encephalitis: A case-control study. *Amer. J. Epidemiol.* **111** 415–424.
HAMILTON, M. A. (1979). Choosing a parameter for 2 × 2 table or 2 × 2 × 2 table analysis. *Amer. J. Epidemiol.* **109** 362–375.
HILL, A. B. (1965). The environment and disease: Association or causation? *Proc. Roy. Soc. Med.* **58** 295–300.
HILLER, R., GIACOMETTI, L. and YUEN, K. (1977). Sunlight and cataract: An epidemiologic investigation. *Amer. J. Epidemiol.* **105** 450–459.
HOLLAND, P. W. (1986a). Which comes first, cause or effect? *New York Statist.* **38** 1–6.
HOLLAND, P. W. (1986b). Statistics and causal inference (with discussion). *J. Amer. Statist. Assoc.* **81** 945–970.
HOLLAND, P. W. and RUBIN, D. B. (1980). Causal inference in prospective and retrospective studies. J. Cornfield Memorial Lecture at the American Statistical Association Meetings. Unpublished manuscript.
HOLLAND, P. W. and RUBIN, D. B. (1983). On Lord's paradox. In *Principles of Modern Psychological Measurement: Festschrift for Frederick M. Lord* (H. Wainer and S. Messick, eds.). Lawrence Earlbaum, Hillsdale, N. J.
KEMPTHORNE, O. (1952). *The Design and Analysis of Experiments.* Wiley, New York.
LEHMANN, E. L. (1959). *Testing Statistical Hypotheses.* Wiley, New York.
LEHMANN, E. L. (1966). Some concepts of dependence. *Ann. Math. Statist.* **37** 1137–1153.
LILIENFELD, A. M. and LILIENFELD, D. E. (1980). *Foundations of Epidemiology*, 2nd ed. Oxford Univ. Press, New York.
MACMAHON, B. (1984). Coffee and cancer of the pancreas: A review. In *Coffee and Health* (B. MacMahon and T. Sugimura, eds.) 109–115. Cold Spring Harbor Laboratory, Cold Spring Harbor, N. Y.
MACMAHON, B. and PUGH, T. (1970). *Epidemiology: Principles and Methods.* Little, Brown, Boston.
MANTEL, N. (1973). Synthetic retrospective studies and related topics. *Biometrics* **29** 479–486.
MANTEL, N. and HAENSZEL, W. (1959). Statistical aspects of retrospective studies of disease. *J. Nat. Cancer Inst.* **22** 719–748.
POPPER, K. (1959). *The Logic of Scientific Discovery.* Harper and Row, New York.
ROSENBAUM, P. R. (1984a). From association to causation in observational studies: The role of tests of strongly ignorable treatment assignment. *J. Amer. Statist. Assoc.* **79** 41–48.
ROSENBAUM, P. R. (1984b). Conditional permutation tests and the propensity score in observational studies. *J. Amer. Statist. Assoc.* **79** 565–574.
ROSENBAUM, P. R. (1984c). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *J. Roy. Statist. Soc. Ser. A* **147** 656–666.
ROSENBAUM, P. R. (1986). Dropping out of high school in the United States: An observational study. *J. Educat. Statist.* **11** 207–224.
ROSENBAUM, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika* **74** 13–26.
ROSENBAUM, P. R. and RUBIN, D. B. (1983a). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55.
ROSENBAUM, P. R. and RUBIN, D. B. (1983b). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *J. Roy. Statist. Soc. Ser. B* **45** 212–218.
ROSENBAUM, P. R. and RUBIN, D. B. (1985). The bias due to incomplete matching. *Biometrics* **41** 103–116.
RUBIN, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educat. Psychol.* **66** 688–701.
RUBIN, D. B. (1977). Assignment to treatment group on the basis of a covariate. *J. Educat. Statist.* **2** 1–26.
RUBIN, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.* **6** 34–58.
RUBIN, D. B. (1980). Discussion of "Randomization analysis of experimental data: The Fisher randomization test," by D. Basu. *J. Amer. Statist. Assoc.* **75** 591–593.
SOLOMON, R. L. (1949). An extension of the control group design. *Psychol. Bull.* **46** 137–150.
WELCH, B. L. (1937). On the z-test in randomized blocks and Latin squares. *Biometrika* **29** 21–52.
YERUSHALMY, J. and PALMER, C. E. (1959). On the methodology of investigations of etiologic factors in chronic diseases. *J. Chronic Dis.* **10** 27–40.

