Evaluation researchers are often confronted with less than optimal conditions in which to design studies. When this occurs, researchers may be forced to utilize relatively weak designs that do not rule out all threats to internal validity. Using archival data from a sales campaign for a state lottery, this article illustrates a multiplist strategy (Cook, 1985) in which several complementary designs are utilized to help rule out the four threats to internal validity associated with the frequently utilized nonequivalent control group design. Specific methods for addressing each of these threats and strengthening the basic nonequivalent control groups design are also illustrated.

A MULTIPLIST STRATEGY FOR STRENGTHENING NONEQUIVALENT CONTROL GROUP DESIGNS

KIM D. REYNOLDS
STEPHEN G. WEST
Arizona State University

Researchers who work in applied settings are often confronted with less than optimal conditions in which to design studies and evaluations. Among these conditions are a lack of control over the assignment of experimental units to treatments and lack of control over crucial measurement issues. When these problems occur, the researcher may be forced to utilize relatively weak designs, limiting the certainty with which causal inferences may be reached about the effectiveness of the treatment program. That is, when programs are evaluated under these conditions it is normally difficult to determine whether changes in

AUTHORS' NOTE: *The authors wish to thank Leona Aiken, Joseph Hepworth, David Kenny, Peter Reingen, and Lee Sechrest for their comments on an earlier version of this article. The authors also thank Charles Buri, Debi Armstrong, Glenn Shippee and Donna Schwartzman at the Arizona Lottery for their assistance in obtaining the data. Requests for reprints should be sent to Stephen G. West, Department of Psychology, Arizona State University, Tempe, AZ 85287.*

EVALUATION REVIEW, Vol. 11 No. 6, December 1987 691-714

the outcome measures are due to the program or to a variety of other potential causal factors that may have coincided with the implementation of the program.

In this article we wish to illustrate how several quasi-experimental designs can be combined to rule out nearly all plausible threats to internal validity. Few examples of such a multiplist strategy have been presented in the evaluation literature (see e.g., Lipsey et al., 1981), although such a strategy has been recommended by several methodologists (Cook, 1985; Cook and Campbell, 1979; West, 1985). We also focus on the utilization of additional design features that strengthen the basic nonequivalent control group design commonly used in evaluation research.

It has traditionally been considered extremely difficult if not impossible to achieve adequate levels of internal validity using other than a randomized experimental design. According to this view, random assignment of subjects to treatments and the control achieved by a laboratory setting are needed to eliminate threats to internal validity (compare Cook, 1983). However, the applied researcher is often not able to implement a random assignment strategy and is constrained in the design of research by any number of practical considerations. Issues of validity and design are often ignored or sacrificed under these conditions leading to designs and conclusions that are weak and confounded. We acknowledge both positions and believe that a fallback methodology can be adopted by applied researchers that allows for the design of relatively strong studies given the constraints within which they operate.

One means of accomplishing this objective that has received considerable attention in the literature is to utilize a multiple indicator, structural modeling approach (Magidson, 1977; Magidson and Sörbom, 1980; Sörbom, 1978, 1981) or an econometric approach (Barnow et al., 1980; Dwyer, 1984; Heckman, 1979; Murnane et al., 1985; Muthén and Jöreskog, 1983). However, these statistical adjustment approaches require (a) the use of very large sample sizes to obtain robust parameter estimates through maximum likelihood techniques (Boomsma, 1982; see also Tukey, 1986, for general comments on the applicability of asymptotic statistical tests), (b) that highly reliable measures of all variables related to the selection process be available, and (c) strong statistical assumptions that must be met before accurate estimates of treatment effects can be made.[1] These conditions often cannot be met by researchers working in an applied setting.

Another way to accomplish this objective is to incorporate design features that at least partially address specific threats to validity and to utilize multiple designs that may each be flawed in some respect but that share no systematic bias. Such a multiple-design strategy has been outlined by Cook and his colleagues (Cook, 1985; Cook and Campbell, 1979; Houts et al., 1986; Shadish et al., 1986). A series of designs that have different strengths and that share no systematic directional bias allow the researcher to eliminate several competing explanations for the results of the program under study. If the direction of the potential bias is different across designs, yet the results converge on the same conclusion, one can have confidence about the program's effectiveness. If the designs do not agree in their conclusion, the researcher is faced with an empirical puzzle and must attempt to explain why this is the case. This form of multiple investigation can then lead to the elucidation of variables that moderate the effectiveness of the program.

The current article is an example of such a multiplistic approach applied to the question of whether or not a given sales program was effective. The present approach may also be useful to the theoretical researcher who wishes to do research in a field setting and is constrained by the practical considerations of the milieu (Higginbotham et al., forthcoming).

