in much of the statistical literature) is routinely infeasible and/or unethical. It is thus extremely tempting for psychopathologists to seek to use analytic methods in an attempt to avoid the interpretative problems that arise when groups differ pretreatment. It is unfortunate that, in the general case, no such analytic method is available, nor can one be

Gregory A. Miller, Departments of Psychology and Psychiatry and Beckman Institute, University of Illinois, Champaign; Jean P. Chapman, Department of Psychology, University of Wisconsin—Madison.

The writing of this article was supported by National Institute of Mental Health Grant MH39628. We thank Loren J. Chapman, Lawrence J. Hubert, Brandy G. Isaacks, Jack B. Nitschke, and E. Keolani Taitano for comments on an earlier draft.

Correspondence concerning this article should be addressed to Gregory A. Miller, Department of Psychology, University of Illinois, 603 East Daniel Street, Champaign, Illinois 61820, or to Jean P. Chapman, Department of Psychology, University of Wisconsin, 1202 West Johnson Street, Madison, Wisconsin 53706. Electronic mail may be sent to gamiller@uiuc.edu or jchapma4@facstaff.wisc.edu.

developed (a point that will be explained below). This logical fact has proven difficult for psychopathologists to accept, perhaps because of the burden that preexisting group differences place on interpretation of experimental results. Despite numerous technical treatments in the literature (e.g., Chapman & Chapman, 1973; Cochran, 1957; Elashoff, 1969; Fleiss & Tanur, 1973; Huitema, 1980; Jin, 1992; Lord, 1967, 1969; Maxwell & Delaney, 1990; Maxwell, Delaney, & Manheimer, 1985; Porter & Raudenbush, 1987; Wainer, 1991; Wildt & Ahtola, 1978) and more accessible statements (e.g., Neale & Oltmanns, 1980; Siddle & Turpin, 1980), together making an overwhelming case against inappropriate attempts to "control for" such group differences, they remain common in the research literature and, if anything, even more common in research grant applications.[1] Given the continuing popularity of inappropriate uses or interpretations of ANCOVA, the present article offers a relatively nontechnical critique, in hopes of helping to popularize the correct use of ANCOVA and helping researchers to avoid its more common abuses.

Understanding Analysis of Covariance

ANCOVA is part of the ANOVA (analysis of variance) tradition. ANCOVA was developed to improve the power of the test of the independent variable, not to "control" for anything. It is helpful here to place ANOVA and ANCOVA in the more general framework of multiple regression and correlation (MRC), understood within the general linear model. In fact, at least one popular ANOVA package, BMDP2V (Dixon, 1992), actually computes its analyses using multiple regression rather than using the equations typically presented in textbooks on ANOVA. Some sources place ANCOVA in the context of MRC, with the covariate in ANCOVA understood as a regression predictor entered before what are, in ANOVA, the main effects and interactions (e.g., Harris, Bisbee, & Evans, 1971; for extended treatments of this viewpoint, see Cohen & Cohen, 1983, and Judd, McClelland, & Smith, 1996). More commonly, other sources have argued that ANCOVA is not strictly equivalent to regressing the dependent variable on the covariate, then doing an ANOVA on the residuals (e.g., Elashoff, 1969; Maxwell & Delaney, 1990; Maxwell et al., 1985; Porter & Raudenbush, 1987). In this view, ANCOVA provides a test of the main effect of group by comparing the error sums of squares resulting from two models rather than a regression. The two models are:

$$Y_{ij} = \mu + \alpha_j + \beta X_{ij} + \epsilon_{ij}$$

$$Y_{ij} = \mu + \beta X_{ij} + \epsilon_{ij}$$

Roughly, Y_{ij} is the dependent variable for the *i*th subject in the *j*th group, μ is the grand mean, α_j is the treatment effect for the *j*th group, βX_{ij} is the product of a population regression coefficient and the score on the covariate for the *i*th subject in the *j*th group, and ϵ_{ij} is an error term for the *i*th subject in the *j*th group. The α_j term obviously differentiates the equations, but the two βX_{ij} terms would not generally be identical, either. Further discussion of these equations and of the *F* test of the main effect of group based on them can be found in Maxwell et al. (1985) and elsewhere. The point in this second view is that the actual main-effect test is not simply the direct outcome of a regression, as the MRC approach implies.

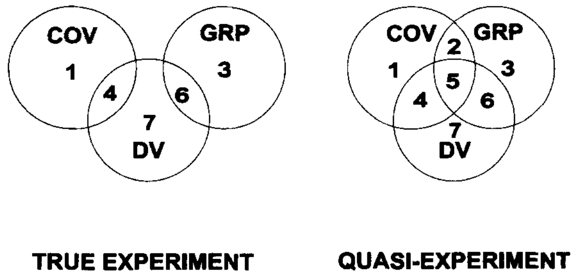

Figure 1. Two (of several) possible relationships between the group (*Grp*), covariate (*Cov*), and dependent (*DV*) variables. In both cases, removing variance in *DV* that is associated with *Cov* will reduce the variance in *DV* that is not associated with *Grp* and thus will enhance the relationship detected between *Grp* and *DV*. In the left panel, *Grp* and *Cov* share no variance. This is the classic situation for a true experiment, with random assignment to groups. Removing the variance associated with *Cov* will not alter *Grp*. Given random assignment, individual characteristics such as height or presence of hallucinations would generally be randomly distributed across the groups, and group means should not differ except by chance. In the right panel, *Grp* and *Cov* do share variance. This is often the case when preexisting groups are studied, such as comparisons of two diagnostic groups—a quasi-experiment rather than a true experiment. In such a case, removing the variance associated with *Cov* will also alter *Grp* in potentially problematic ways.

However, the distinction between the classical and MRC-based perspectives on ANCOVA is not important for the issues addressed in the present article, and the MRC approach has some expository advantages. For illustrative purposes, we assume a simple design having one covariate, *Cov*, one grouping variable, *Grp*, and one dependent variable, *DV*, discussed in the MRC framework advocated by Cohen and Cohen (1983). Figure 1 shows two of several possible relationships among the three variables, with overlap indicating shared variance and, equivalently, a nonzero correlation. In the left panel, *Cov* and *Grp* share no variance, reflecting random assignment to group. On the right, they are correlated.