Comment

Paul W. Holland

I am pleased to see that Rosenbaum's work is included in *Statistical Science*—for two reasons. First, observational studies are very common in scientific work and yet from a theoretical perspective they are poorly understood, often maligned and rarely subjected to serious formal analysis. Rosenbaum's discussion here and elsewhere shows that a formal analysis can lead to useful, practical tools that can help in the design and analysis of nonrandomized studies. Such work ought to be widely publicized and *Statistical Science* is an attractive forum. Second, the particular formal analysis used here by Rosenbaum elaborates and extends the approach I call "Rubin's model" (Holland, 1986a, 1986b) and which I personally feel needs to become wider known and used by mathematical statisticians. My experience over the last 10 years has been that *any* problem involving causal inferences (e.g., inferences about the effects of treatments) is

Paul W. Holland is Distinguished Research Scholar and Director of the Research Statistics Group, Educational Testing Service, 21-T Rosedale Road, Princeton, New Jersey 08541.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *Statistical Science.* www.jstor.org

clarified and illuminated by applying an appropriate version of the analysis advocated in Rubin (1974). The present paper does exactly this in two ways. First, by clarifying what two different control groups can really contribute to a nonrandomized study and how best to choose them. Second, by illustrating how to test the type of implicit assumptions one makes in the analysis of data from such studies. I used to think that these assumptions were untestable, but Paul has convinced me that, in a limited way, they can be tested and our conclusions from such studies improved accordingly. Such results show the power of Rubin's model and are just the most recent in a long list of such results that convince me that Rubin's model ought to be a standard member of every statistician's tool kit—like the normal distribution or the general linear model! In short, I have little but praise for this paper and hope that it interests others in the work that Rosenbaum, Rubin and others are doing in this important area.

My own experience with multiple control groups in an observational study is limited to a single study of computer-assisted instruction (CAI) carried out during the late 1970s in Los Angeles (Ragosta, Holland and Jamison, 1982). In this study there were four "CAI" schools (where CAI equipment was made available) and two "comparison" schools. In the CAI schools, for the first year of the study, students in grades 2, 4 and 6 were assigned to 20 minutes a day of CAI in one of three curricula—mathematics (M), reading (R) or language arts (L). In the second year of the study, only grades 1, 3 and 5 of the CAI schools were assigned to the CAI conditions. All students in all six schools were tested in the Fall and Spring in all three subjects areas by using appropriate standardized tests. The Spring testing (the posttest) was part of the normal standardized testing in the schools, whereas the Fall testing (the pretest) was introduced especially for the study. This design yielded two nonrandomized control groups for, say, the grade 4 CAI students in the first year of the study. These were:

C1 = comparable (on pretest) grade 4 students in the comparison schools;
C2 = comparable (on pretest) students in grade 4 in the CAI schools in the second year of the study.

The C1 controls were in different schools but were tested in the same year as the CAI students. The C2 controls were in the same schools but were tested in a different year than the CAI students. An application of Rosenbaum's proposal for testing the assumption of X-adjustable treatment (i.e., CAI) assignment (here X is the pretest score and the response in the posttest score) is a comparison of the regressions of posttest on pretest in the two control groups, C1 and C2. This was not done, so I cannot show how Rosenbaum's idea would look in such a case, but it might have been a useful analysis. I was always troubled by the results obtained by these two nonrandomized comparisons because they tended to be unrelated to the consistent and clear-cut results we obtained upon examination of the randomized comparisons in the study (Holland, Jamison and Ragosta, 1979). Within classrooms in the CAI grades we randomly assigned the students to one of the possible CAI curricula. Hence, R students, assigned to the reading curriculum, served as controls for M students, assigned to the mathematics curriculum. In such a comparison the response would be a mathematics test. Then, the M students served as controls for the R students for performance on a reading test.

I came away from this study wondering how people could place much confidence in nonrandomized controls in studies of this sort—in such studies it is much more common to use nonrandomized rather than randomized controls. It was not that pretests are not good covariates—the pretest-posttest correlations are often over .7—rather it was the large teacher differences within schools and the substantial differences in the actual year to year testing times that we experienced.

These problems identify two assumptions that Rosenbaum makes but does not comment on. The first has to do with the responses, R_T and R_C, and the second has to do with the treatment to which the various control groups are exposed. I discuss these briefly and perhaps Rosenbaum will add comments in his rejoinder.