PROGRAM DESCRIPTION

The Arizona lottery is a state-run agency that raises revenue for Arizona by selling lottery tickets. The lottery contributes approximately $22 million each year to Arizona for the improvement of public transportation. This money is distributed to cities and towns throughout Arizona, proportionally according to population.

During 1983 and 1984 the Arizona lottery held a campaign called "Ask for the Sale" to promote sales of its instant game tickets.[2] In this program, signs that read "Did we ask you if you want a Lottery ticket? If not, you get one free" were placed near the point of sale in business establishments. Employees were then instructed to ask customers if they would like to purchase a lottery ticket. Each customer who was not asked and who then demanded a free ticket was given one.

The program was fairly widely implemented among various retailers and involved promotional and free ticket costs to the lottery and to participating retailers. The Arizona lottery agency requested an eval-

uation to determine the effectiveness of this program at increasing ticket sales.

EVALUATION OVERVIEW

In order to assess the effectiveness of this campaign, four separate quasi-experimental designs were utilized. The evaluation was conducted using archival sales data; the rule used to assign stores to the treatment or control groups was unknown and presumed to be nonrandom. Following the theory of quasi experimentation outlined by Cook (1983), we sought to identify plausible alternative hypotheses that could account for apparent program effects and then probed these hypothesis using the four designs. Each design was capable of eliminating certain specific alternative hypotheses. After all four designs have been presented, the results will be summarized and the status of each of the plausible alternative hypotheses for the treatment effect will be assessed.

The emphasis of this evaluation is clearly on *internal validity*. That is, we sought to determine whether the "Ask for the Sale" program was *causally* related to the observed increase in lottery ticket sales in those stores that participated. In order to establish the causal linkage, we have outlined plausible alternative explanations for why an increase in sales might have been observed in the treatment stores. The plausible alternative explanations center around the four basic threats to internal validity that may be present in non-equivalent control group designs (Cook and Campbell, 1979; West, 1985). These four threats are (a) *selection by maturation*, (b) *instrumentation*, (c) *differential statistical regression*, and (d) *local history*.

(a) Selection by maturation typically occurs when persons in one group are maturing or developing at a faster rate than persons in another group. For example, personnel in the treatment condition may have been improving faster in their ability to sell *any* product (including lottery tickets) prior to the beginning of the "Ask for the Sale" campaign than personnel in the control condition.

(b) Instrumentation effects may occur when the scales used to measure the dependent variable have different properties in the two groups. Although Cook and Campbell (1979) outline a variety of problems, such as differential reliability, validity, and scale usage that may produce instrumentation effects, audited sales and many other forms of economic data are typically not subject to most of these problems (see Campbell and Boruch, 1975). The major potential

instrumentation problem with lottery sales data is that of ceiling effects. For example, if salespersons in stores in the treatment program were already selling close to the maximum possible number of lottery tickets, improvements in their sales ability as a result of the program might not be detected, leading to the incorrect conclusion that the program was not effective.

(c) Differential statistical regression is most likely to occur when the experimental units are classified into treatment groups based on their performance on a pretest or correlate of a pretest score. If this selection occurs on the basis of an unreliable measure, groups selected because of their high scores on the measure will obtain lower scores on the posttest, whereas groups selected because of their low scores on the measure will obtain higher scores on the posttest in the absence of any treatment effect. In the present design, stores may have been selected into treatment conditions based on an unreliable measure of prior lottery ticket sales, such as (a) the fallible opinion of management that these stores would have the potential for a high volume of ticket sales or (b) data from a single week in which ticket sales in these stores were unusually low.

(d) Finally, local history occurs when some causal agent that is unrelated to the treatment operates at the same time as the treatment to produce an increase in sales in the treatment stores or a decrease in sales in the control stores. For example, if an increase in foot traffic occurred in the treatment stores at the same time as the "Ask for the Sale" program, any change in ticket sales could not be confidently attributed to the Ask for the Sale program.

DESIGN 1

METHOD

Design overview. In the terminology of Cook and Campbell (1979), Design 1 is an Untreated Control Group Design with Pretest and Posttest. That is, it is the basic nonequivalent control group design. The end-of-game market share was recorded for each participating outlet for the Arizona lottery's instant games 10 and 11. The program was implemented during game 11. Therefore, game ten marketshare serves as the pretest and game eleven marketshare serves as the posttest.