Following the approach of Cohen and Cohen (1983), *Cov* is entered first in the regression. This removes variance from *Grp* that *Cov* shares with *Grp*, if any (area 2 + area 5 in Figure 1), leaving a residual portion of *Grp* with which it is not correlated, Grp_{res} (area 3 and area 6). Entry of *Cov* also removes variance it shares with *DV* from *DV* (area 4 + area 5), leaving a residual portion of *DV* with which it is not correlated, DV_{res} (area 6 + area 7). For either panel in Figure 1, removal of *Cov* leaves areas 3, 6, and 7. The regression then enters Grp_{res} into the model and computes its correlation with DV_{res}. What the investigator may view as the conventional *F* test of *Grp* is actually, instead, an evaluation of the variance shared by Grp_{res} and DV_{res}: How large area 6 is, compared with the sum of area 6 + area 7.

[1] Gregory Miller was Chair of the NIMH Clinical Psychopathology Review Committee when this article was written. Approximately one third of the grant applications he reviewed proposed a questionable or clearly invalid use of ANCOVA.

ANCOVA When Groups Do Not Differ on the Covariate

In the classical treatment of random assignment to groups, *Cov* should, in principle, share no variance with *Grp,* as a direct result of the random assignment (Figure 1, left panel). That is, the expected value of *Cov* will be the same for every group, and, except for random error, the group means for *Cov* will be identical. As a consequence, entry into the regression of *Cov* before *Grp* will remove no variance from *Grp.* Thus, $Grp = Grp_{\text{res}}$. (In practice, some nonzero correlation between *Cov* and *Grp* may be observed, but generally the correlation will be negligible. Viewpoints on cases in which it is not negligible are discussed below.) When $Grp = Grp_{\text{res}}$, the only effect of ANCOVA is to remove variance (area 4) from *DV* which, from the standpoint of *Grp,* is simply noise (error). Thus, *Grp* will correlate more highly with DV_{res} than with *DV,* resulting (all things being equal) in a larger effect size and a more powerful significance test. Specifically in the left panel of Figure 1, *Grp* overlaps with a higher proportion of (the smaller) DV_{res} (area 6 + area 7) than with *DV* (area 4 + area 6 + area 7), so the correlation of *Grp* and DV_{res} is higher than that of *Grp* and *DV.*

Used this way, ANCOVA serves as a legitimate and appealing noise-reduction technique for evaluating the relationship of *Grp* and *DV.* Other than the loss of a degree of freedom associated with inclusion of *Cov* in the analytic model, ANCOVA would appear to be valuable in improving the power of the test of *Grp* (see Bock, 1975, and Porter & Raudenbush, 1987, for technical treatments of the gain in efficiency from an appropriate use of ANCOVA). For example, a study comparing simple phobic and social phobic patients' response to phobia-relevant material might use individually developed videos that differ in duration, and duration itself might affect peak anxiety rating. If the groups do not differ on mean video duration, duration might serve as a covariate, reducing variance in anxiety rating that is driven by video duration and unrelated to diagnostic group. In general, this will improve the observed relationship between *Grp* and *DV.* Given this benefit, ANCOVA is much underutilized in the psychopathology literature.

It is important to measure *Cov* as reliably as possible to maximize its ability to capture noise variance in *DV* and to ensure that the adjusted *DV,* DV_{res}, is not contaminated by noise associated with the measurement of *Cov.* Measurement error could distort the resulting mean effects and significance test because the adjustment is necessarily for observed scores rather than true scores on the covariate (Reichardt, 1979; Elashoff, 1969; Fleiss & Tanur, 1973; Maxwell & Delaney, 1990; Richards, 1980; although see Huitema, 1980; Overall & Woodward, 1977; Porter & Raudenbush, 1987; Reichardt, 1979; and Reichardt & Bormann, 1994, for discussions of how this problem might be dealt with). This issue of measurement error is particularly problematic in studies with a small sample size.

With this MRC perspective on ANCOVA in mind, the assumptions in ANCOVA can be readily summarized (for more extended, technical treatments of the assumptions of ANCOVA, see Elashoff, 1969; Fleiss & Tanur, 1973; Huitema, 1980; Maxwell et al., 1985; Porter & Raudenbush, 1987; Wildt & Ahtola, 1978). Widely noted is that ANCOVA assumes that groups do not differ in the regression of *DV* on *Cov.* This is often referred to as *the assumption of homogeneity of regression slopes.*

Violation of this assumption is not as disabling as nonequivalence of groups on the covariate. When faced with heterogeneity of regression slopes, the investigator is encouraged (e.g., Cohen & Cohen, 1983) simply to frame the analysis as a hierarchical or simultaneous regression and in that context to include an interaction term consisting of the product of *Cov* and *Grp* (see Cohen & Cohen, 1983, for methods of representing categorical variables in such an approach). In effect, *Cov* is no longer viewed as a qualitatively distinct covariate with a purely methodological role in the analysis but as a meaningful, substantive part of the analysis. Such interactions may be theoretically interesting. Given that the investigator's goal is to identify sources of variance in *DV,* the test of such an interaction may be fruitful. Rogosa (1980) noted that the typical all-or-none framing of the assumption of equality of regression slopes is simplistic and often inappropriate. For example, with very large sample sizes, differences in slope might be "significant" but trivially small; with small sample sizes, functionally important differences in slopes might not be "significant." He discussed other ways to frame the issue and some alternative analytic strategies in the face of nonparallel regression slopes.

Invalid ANCOVA When Groups Differ on the Covariate: Consensus of the Technical Literature

The assumption of homogeneity of regression slopes is fairly well known. In contrast, the importance of groups not differing on *Cov* is not widely recognized in the psychopathology literature. Mistakenly, investigators frequently turn to ANCOVA in hopes of "controlling for" group differences on the covariate. There is no statistical means of accomplishing this "control" (Chapman & Chapman, 1973; Fleiss & Tanur, 1973; Lord, 1967).