The Response Variable. In the control group C2, described above, there was a real possibility that the meaning of the response (i.e., the Spring testing results) was not the same as it was for the treatment group. This was due to year to year variation in what "Spring testing" means. First of all, the testing dates can vary from year to year for a variety of reasons. Secondly, something like a "flu" outbreak in the middle of one year can significantly change the amount of time students are in school between "Fall" and "Spring" testing. In Rosenbaum's medical examples, the values of the response, R_T and R_C, may be more clear-cut, but I am sure this depends on how deeply one delves into the definition of the response measure.

The Control Groups' Treatments. Rosenbaum assumes that in the control groups the value of R_C is observed on each subject, as opposed to the value of R_T. In my CAI example, this could be seriously questioned because the teachers and schools are different in C1 and C2 and one might expect these to make a difference in the schooling the students experience. In fact, the assumption which Rubin (1980, 1986) calls the SUTVA (stable unit-treatment value assumption) is probably false in the CAI example if teachers and schools can affect students' test scores differentially.

Again, the SUTVA is implicit in Rosenbaum's analysis and its violation might make the application of his results ineffective. For his medical examples, the SUTVA is not satisfied if the model only specifies a single control response, R_C rather than an R_{C_i} for the ith control group and, in fact, subjects in the first control group are exposed to a different kind of treatment than are those in the second control group. On the other hand, Rosenbaum shows that when such an assumption holds there is a clear benefit for the design and analysis of observational studies.

Because Rosenbaum focuses on the problem of assessing biases due to pretreatment differences, it is a little unfair to ask him to solve these other problems as well, but from the excellence of the present paper I am sure he is up to the task.

ADDITIONAL REFERENCES

HOLLAND, P. W., JAMISON, D. and RAGOSTA, M. (1979). *Computer-assisted Instruction and Compensatory Education: The ETS/LAUSD Study. Data Analysis: Fiscal year 1978.* Educational Testing Service, Princeton, N. J.

RAGOSTA, M., HOLLAND, P. W. and JAMISON, D. (1982). *Computer-assisted Instruction and Compensatory Education: The ETS/LAUSD Study. Final Report.* Educational Testing Service, Princeton, N. J.

RUBIN, D. B. (1986). What ifs have causal answers, discussion of "Statistics and causal inference," by P. Holland. *J. Amer. Statist. Assoc.* **81** 961–962.

Comment: The Use of Multiple Control Groups in Designed Experiments

Barry H. Margolin

I want to commend Dr. Rosenbaum on a most lucid presentation regarding the role of a second control group in an observational study. My contribution to this discussion is motivated by a remark of Cochran (1965), that in discussions of topics in observational studies, "it is relevant to indicate how the problem is tackled in controlled experimentation . . . . " With this in mind, I will discuss briefly and illustrate the roles that multiple control groups have played in designed experiments; these illustrations draw heavily upon my own research simply because they are readily accessible to me. Three distinct roles are discernible: (i) to detect the presence of unsuspected systematic effects; (ii) to determine whether there are hidden sources of extraneous random variability; and (iii) to assemble sufficient control data to permit a meaningful assessment of sampling model assumptions.

The earliest experiments whose design included features resembling multiple control groups appear to be uniformity trials in agricultural research (Cochran, 1937). These are agricultural experiments in which the land is divided into a number of plots of the same size. A single variety of the crop of interest is planted, although other factors such as fertilizer are kept constant from plot to plot, and the yield of each plot is observed. As Cochran (1937) noted, the primary purpose of a uniformity trial is to study the effects of amalgamation of the original plots into "larger plots of various sizes and shapes" and "to provide information on the optimum size and shape of plot" with regard to experimental error. As such, a uniformity trial is viewable as an experiment with a control group but no treatment group.

The analogy with multiple control groups becomes clearer, however, when one observes that uniformity trials are also conducted to validate the applicability of tests of significance that are based on analysis of variance (ANOVA). As Cochran (1937) writes,

> "A preliminary requirement for the application of the analysis of variance to be possible is that the experimental design used should be chosen at random from a set of designs such that, in the absence of any treatment effect, the average treatment mean square over the set should equal the average error mean square. . . . The further question arises: how good an approximation to the tabulated z distribution is generated by the process of randomization used? There again the question may be tested from uniformity trial data."

When any particular design is imposed on the uniformity trial data, which involve no true treatments, the results hopefully appear as if they derived from a

Barry H. Margolin is Head, Statistical Methodology Section, National Institute of Environmental Health Sciences, P. O. Box 12233, Research Triangle Park, North Carolina 27709.