In the analyses for this design, marketshare and not raw ticket sales was utilized as the dependent measure. This measure corrects for length of game and number of tickets distributed during a specific game. The marketshare figure is arrived at by dividing the raw ticket sales of a particular store by the total number of tickets sold in the state for that game. Several million tickets are sold for each game through thousands of retail outlets. The marketshare for the treatment stores was adjusted to remove tickets given away as part of the "Ask for the Sale" program. These tickets constituted an element in the sales program and could not be considered an outcome of the program.[3]

Contribution of design. The basic nonequivalent control group design is open to the four threats to internal validity (selection by maturation, differential statistical regression, instrumentation, and local history) identified above. This design is weak but can often be constructed when the researcher has minimal control over the design of an evaluation or study in the applied environment. One form of the local history threat, differential implementation of other sales training programs in the treatment and control stores, was directly checked. Assessment of the plausibility of each threat depends in large measure on the pattern of the results that are obtained (West, 1985).

Participants. Forty-four chain convenience stores participated in the program during instant lottery game 11, which lasted for a period of 10 weeks. Of the 44 chain outlets, 34 were retailers from one chain (chain A) and 10 were retailers from another chain (chain B).

Matching of control stores. Control stores were selected by matching each treatment store with a store in the *same chain* that had not received the treatment and that was closely similar on game 10 (pretest) marketshare. When multiple matches were possible, store zip codes were utilized as a second matching variable. Zip codes were utilized to help equate treatment and comparison stores on socioeconomic variables.

RESULTS

A t-test was performed on the combined chain retailer pretest scores comparing the treatment stores to the control stores. This analysis indicated that the treatment and control stores did not differ significantly in marketshare prior to the implementation of the program, $t(1, 64) =$

.02, ns. An analysis of covariance (ANCOVA) was then conducted on the combined chain store data using marketshare for game 11 as the dependent variable and marketshare for game 10 as the covariate. The ANCOVA indicated a significant treatment effect for the combined chain stores, $F(1, 63) = 34.42$, $p < .0001$, with the treatment stores showing an increase in sales over the control stores. This effect was also found in both chains: chain A, $F(1, 43) = 18.59$, $p < .0001$ (adjusted $M_t = .00052$, $M_c = .00037$); chain B, $F(1, 17) = 11.62$, $p < .003$ (adjusted $M_t = .00067$, $M_c = .00044$). These results are depicted for chain A, the larger of the two chains, in Figure 1. The analyses indicated that the mean marketshare was higher in the treatment stores than control stores during the same game in which the program was implemented.

In their discussions of analyses of the basic nonequivalent control group design, most authors (e.g., Judd and Kenny, 1981; Reichardt, 1979; West, 1985) recommend that additional sets of analyses be performed to probe two potential artifacts for which the simple ANCOVA does not correct. First, ANCOVA underadjusts for pretest differences in the groups if there is unreliability in the pretest measure (Campbell et al., 1970; Reichardt, 1979). Consequently, a second ANCOVA using Porter's true score correction (Huitema, 1980) is normally recommended to avoid biased treatment estimates.[4] Second, a common form of the selection by maturation threat ("fan spread" growth model) can lead to spurious estimates of treatment effects. This problem occurs when there are pretest differences between the treatment and control groups and differences in rates of maturation are a function of the group's level on the pretest (Kenny, 1975, 1979). Standardized gain score analysis (Huitema, 1980; Kenny, 1975) is normally recommended to correct for this potential problem. Convergence of the results of this set of analyses rules out many, but not all (see Bryk and Weisberg, 1977; Judd and Kenny, 1981), forms of bias in the estimation of the treatment effect.

In the present design, the inclusion of two design features led to the outcome that the results of the additional analyses would not differ appreciably from the results of the basic ANCOVA. (a) Treatment and comparison stores were matched on pretest marketshare so there were no pretest differences between groups in marketshare. (b) The reliability of sales data audited by both the stores and the Arizona lottery was extremely high. These design features minimized the informativeness of the additional analyses in the present case.