In fact, "control" is altogether the wrong metaphor for understanding what ANCOVA accomplishes. We have found that investigators are frequently surprised when this is pointed out. Some assert that "controlling" or "removing" nontrivial group differences on the covariate is the primary use of ANCOVA. It is important to establish that relevant literature roundly condemns this view, before attempting to provide an accessible explanation for why this view is mistaken and before considering some alternatives. Modern literature on ANCOVA began with Cochran (1957), who stated, "[I]t is important to verify that the treatments have had no effect on" the covariate and "a covariance adjustment . . . may remove most of the real treatment effect" (p. 264). Campbell and Stanley (1963) spoke favorably of ANCOVA in general but cautioned, "The usual statistics [including ANCOVA, cited earlier on the same page] are appropriate only where individual students have been assigned at random to treatments" (p. 23). In contrast, ANCOVA "to compare naturally occurring groups . . . , which is contrary to the admonitions of many experts in experimental design, can yield statistically significant results which are entirely spurious" (Evans & Anastasio, 1968, p. 225). Elashoff (1969) explained:

> A basic postulate underlying the use of analysis of covariance to adjust treatment means for the effects of the covariate x is that the x variable is *statistically* independent of the treatment effect. In other words, this means that the distribution of covariate values is not affected by the treatments either through direct causation or through correlation with another affected character (and the x variable does not affect the treatment). . . . Therefore, if . . . treatments are not manip-

> ulated as independent variables but are classifications of naturally occurring groups this assumption will not be valid. . . . Analysis of covariance is inappropriate if the covariate is not independent of the treatment. (pp. 388–389)

Chapman and Chapman (1973) stated that there is no statistical method that can address the question of whether two groups that differ on variable *A* would differ on variable *B* if they did not differ on variable *A* and added, "The only legitimate use of analysis of covariance is for reducing variability of scores in groups that vary randomly. Its use is invalid for preexisting disparate groups that differ on the variable to be covaried out" (p. 82). Cohen and Cohen (1975) stated the principle forcefully:

> [O]ne does not answer such questions with [ANCOVA]. . . . What are clearly *not* warranted by the results of [ANCOVA] are subjunctive formulations like: '*If* black and white varieties [of corn that differ in height] *were* of equal height, *then* they would have equal yields.' . . . [S]uch subjunctive questions can not be answered by [ANCOVA], or indeed, by any method of analyzing data. (pp. 396–397, 398, 399; emphasis in original).[2]

Jin (1992) provided this portrayal of the issue: "[Investigators ask], 'What would the results be if the subjects were at the same basal level?' However [given nonrandom group assignment], this is an unrealistic question" (p. 182). Porter and Raudenbush (1987) provided an extended technical treatment and concluded, "It is crucial that the covariable be unaffected by the treatment" (p. 385); "In short, ANCOVA cannot be counted on to estimate the right effects or test the correct hypothesis when it is used to analyze nonrandomized experiments" (p. 391); and "ANCOVA will . . . remove some or all of the treatment effect for the dependent variable" (p. 391). In their monograph on ANCOVA, Wildt and Ahtola (1978) stated unequivocally:

> [T]his assumption, that the covariate is independent of the treatment, is a basic tenet of the analysis of covariance model. When the covariate and the treatment are not independent, the regression adjustment may obscure part of the treatment effect or may produce spurious treatment effects. . . . [V]iolation of this assumption may seriously affect the interpretation of results. (p. 90)

Fleiss and Tanur (1973) took a similarly stark position:

> [N]o amount of statistical manipulation can tell one what might have been had certain differences been non-existent. . . . The overwhelming weight of logic is on the side of those who warn that neither the analysis of covariance nor any other statistical technique can undo systematic differences which were out of the investigator's control. (p. 513, 517)

Huitema (1980) agreed:

> If a nonrandomized design other than the biased assignment design is employed and the covariate is measured after treatments are administered [i.e., if the groups' covariate means could have been affected by—correlated with—the grouping variable], the ANOVA on the covariate as well as the ANOVA and the ANCOVA on [the dependent variable] will be essentially uninterpretable because treatment effects and pretreatment differences among the populations will be confounded. (Huitema, 1980, p. 109)

Finally, "The basic desideratum is that the covariate and the treatment be statistically independent," and "The basic concern is that if the treatments differentially affect the covariate scores, then an ANCOVA . . . would in fact remove from the treatment sum of squares part of the treatment effect you really want included" (Maxwell & Delaney, 1990, pp. 380, 382–383).

Invalid ANCOVA When Groups Differ on the Covariate: The Substantive Problem

Granting this highly consistent sentiment in the technical literature, what is it that makes ANCOVA unacceptable in the face of group differences on *Cov?* Technical discussions of this issue, known as Lord's Paradox, have long been available (e.g., Bock, 1975; Fleiss & Tanur, 1973; Holland & Rubin, 1983; Lord, 1967, 1969; Maris, 1998; Maxwell & Delaney, 1990). It may be useful, in persuading researchers on this issue, to convey the issue in terms of theoretical substance instead of mathematical proof. The central problem is that often one does not know what Grp_{res} represents when *Cov* and *Grp* are related. "When the covariate . . . is affected by the treatment, the regression adjustment may remove part of the treatment effect or produce a spurious treatment effect" (Elashoff, 1969, p. 388). The grouping variable, its essence, has been altered in some substantive way that is frequently not specifiable in a conceptually meaningful way (see also Evans & Anastasio, 1968). Thus, Grp_{res} is not a good measure of the construct that *Grp* is intended to measure: "the adjustment made on the dependent variable is biased because some effects attributable to the treatment are eliminated from the dependent variable" (Wildt & Ahtola, 1978, p. 15). In the right panel of Figure 1, this problem is manifested in the fact that area 2 and area 5 are no longer part of the (residualized) grouping variable.

This problem is often not acknowledged in statistical textbooks on ANCOVA and appears to be largely unknown in the psychopathology literature, but it is quite important. For example, the considerable diagnostic comorbidity of depression and anxiety and the assumption that they share symptoms, psychological processes, and even some neural processes (e.g., Keller et al., 2000) renders complex any attempt to separate their components. If we compare a sample of depressed patients with nonpatient controls and covary out anxiety, which happens to be higher in the patients, it is not necessarily the case that the residual group difference is a clear, clean representation of depression as it would exist without the comorbid anxiety. What we should believe about that depends on our model of the relationship between depression and anxiety. If they happen to co-occur because of nonspecific severity factors that themselves are not specifically related to depression, our ANCOVA might be effective in removing such variance, leaving "pure" depression. If, however, we believe that the negative affect that depression and anxiety share is central to the concept of depression, then removing negative affect (by removing anxiety) will mean that the group variance that remains has very poor construct validity for depression.

Turning to a common example outside the psychopathology literature, one on which many technical treatments rely, we offer a modest, nonmathematical elaboration of Lord's Paradox. Lord (1967) contrasted the approach used by two hypothetical statisti-

[2] Most of this is repeated in Cohen and Cohen (1983, p. 425), with the verbatim conclusion drawn.

cians analyzing a hypothetical data set. The data are for boys and girls at the beginning and end of an academic year. The boys, as a group, start and end the year weighing more than the girls, and neither groups' average weight changes over time. An issue is whether diet affected boys and girls differentially during the school year. The statistician who favored an ANCOVA (incorrectly so, in the opinion of Lord and numerous subsequent authors quoted above) used initial weight as a covariate and concluded:

> If one selects on the basis of initial weight a subgroup of boys and a subgroup of girls having identical frequency distributions of initial weight, . . . the subgroup of boys is going to gain substantially more during the year than the subgroup of girls. (Lord, 1967, p. 305)

Lord's pro-ANCOVA statistician concluded from this that there is a meaningful differential effect of diet on boys and girls. However, the two selected subgroups are not representative of the larger groups of boys and girls. In effect, the two levels of the gender effect no longer represent the samples of boys and girls in the original analysis. What this statistician observed is merely an example of the principle of regression toward the mean. By selecting subsets of boys and girls matched on weight, the statistician will have selected boys weighing less than the boys' mean and girls weighing more than the girls' mean. Regression toward the mean, when the participants are reassessed, would be expected solely as a result of the lack of a perfect correlation between initial and final weight. Thus, comparison of the subset of boys gaining weight and the subset of girls losing weight is not evidence for systematic differential effects of diet on boys and girls.

Lord's (1967) example is compelling, in part because he constructed a case in which it is known that the groups do not change over time, so they cannot change differentially over time, as a function of diet or any other factor. We suggest that the more general point in this example is that, because gender and weight are confounded (correlated) to begin with, there is no statistical means to unconfound them in these examples—here, to study the potential differential effect of diet on the two groups.

As another illustration, consider a data set in which two groups are older men and younger women, and gender is of interest as an independent variable, *Grp*. Using age as a covariate does indeed remove age variance. The problem is that, because age and gender are correlated in this data set, removing variance associated with *Cov* will also remove some (shared) variance due to *Grp*. Within this data set, there is no way to determine what values of *DV* men younger than those tested or women older than those tested would have provided. Far from "controlling for" age, the ANCOVA will systematically distort the gender variable. As in our presentation of Lord's Paradox above, Grp_{res} will not be a valid measure of the construct of gender. Again, this is seen in the right panel of Figure 1, when area 2 and area 5 are removed from *Grp*.

Consider a data set consisting of childrens' age, height, and weight. If we conduct an ANCOVA in which height is the covariate, age is the grouping variable, and weight is the dependent variable, we are attempting to ask whether younger and older children would differ in weight if they did not happen to differ in height. If the groups indeed do not differ on the covariate, this question can be asked. But if there is something about the construct of age in childhood that inherently involves differences in height, the question makes no sense, because then age with height partialed out would no longer be age. There is no way to "equate" older and younger children on height, because growth is an inherent (not chance or noise) differentiation of the two groups.

As noted above, it would be entirely reasonable, given such a data set, to explore the relationships among age, height, and weight (Cohen & Cohen, 1983). The point is that statistical control, in the sense of cleanly removing the effect of *Cov*, is not what one would be able to accomplish with ANCOVA. Cohen and Cohen (1983) provided the following extreme example: "Consider the fact that the difference in mean height between the mountains of the Himalayan and Catskill ranges, adjusting for differences in atmospheric pressure, is zero!" (p. 425), the point being that one has not in any sense "equated" the two mountain ranges by using atmospheric pressure as a covariate.

Invalid ANCOVA in Psychopathology Research

One can readily imagine additional examples in psychopathology. If likelihood of diagnosis of schizophrenia rises with age, it would not be appropriate to try to "control for" age differences between two samples to evaluate the relationship between proportion of each sample receiving a diagnosis and some dependent variable such as current employment status. Age would be systematically related to the defining characteristic of the groups, so removing variance associated with age would, in effect, corrupt the grouping variable itself.

In a real-world example, Deldin (1996) wanted to investigate whether a brain-wave measure known as P300 is reduced in depression. This question is complicated by the fact of considerable comorbidity between anxiety and depression, noted above. Two groups of participants, diagnosed as depressed and nondepressed, completed a self-report anxiety measure. Anxiety score might be used as a covariate in an ANCOVA with diagnosis as the independent variable and P300 as the dependent variable. The hope in such an analysis would be "to control" anxiety and thus be able to observe the relationship between pure depression (not confounded with anxiety) and P300. However, this would make no sense on clinical, substantive grounds (fortunately, Deldin did not conduct an ANCOVA). The literature does not generally view anxiety as an independent, confounding variable in depression but as intimately related to depression and perhaps, in part, indistinguishable from it (e.g., Heller, Etienne, & Miller, 1995; Watson et al., 1995). Statistical methods cannot remove the "effect" of anxiety from depression if conceptually they are overlapping constructs.