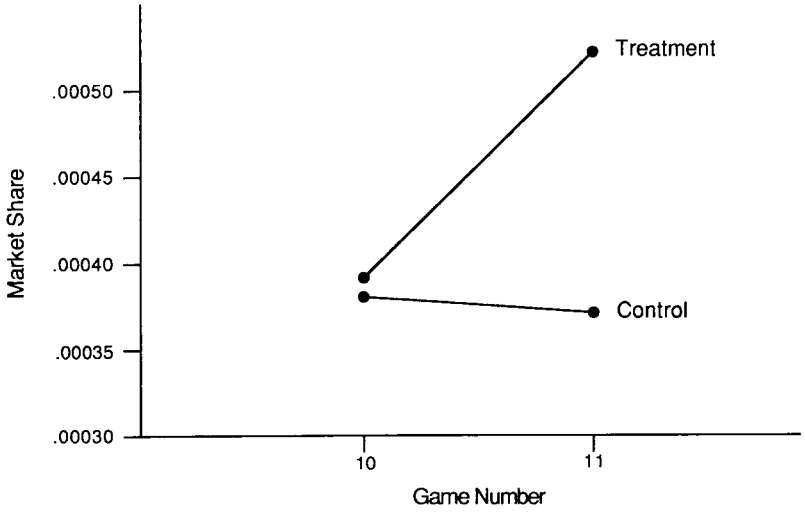

Figure 1: Mean Marketshare for Treatment and Control Stores in Chain A
Note: The program was implemented in treatment stores during game 11.

DESIGN 2

METHOD

The second study utilized a Nonequivalent Control Dependent Variables design (Cook and Campbell, 1979). That is, a design was constructed in which several dependent variables were measured in addition to lottery sales and in which no separate control group was constructed. A chain of gasoline stations that also sold food and other small items implemented the "Ask for the Sale" program in 20 outlets for one month. Sales data were available on five product categories (gasoline, taxable groceries, cigarettes, nontaxable groceries, and lottery tickets) one month prior to and one month after the start of the program. Comparison of the effect of the "Ask for the Sale" campaign on lottery ticket sales with changes in sales for several other product

categories over the same period provided partial checks on a number of threats to internal validity. In particular, forms of the (local) history threat related to changes in local economic conditions and increased customer traffic in the stores that participated were examined.

The dependent measure consisted of raw monthly ticket sales rather than marketshare. Raw ticket sales constitute the appropriate dependent measure because marketshare cannot be accurately computed on a monthly basis and because all comparisons were made within a single game, making corrections for ticket distribution and length of game unnecessary. Control stores could not be selected in the present design for two reasons: (a) The entire chain of 20 gasoline stations participated in the "Ask for the Sale" program and (b) no similar chain(s) of stores were available from which control groups could be selected. Attempts to construct, or in a sense to "force" control groups under these conditions would have led to the creation of large systematic differences between the treatment and control groups enhancing the plausibility of threats due to interactions between selection and other threats to internal validity (e.g., selection by regression).

Two specific predictions were made concerning the dependent variables. Sales for lottery tickets were expected to increase significantly following implementation for the "Ask for the Sale" program. In contrast, sales for the remaining four dependent variables were not expected to increase significantly during the same time period.

Contribution of design 2. The present design, though it is quite weak when taken by itself, can strengthen the overall evaluation of the program and assist in making causal interpretations despite the fact that no control groups could be assigned. The four dependent variables of cigarette sales, gasoline sales, taxable grocery sales and nontaxable grocery sales theoretically should not be affected by the implementation of the program. Thus, the four "nonreactive" dependent variables serve as a crude control against which to compare the treatment-related variable of lottery ticket sales. Theoretically, each of the four "nonreactive" dependent variables would be affected by the most plausible threats to internal validity. In particular, these dependent variables should be sensitive to history effects such as changes in local economic conditions, increases in customer traffic and employee sales training programs.

The strength of the present design is enhanced by the number of alternative dependent variables measured and the specificity of a priori predictions concerning the dependent variables. The design is also

strengthened if the reliability of each of the "nonreactive" dependent measures is high and if a large increase in lottery ticket sales is observed relative to changes in sales for the other dependent measures.

RESULTS

The pretest data for each of the dependent variables were transformed to z-scores; the posttest data were transformed to z-scores using the *pretest* mean and standard deviation for each of the dependent variables. A series of univariate and multivariate analyses of variance were then conducted. For the first analysis, the z-scores for premeasures of the four "nonreactive" variables (e.g., gasoline sales) were summed to form a single "nonreactive" premeasure z-score. This procedure was again completed on the "nonreactive" postmeasures. A repeated measures analysis of variance was then conducted to compare the pre-post change in lottery ticket sales against the pre-post change in the combined nonreactive measures. A significant time by treatment interaction was obtained, $F(1, 19) = 40.60, p < .0001$. This test indicated that the lottery ticket sales increased significantly more than sales for the other four dependent variables combined after the program was implemented.

A second MANOVA was conducted using the z-scores for each premeasure variable (e.g., transformed premeasure on gasoline sales) and each postmeasure (e.g., transformed postmeasure on gasoline sales). A significant multivariate effect was obtained, $F_{mult}(5, 15) = 14.19, p < .0001$. Examination of the univariate F tests for change in each variable over time showed that three of the dependent variables exhibited changes in sales between the pretreatment and posttreatment periods. These variables included ticket sales, $F(1, 19) = 51.85, p < .0001$, nontaxable groceries, $F(1, 19) = 5.35\, p < .04$ and gasoline sales, $F(1, 19) = 3.81, p < .07$. As can be seen in Figure 2, nontaxable grocery sales exhibited a significant *decrease* after implementation of the program while gasoline sales showed a marginally significant increase. Lottery ticket sales increased substantially during this same period of time.

A final ANOVA was conducted making a direct comparison between the nonequivalent dependent variable showing the largest increase, gasoline sales, and lottery ticket sales. It showed that ticket sales did increase significantly more than gasoline sales during the treatment period, $F(1, 18) = 24.90, p < .001$.

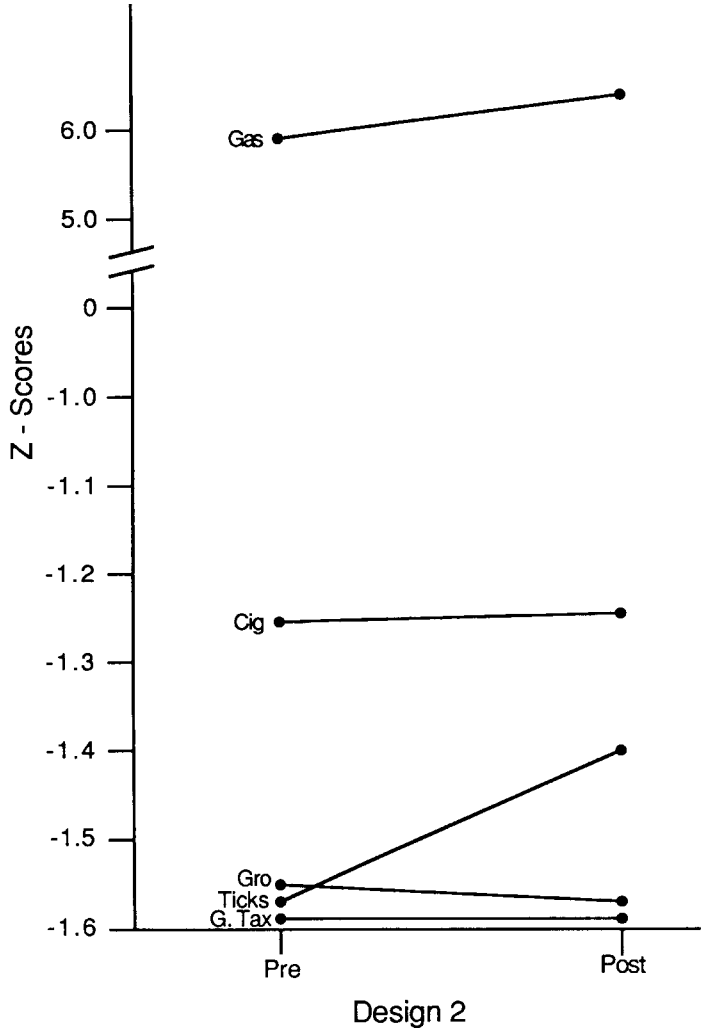

Figure 2: Pre-Post Change in Sales for Each Dependent Variable in Design 2
Note: Department variables are gasoline sales (Gas), cigarette sales (Cig), non-taxable groceries (Gro), lottery ticket sales (Ticks), and taxable groceries (G. Tax).

DESIGN 3

METHOD

Design 3 is a Short Multiple Group Time Series. This design is similar to the basic nonequivalent control groups design utilized in Design 1,

but is enhanced by the inclusion of a brief series of before and after measures on both the treatment and control groups. Eight weekly measurements were collected on the 34 stores that participated in the "Ask for the Sale" program from chain A in Design 1. Thus, Design 3 is a second set of analyses conducted on the same stores that participated from chain A in Design 1. Additional weekly measurements were collected during lottery game 11, and Design 3 was formulated using these measurements. Four of these measurements were taken before the program was implemented and four were collected after the program had started. In addition, 13 control stores were selected by the management of chain A and had weekly sales measures collected at time intervals identical to those during which the treatment stores were measured. The dependent measure consisted of raw weekly ticket sales following the rationale outlined in Design 2.

Contributions of design 3. The threats of selection by maturation and differential statistical regression are both specifically addressed through the use of the extended preseries and postseries measurements. These measurements allow us to statistically test for these two threats and possibly to eliminate them as alternative explanations for any observed effect. Instrumentation and local history are both difficult to eliminate using this design. Instrumentation is plausible if the stores in the control, but not in the treatment group, were selling tickets at the maximum possible level prior to the implementation of the program. This threat may be addressed by examining the data to determine how consistent the hypothesis that a ceiling effect occurred only in the control group is with the obtained pattern of results. Local history must be checked by obtaining additional information on variables that might lead to increases in sales independently of the program, such as sales training programs and increases in foot traffic. Many plausible local history threats are possible, making it difficult to eliminate this threat with certainty.