As a final example, consider the effects of using gender as a covariate in a comparison of schizophrenic and depressed groups, diagnoses in which it is well established that men and women, respectively, are overrepresented (Kessler et al., 1994; Walker & Lewine, 1993). Removing gender variance could systematically alter the apparent nature of and relationships between the diagnostic groups. If, further, the dependent variable were performance on a task on which women are believed to be superior (i.e., if *Cov* correlates with both *Grp* and *DV*), such an ANCOVA would corrupt both independent and dependent variables, providing a meaningless *F* test.

Misuse or misinterpretation of ANCOVA is so widespread in the psychopathology literature, in our experience, that it is difficult to cite examples without stepping on toes. The report by Rosvold, Mirsky, Sarason, Bransome, and Beck (1956) may be used without causing offense, because it was published in a respected journal before the problem was recognized (Cochran, 1957) and because it is a deservedly famous article for other reasons, having recently been republished as a landmark (*Journal of NIH Research,* 1997). Among other important contributions, this paper introduced the still widely used Continuous Performance Test (CPT) to the neuropsychology and psychopathology literatures. The design included several groups of child and adult brain-damaged and control samples, with X and AX denoting two types of trials.

> Since there was a significant age difference between the Child subgroups and a significant IQ difference between the Adult subgroups . . . , the differences between the paired subgroups means on X and AX in these two groups were evaluated by means of an analysis of covariance. (Rosvold et al., 1956, p. 346)

Unfortunately, IQ would be very likely to be meaningfully related to brain damage, so using IQ as a covariate would disrupt any comparison of brain-damaged and control groups' performance: IQ differences would almost certainly be *part of* group differences in brain-damage status. As a consequence, removing variance associated with IQ would alter the diagnostic group variable substantively, and Grp_{res} would not merely be a de-noised surrogate for *Grp.* Age as a covariate presents the same potential problem, but it is not as likely that age is substantively related to brain damage, so ANCOVA might be viable. (The legitimacy of ANCOVA in the face of group differences on the covariate if the differences arose by chance is discussed below.)

This range of examples serves to convey the substantive (not merely mathematical) problem that arises when groups differ meaningfully on the covariate. ANCOVA does indeed "remove" the variance due to the *Cov,* but it does not successfully "control" for *Cov* if *Cov* is a meaningful part of *Grp.* On the contrary, and unfortunately for well-intentioned psychopathology researchers, ANCOVA removes meaningful variance from *Grp,* leaving an undercharacterized, vestigial Grp_{res} with an uncertain relationship to the construct that *Grp* represented. The relationship of such a Grp_{res} to *DV* or DV_{res} is often not interpretable. In general, ANCOVA is appropriate when the groups do not differ on the covariate—when inclusion of the covariate serves merely to remove noise variance unrelated to the grouping variable (left panel of Figure 1).

Possibly Valid ANCOVA When Groups Differ on the Covariate

Beyond this basic point, there are some gray areas regarding the use and misuse of ANCOVA. Bock (1975) and Wainer (1991) noted that the appropriateness of an ANCOVA depends not only on meeting statistical assumptions but also on the nature of the question posed. Heckman (1989, p. 166) stated, "A decision about the appropriate statistical procedure requires information outside of statistics." Fleiss and Tanur (1973) concluded that it is not the analysis but, in part, "the phrasing of the hypotheses and of the inferences which are usually invalid" (p. 518) in the misuse of ANCOVA, suggesting that if properly framed the analysis can be appropriate (see Fleiss & Tanur, 1973, pp. 518–521, for further discussion). Cochran (1957), Evans and Anastasio (1968), Maxwell and Delaney (1990), and Porter and Raudenbush (1987) also discussed narrow purposes or conditions under which ANCOVA might be interpretable despite nonrandom assignment.

There can thus be a range of views on how to interpret certain cases in which groups differ on *Cov.* Even in the case of random assignment to groups, nontrivial differences on *Cov* will occasionally arise by chance. In the right panel of Figure 1, the issue is whether area 2 and area 5 are substantively important parts of *Grp.* If not, then after their removal Grp_{res} would still be a valid measure of the Group construct.

How is an investigator to know when a Type I error has occurred in a test of group differences on *Cov?* A moderate position might be that, if the investigator has good reason to believe that group differences on *Cov* truly arose by chance, ANCOVA is appropriate (Maxwell & Delaney, 1990). The rationale for this view is that ANCOVA would only be removing noise variance from *Grp,* not anything substantive about *Grp.* Of course, this rationale turns on the strength of the assumption that the differences on *Cov* did, in fact, arise by chance. In other words, the investigator must be convinced (and must convince readers) that, among the populations from which the groups are sampled, there is no relationship with *Cov* and that the available samples are sometimes nevertheless a good representation of those populations. In the case of random assignment, the basis for this judgment is quite strong, assuming good execution of the group assignment process. In the case of nonrandom assignment—almost invariably the case in studies of psychopathology—this is often a difficult judgment to defend. However, investigators should be free to consider it and to make their case.

Overall and Woodward (1977) went a step beyond the arising-randomly criterion and argued that what matters is whether *Grp* could have caused the group differences on *Cov.* If not, in their view, then ANCOVA may be legitimate, even if the group differences on *Cov* are substantive. Similarly, Wildt and Ahtola (1978) suggested that ANCOVA might be acceptable if the investigator is certain that the *Grp* could not have affected *Cov.* We (see also passages quoted above from Elashoff, 1969; Maxwell & Delaney, 1990; and Porter & Raudenbush, 1987) do not find this argument compelling, in general. It will still happen that, because *Cov* will be entered into the regression before *Grp, Cov* will in effect get credit for any relationship of their shared variance that is also shared with *DV.* Thus, not only will Grp_{res} be a much-diminished representation of the construct that *Grp* measures (a conceptual problem) but also the relationship of Grp_{res} and DV_{res} may be underestimated (a statistical problem). Cohen and Cohen (1983) offered another objection, in that it is often impossible to determine causal relationships between *Grp* and *Cov.* Overall and Woodward offered a more specific analysis, however, which is sound: In the absence of random assignment, one might be able to render ANCOVA legitimate given an experimentally appropriate assignment to group on the basis of scores on the covariate. Their argument turns on particular patterns of relationships among *Cov, Grp,* and *DV.* In effect, the point is that the legitimacy of ANCOVA depends on a careful examination of these relationships. Unfortunately, this suggestion obviously would not apply in studies of preexisting groups typical in psychopathology.

One other gray area to consider is the context of exploration of a data set, in contrast to the more commonly explicit goal of hypothesis testing. As Huitema (1980, p. 108) outlined, it may be useful to conduct a series of ANCOVAs, trying different variables as the covariate, as a means of understanding the patterns of shared variance in a data set. He hastened to add that, when groups differ on the covariate, "it would be reasonable to *speculate* that the treatment effect on the dependent variable is mediated by the covariate. It couldn't be *concluded*. . . . " In effect, Huitema was describing the strategy of Cohen and Cohen (1983) in using a variety of hierarchical regression analyses to evaluate the relationships among variables. Procedures such as those discussed by Baron and Kenny (1986) might then be used to confirm mediational relationships.

Beyond Statistical Concerns

We have emphasized that the problem with ANCOVA in the face of group differences on the covariate is as much a substantive and interpretative issue as it is a mathematical issue. It can be noted that the problem of misuse of ANCOVA putatively to "control for" differences on the covariate is not confined to academe. An article in a prominent weekly political magazine stated:

> On the average, students in predominantly White districts did much better on reading tests than those in predominantly Black districts. The reason? Not poverty, parent education, or class size. The White districts were, of course, generally better off, the parents of pupils were more highly educated, and so on. But *when these differences were held constant statistically,* the difference that made a difference was the quality of the teaching staff, as measured by a language skills exam. (Thernstrom, 1991, p. 22; emphasis added)

It is apparent from this example that the public-policy stakes associated with the misuse or misinterpretation of ANCOVA can be quite high. As we have argued on substantive grounds, complementing previous technical discussions, it is simply not possible for such differences to be "held constant statistically" as this article claims was done, unless we were to suppose that teachers were assigned to White and Black school districts randomly or that differences in poverty, parental education, and class size existed only by chance. Surely, teacher recruitment (and subsequent performance) is in part driven by differences in district poverty, parent education, class size, etc. All of these variables are correlated, and they must be understood as inherently confounded if one adopts the superficially appealing but inappropriate goal of "controlling for" some of them. Ultimately, "quality of teaching staff" is not a variable substantively left in the analysis by the time all of those covariates that are surely meaningfully correlated with it have been partialed out. Only a highly residualized variable given the same name is being tested. One could not conclude anything about the original variable, "quality of teaching staff." At best, one could speculate, as Huitema (1980) suggested. Such nuances are often lost in public debate and, in our experience, in the psychopathology literature.

Some Alternatives to ANCOVA

Methods

Once again, analysis of covariance cannot tell us how groups would differ if they did not differ on the covariate. What then should the investigator do instead of analysis of covariance? Maxwell and Delaney (1990) discussed the analysis of gain scores and the blocking of subjects on the covariate, noting limitations in both strategies relative to ANCOVA (see also Reichardt & Bormann, 1994). Cohen and Cohen (1983) and Harris et al. (1971) recommended incorporating the covariate into the analysis—no longer conceived as a covariate but as another substantive variable. Both of these are sound, relatively familiar strategies. Rosenbaum and Rubin (1984; Rosenbaum, 1995; Rubin, in press) discussed the less widely known concept of *propensity score,* the conditional probability of assignment to a particular group or treatment given a set of observed covariates. They noted that there are circumstances under which propensity score analysis can balance the group on the covariates. It cannot address unobserved differences in covariates, although there are methods of determining the extent of bias arising from unobserved covariates (Rosenbaum, 1984). Less specific but potentially fruitful is the aggregation (perhaps via formal meta-analysis) of a large number of individually compromised studies, none of which may have random assignment but collectively present a strong inferential case (Rosenbaum, 1987). For example, no study of lung cancer has undertaken random assignment to chronic smoking and nonsmoking groups, but a strong case for causal factors in the face of potentially confounded covariates (such as hypertension, on which smokers and nonsmokers might vary and on which smoking can have causal effects) has been made on the basis of a variety of different studies. Importantly for psychopathology research, Rosenbaum's suggestion can be generalized beyond traditionally defined control groups. For example, in a study comparing symptoms of psychosis observed in schizophrenia and bipolar patients, there are advantages to recruiting multiple samples, known to be different, within each diagnosis.

Placement of group means on the regression line. Fleiss and Tanur (1973) suggested another alternative to ANCOVA. They reasoned from the principle that a problem in using analysis of covariance for the comparison of intact groups is that the regression line within groups cannot legitimately be used for between-groups comparisons. They inferred from this that the investigator should consider the performance of many differing groups of participants, examining how different groups place on the regression line.

As a simple example to illustrate their method, an investigator might wish to test the hypothesis that schizophrenic patients show a greater concreteness of proverb interpretation than is accounted for by their lower Verbal IQ score on the Wechsler Adult Intelligence Scale (WAIS). (Schizophrenic individuals' thinking has often been characterized as more concrete and less abstract than nonpatients'. Judging that a hammer and a screwdriver are both tools is more abstract than that they are both made of metal.) The investigator recognizes that she or he cannot simply match subgroups of schizophrenic and control participants on WAIS IQ and then compare the matched subgroups on concreteness because such a comparison would be contaminated by regression toward the mean (see discussion of Lord's Paradox above). Fleiss and Tanur's (1973, p. 522) solution would be to extend the study of IQ and concreteness to "many different kinds of subjects (normals, neurotics, depressives, etc.), all having in common the fact that they are not schizophrenic." Then, after obtaining mean scores on IQ and on WAIS Verbal IQ for each group, the investigator finds a transformation of the scores such that a nearly linear relation can be fitted to all pairs of group means. In our example, a straight line would be fitted to the relation of mean group concreteness score to mean group IQ score. Then the means of the schizophrenic group are examined in relation to the fitted line. If the schizophrenic patients score more deviantly on concreteness than predicted by the multigroup regression line, one infers that their concreteness is unusual.