RESULTS

The data were analyzed following a general strategy outlined by Algina and Swaminathan (1979) for the analysis of the multiple group time-series design.[5] First, we tested the overall treatment by time interaction for the full 8 week series, finding a significant effect, $F_{mult}(7, 39) = 4.06, p < .002$. We then sought to determine where in the

series the deviations from parallelism occurred. Analysis of the data for the 4 weeks prior to program implementation indicated no significant effect of treatment, $F(1, 45) = 0.06$, ns, nor a significant treatment by time interaction, $F_{mult}(3, 43) = 0.72$, ns. Analysis of the data for the 4 weeks following the implementation of the "Ask for the Sale" program indicated a significant effect of treatment, $F(1, 45) = 21.24$, $p < .0001$; however, the treatment by time interaction did not approach statistical significance $F_{mult}(3, 43) = 0.98$, ns. Examination of the series for the treatment and control group did not suggest additional hypotheses related to differential rates of maturation in the treatment and control groups that should be probed (see Figure 3). Finally, a test of the treatment group by period (pre versus post) interaction was performed in which the data were averaged across the 4 weeks prior to and the 4 weeks following the intervention within each treatment group. This result was significant, $F(1, 45) = 23.30$, $p < .0001$, indicating that the deviation from parallelism of the treatment and control series occurred at the point of implementation of the treatment. This particular pattern of results across the sequence of hypothesis tests allow us to rule out virtually all competing statistical hypotheses concerning differential rates of maturation in the treatment and control groups. Therefore it may be concluded that a significant treatment effect occurred such that treatment stores sold more tickets than control stores while implementing the program.

DESIGN 4

METHOD

Design 4 combines both the Removed Treatment and the Repeated Treatment designs (see Cook and Campbell, 1979). The present design uses three separate groups of stores that implemented and removed the "Ask for the Sale" program during three different instant lottery games. In addition, an extended preseries of measurements has been included for each set of stores to enhance the strength of the design. Each separate group of stores has been matched with a corresponding control group of stores that did not receive the program. These stores were matched on the basis of sales in the game just prior to the initiation of the "Ask for the Sale" program. Group 1 received the program during game 13; the

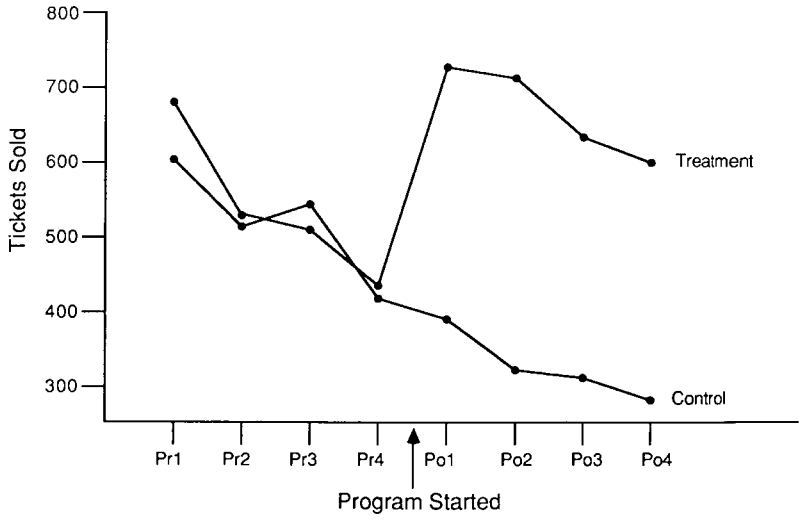

Figure 3: Mean Weekly Ticket Sales Pre- and Posttreatment
Note: Raw mean weekly ticket sales for the 4 weeks prior to treatment (Pr1 – Pr4) and the 4 weeks after treatment (Po1 – Po4) are plotted for treatment (T) and control (C) stores.

program was removed during game 14. Group 2 initiated the program during game 13 and continued implementation through the end of game 14. Group 3 did not implement the program until the beginning of game 14. Final sales measurements were taken at the completion of game 14.

All stores came from a single large chain of convenience stores that had not previously participated in the "Ask for the Sale" program. Group 1 consisted of 43 treatment stores, Group 2 had 21 treatment stores and Group 3 had 90 treatment stores. Market share was utilized as the dependent variable and was available for each of the games in which stores participated.

Contributions of design 4. As with the other nonequivalent control group designs presented earlier, selection by maturation, selection by instrumentation, selection by history and selection by statistical regression are the primary threats to internal validity in the present design. However, each of these threats is diminished through the use of several independent groups and the statement of explicit and complex predictions for the performance of these groups. That is, if the program is

effective at increasing sales, a specific a priori pattern of sales performance can be predicted for these stores. It is highly unlikely that a group of threats to internal validity would combine to produce the predicted set of results. We are, in effect, making a complex set of predictions based on theoretical knowledge of the program and past empirical evidence. In addition, the extended series of premeasures substantially increases our ability to eliminate selection by maturation and selection by regression as plausible threats.

RESULTS

For ease of presentation, three sets of analyses were completed, one for stores that only participated during game 13 (Group 1), another for stores that participated during both game 13 and game 14 (Group 2), and one set for stores that participated only during game 14 (Group 3). Figure 4 presents the mean marketshare in games 8-14 for treatment and control stores in each of the three groups.

Group 1. Treatment and control stores were compared on game 12 marketshare to determine whether these two groups differed in sales prior to the implementation of the program. No significant difference was found between treatment and control stores, $t(1, 84) = .11$, ns. Visual inspection of the preseries shows that treatment and control stores closely parallel each other in sales prior to the start of the game. In addition, the control stores showed a decrease in sales during both games 13 and 14. A repeated measures MANOVA conducted on the pretest series showed no significant effect of treatment, $F(1, 84) = 0.15$, ns., nor a significant treatment by time interaction, $F_{mult}(4, 81) = 0.75$, ns.

An analysis of covariance was then conducted using the game 13 marketshare as the dependent variable and game 12 marketshare as the covariate. Game 13 marketshare reflects sales for the game during which group 1 stores participated in the "Ask for the Sale" program. The treatment effect was significant, $F(1, 83) = 83.84$ $p < .0001$. Sales increased substantially during game 13 with mean adjusted marketshare being higher for treatment stores (adjusted $M = .00087$) than for control stores (adjusted $M = .00053$).

A second ANCOVA was run on the group 1 stores using marketshare for game 14 as the dependent variable. This analysis permits an examination of the carryover effects from the program. The analysis showed that a carryover effect did occur, although the increase in sales

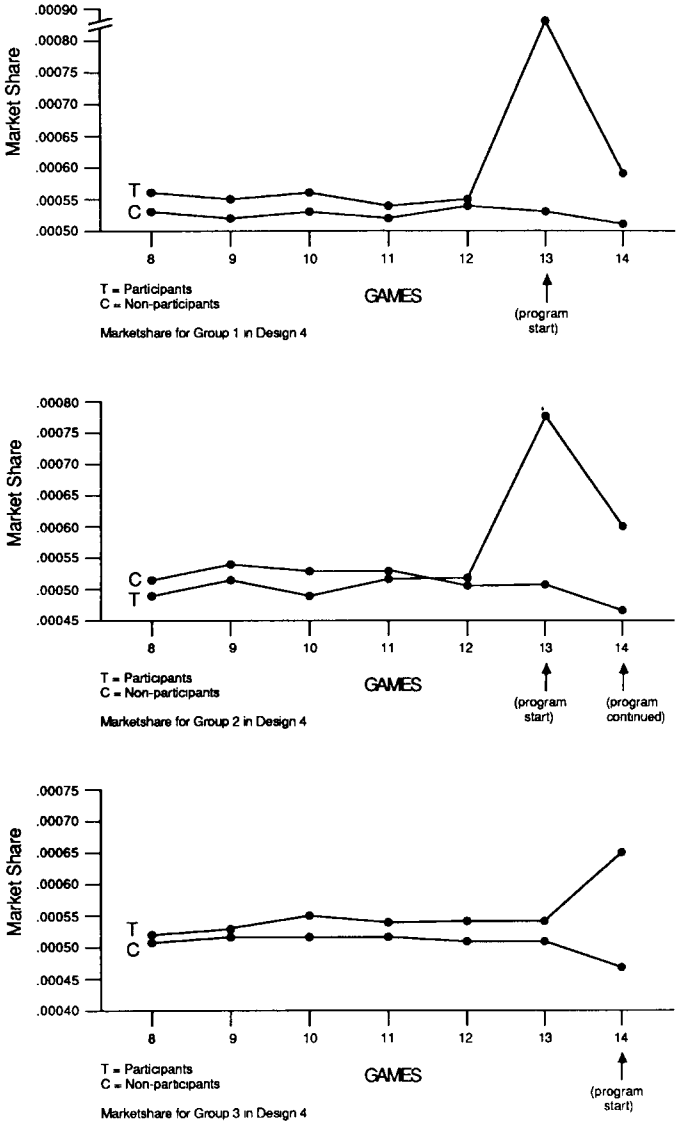

Figure 4: Mean Marketshare for Treatment and Corresponding Control Stores in Group 1, Group 2, and Group 3