This solution is problematic. It is, of course, rather impractical to test many groups in most such studies. More importantly, the method may be flawed in some cases. If, as Fleiss and Tanur suggested, only transformation data rather than original data fit a straight line, the meaning of that linearity is unclear. Some transformations might alter the data considerably, and the transformation is selected to fit all of the pairs of group means to a straight line except those of the schizophrenics. This procedure appears vulnerable to random error that could foster the predicted result that the means of the schizophrenic group, but not the other groups, do not fit that line. It is quite possible that, if one included the schizophrenic group among those for which a straight line is fitted, but excluded some other group, such as depressed patients, the means for that excluded group would fail to fall on the line, as a procedural artifact.

Regressing the dependent variable on the independent variable for control participants of a wide range of performance. M. B. Miller, Chapman, and Chapman (1993) suggested supplementing the "normal" control group with additional persons from a kind of group that is not grossly pathological but is known to be impaired on a variable that might be a confound in the results. They studied the prior preparatory interval (PPI) effect in schizophrenia (wherein reaction time depends on the length and predictability of the intertrial interval) and wished to determine whether schizophrenic patients' greater overall slowness might be a confound. Accordingly, the investigators added elderly individuals to their normal control group because the elderly tend to demonstrate overall slowness on reaction time. They found a very similar regression slope of the PPI effect score as a function of overall slowness for elderly as for younger normal participants. Accordingly, they computed a combined regression line. Using the slope and intercept of that line, they computed residualized scores for the schizophrenic patients' PPI effect.

To apply this method to the example of concreteness of proverb interpretation in schizophrenia, one might add borderline retarded individuals to the control group, compute the regression of concreteness score on IQ score for this expanded group of nonpatients, and then use the regression line to compute residualized scores of concreteness for the schizophrenic patients. A finding of deviantly high residualized concreteness scores for the schizophrenic patients would indicate that their concreteness is greater than expected by the standards of a group that includes borderline retarded persons.

This method has several limitations. First, it is based on the hypothesis that the regression line for the original normal group and that for the additional impaired nonschizophrenic group have the same slope and intercept. If either the slopes or intercepts should differ much, the combining of the two groups to compute a joint regression line would appear inappropriate. In addition, the range of scores on the predictor variable (IQ in the example) must be as great for the combined control group as for the schizophrenic patients. This is crucially important for the residualized scores to be meaningful.

A third problem is that the answer yielded by the study does not precisely match the original question. The investigators' hypothesis was probably not concerned with a comparison of the concreteness of schizophrenia with that of borderline retardation. In our view, however, this kind of answer is often better than the available alternatives.

Substantive Questions

Future work may produce more general or more satisfactory means to address the question psychopathologists usually attempt to answer with ANCOVA, a question for which the technical literature shows ANCOVA to be inappropriate: How would the groups differ on *DV* if they did not differ on *Cov?* However, we have argued that fundamental logical problems with such a question preclude its being meaningful.

Rather than pursue such a question, we believe that psychopathologists would do well to frame questions that a rich experimental design can address. For example, the high comorbidity between depression and anxiety need not be seen as a barrier to research. The comorbidity suggests that depression and anxiety are not entirely distinct concepts and not phenomena that should be separated or can be represented in a fully factorial design. Instead, one can articulate a specific relationship between them and then design an appropriate study. An appealing research design might include depressed individuals varying in level of anxiety and/or anxious individuals varying in level of depression. A nonadditive model of the characteristics and dynamics of comorbid depression and anxiety may best fit the resulting data.

Similarly, variables that have often been viewed as confounds in research on schizophrenia may be incorporated into models as significant conceptual players. For example, given the large literature on cognitive deficits in schizophrenia, reduced IQ need not be viewed as something to "control for" but rather as a feature of the disorder in many cases, to be studied as one studies more traditional clinical symptoms. The broader lesson here is that, if one's questions cannot be served by one's methods, one can reconceptualize the questions to suit the best available methods.

It is our hope that the present discussion provides an accessible portrayal of the proper use of ANCOVA. Relevant technical literature overwhelmingly condemns its uncritical use in the face of group differences on the covariate despite its continuing popularity in such situations. Nonrandom assignment to groups presents psychopathologists with design, statistical, and interpretative challenges, and we have provided some comments on attempts to address those challenges. Studies of psychopathology often involve group comparisons without random assignment. Investigators should consider ANCOVA and the alternatives briefly reviewed here to improve their designs. A more fundamental point is that investigators should reexamine the issues they may have wished to address with ANCOVA and consider whether other issues are more fruitful theoretically and more achievable methodologically.

References

Baron, R. M., & Kenny, D. A. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51,* 1173–1182.

Bock, D. (1975). *Multivariate statistical methods in behavioral research.* New York: McGraw-Hill.

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

Chapman, L. J., & Chapman, J. P. (1973). *Disordered thought in schizophrenia.* New York: Appleton-Century-Crofts.

Cochran, W. G. (1957). Analysis of covariance: Its nature and uses. *Biometrics, 44,* 261–281.

Cohen, J., & Cohen, P. (1975). *Applied multiple regression/correlation analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd edition). Hillsdale, NJ: Erlbaum.

Deldin, P. J. (1996). *Information processing in major depression: The ERP connection.* Unpublished doctoral dissertation, University of Illinois, Champaign.

Dixon, W. J. (1992). *BMDP statistical software manual.* Los Angeles: University of California Press.

Elashoff, J. D. (1969). Analysis of covariance: A delicate instrument. *American Educational Research Journal, 6,* 383–401.

Evans, S. H., & Anastasio, E. J. (1968). Misuse of analysis of covariance when treatment effect and covariate are confounded. *Psychological Bulletin, 69,* 225–234.

Fleiss, J. L., & Tanur, J. M. (1973). The analysis of covariance in psychopathology. In M. Hammer, K. Salzinger, & S. Sutton (Eds.), *Psychopathology: Contributions from the social, behavioral, and biological sciences* (pp. 509–527). New York: Wiley.

Harris, D. R., Bisbee, C. T., & Evans, S. H. (1971). Further comments: Misuse of analysis of covariance. *Psychological Bulletin, 75,* 220–222.

Heckman, J. J. (1989). Causal inference and nonrandom samples. *Journal of Educational Statistics, 14,* 159–168.

Heller, W., Etienne, M. A., & Miller, G. A. (1995). Patterns of perceptual asymmetry in depression and anxiety: Implications for neuropsychological models of emotion and psychopathology. *Journal of Abnormal Psychology, 104,* 327–333.

Holland, P. W., & Rubin, D. B. (1983). On Lord's paradox. In H. Wainer & S. Messick (Eds.), *Principles of modern psychological measurement* (pp. 3–35). Hillsdale, NJ: Erlbaum.

Huitema, B. (1980). *Analysis of covariance and alternatives.* New York: Wiley.

Jin, P. (1992). Toward a reconceptualization of the Law of Initial Value. *Psychological Bulletin, 111,* 176–184.

Judd, C. M., McClelland, G. H., & Smith, E. R. (1996). Testing treatment by covariate interactions when treatment varies within subjects. *Psychological Methods, 1,* 366–378.

Keller, J., Nitschke, J. B., Bhargava, T., Deldin, P. J., Gergen, J. A., Miller, G. A., & Heller, W. (2000). Neuropsychological differentiation of depression and anxiety. *Journal of Abnormal Psychology, 109,* 3–10.

Kessler, R. C., McGonagle, K. A., Zhao, S., Nelson, C. B., Hughes, M., Eshleman, S., Wittchen, H., & Kendler, K. S. (1994). Lifetime and 12-month prevalence of *DSM-III-R* psychiatric disorders in the United States. *Archives of General Psychiatry, 51,* 8–19.

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin, 68,* 304–305.

Lord, F. M. (1969). Statistical adjustments when comparing preexisting groups. *Psychological Bulletin, 72,* 336–337.

Maris, E. (1998). Covariance adjustment versus gain scores—revisited. *Psychological Methods, 3,* 309–327.

Maxwell, S. E., & Delaney, H. D. (1990). *Designing experiments and analyzing data: A model comparison perspective.* Belmont, CA: Wadsworth.

Maxwell, S. E., Delaney, H. D., & Manheimer, J. M. (1985). ANOVA of residuals and ANCOVA: Correcting an illusion by using model comparisons and graphs. *Journal of Educational Statistics, 10,* 197–209.

Miller, M. B., Chapman, L. J., & Chapman, J. P. (1993). Slowness and the preceding preparatory interval effect in schizophrenia. *Journal of Abnormal Psychology, 102,* 145–151.

Neale, J. M., & Oltmanns, T. F. (1980). *Schizophrenia.* New York: Wiley.

Overall, J. E., & Woodward, J. A. (1977). Nonrandom assignment and the analysis of covariance. *Psychological Bulletin, 84,* 588–594.

Porter, A. C., & Raudenbush, S. W. (1987). Analysis of covariance: Its model and use in psychological research. *Journal of Counseling Psychology, 34,* 383–392.

Reichardt, C. S. (1979). The statistical analysis of data from nonequivalent group designs. In T. D. Cook & D. T. Campbell (Eds.), *Quasi-experimentation: Design and analysis issues for field settings* (pp. 147–205). Boston: Houghton Mifflin.

Reichardt, C. S., & Bormann, C. A. (1994). Using regression models to estimate program effects. In J. S. Wholey, H. P. Hatry, & K. E. Newcomer (Eds.), *Handbook of practical program evaluation* (pp. 417–455). San Francisco: Jossey-Bass.

Richards, J. E. (1980). The statistical analysis of heart rate: A review emphasizing infancy data. *Psychophysiology, 17,* 153–166.

Rogosa, D. (1980). Comparing nonparallel regression lines. *Psychological Bulletin, 88,* 307–321.

Rosenbaum, P. R. (1984). From association to causation in observational studies: The role of strongly ignorable treatment assignment. *Journal of the American Statistical Association, 79,* 41–48.

Rosenbaum, P. R. (1987). The role of a second control group in an observational study (with discussion). *Statistical Science, 2,* 292–316.

Rosenbaum, P. R. (1995). *Observational studies.* New York: Springer-Verlag.

Rosenbaum, P. R., & Rubin, D. (1984). Reducing bias in observational studies using subclassifications on the propensity score. *Journal of the American Statistical Association, 79,* 516–524.

Rosvold, H. E., Mirsky, A. F., Sarason, I., Bransome, E. D., Jr., & Beck, L. H. (1956). A continuous performance test of brain damage. *Journal of Consulting Psychology, 20,* 346–350.

Rubin, D. B. (in press). Propensity score methods. In L. Bickman (Ed.), *Contributions to research design: Donald Campbell's legacy* (Vol. 2). Thousand Oaks, CA: Sage.

Siddle, D. A. T., & Turpin, G. (1980). Measurement, quantification, and analysis of cardiac activity. In I. Martin & P. H. Venables (Eds.), *Techniques in psychophysiology* (pp. 139–246). London: Wiley.

Thernstrom, A. (1991, December 16). Beyond the pale. *New Republic.*

Wainer, H. (1991). Adjusting for differential base rates: Lord's Paradox again. *Psychological Bulletin, 109,* 147–151.

Walker, E., & Lewine, R. (1993). Sampling biases in studies of gender and schizophrenia. *Schizophrenia Bulletin, 19,* 1–7.

Watson, D., Weber, K., Assenheimer, J. S., Clark, L. A., Strauss, M. E., & McCormick, R. A. (1995). Testing a tripartite model: I. Evaluating the convergent and discriminant validity of anxiety and depression symptom scales. *Journal of Abnormal Psychology, 104,* 3–14.

Wildt, A. R., & Ahtola, O. T. (1978). *Analysis of covariance.* Beverly Hills, CA: Sage.

Received November 3, 1998
Revision received September 7, 2000
Accepted September 7, 2000 ■