Note: Treatment stores in Group 1 began the program during game 13 and discontinued the program during game 14. Treatment stores in Group 2 began the program during game 13 and continued during game 14. Treatment stores in Group 3 began the program during game 14.

experienced by treatment stores was greatly reduced for game 14, $F(1, 83) = 7.77, p < .007$. Once again sales for treatment stores (adjusted $M = .00059$) exceeded those for control stores (adjusted $M = .00051$).

Group 2. Treatment and control stores were compared on marketshare for game 12 to determine whether these stores differed in sales prior to implementation of the program. This test showed no significant difference in sales on game 12, $t(1,40) = .03$, ns. Repeated measures MANOVA on the preseries measures also indicated that neither the treatment main effect, $F(1,40) = 0.13$, ns., nor the treatment by time interaction, $F_{mult}(4, 37) = 0.54$, ns., were significant prior to treatment. Consistent with these results, visual examination of both the treatment group and the control group preseries measurements indicated that the sales in the two groups were highly similar prior to treatment.

An ANCOVA was conducted using marketshare for game 13 as the dependent variable. This analysis showed a significant difference in sales between treatment and control stores, $F(1, 39) = 44.00, p < .0001$. The mean adjusted marketshare for treatment stores (adjusted $M = .00078$) was higher than for control stores (adjusted $M = .00051$).

A second ANCOVA was run using marketshare for game 14 as the dependent variable. Since the program was implemented in both games 13 and 14 for group 2, this analysis provides a test of the program for the second continuous game in which it was utilized. The treatment and control stores significantly differed during this second game, $F(1, 39) = 13.39, p < .001$. The effect is less striking relative to the differences seen for this group during game 13. Mean adjusted marketshare for treatment stores (adjusted $M = .00060$) was higher than mean adjusted marketshare for control stores (adjusted $M = .00047$).

Group 3. A t-test conducted on the game 13 marketshare indicated that treatment and control stores did not differ in sales prior to the implementation of the program, $t(1, 178) = .72$, ns. A repeated measures MANOVA comparing the treatment and control preseries and indicated that the two sets of premeasures did not differ: treatment main effect, $F(1, 177) = 0.37$, ns.; treatment by time interaction, $F\text{mult}(5, 173) = 0.26$, ns.

An ANCOVA was performed using marketshare for game 14 as the dependent variable. This analysis showed that treatment and control stores differed significantly in sales during game 14, $F(1, 177) = 19.44 p < .001$, with treatment stores (adjusted $M = .00065$) selling significantly more tickets than control stores (adjusted $M = .00047$).

Visual examination of the preseries shows a close parallel in

marketshare between treatment and control stores through game 13. The control series decreased from game 13 to game 14, whereas the treatment series increased.

For purposes of illustration, the sales patterns of the three groups of treatment stores are presented together in Figure 5. It can be seen that, as predicted, sales increased in game 13 for both of the groups that initiated the program during this game while sales increased in the third group only during game 14.

GENERAL DISCUSSION

The present set of designs converge on the result that the "Ask for the Sale" program has been effective at increasing sales in those stores that have participated. There is some indication, however, that this effect does not persist without the continued implementation of the program. In addition, there is evidence that the program loses strength over time even if it is implemented continuously.

In Design 1, the use of audited sales data ruled out all but one form of the instrumentation threat. The remaining form of this threat, ceiling effects, is unlikely given the close match of treatment and control stores on game 11 sales. The most common forms of the selection by maturation threat are also implausible in light of the pattern of obtained results in which the treatment stores increased their sales, whereas the control stores showed a slight decrease (Cook and Campbell, 1979: 104-105). The high reliability of the sales measure decreases the likelihood of differential statistical regression. The use of a geographically dispersed sample of stores diminished the plausibility of most local history threats; one form of this threat, differential implementation of sales training programs, was also checked directly. Nonetheless, other forms of the local history threat (e.g., rate of sales force turnover) remain viable.

The primary contribution of Design 2 is that it addresses forms of the history threat remaining after Design 1. The robustness of the increase in lottery ticket sales relative to the changes in sales of a number of other product categories together with the high reliability of the "nonreactive" dependent measures (test-retest r_{xx} = .96 to .99) combine to render implausible nearly all forms of the history threat (e.g., employee

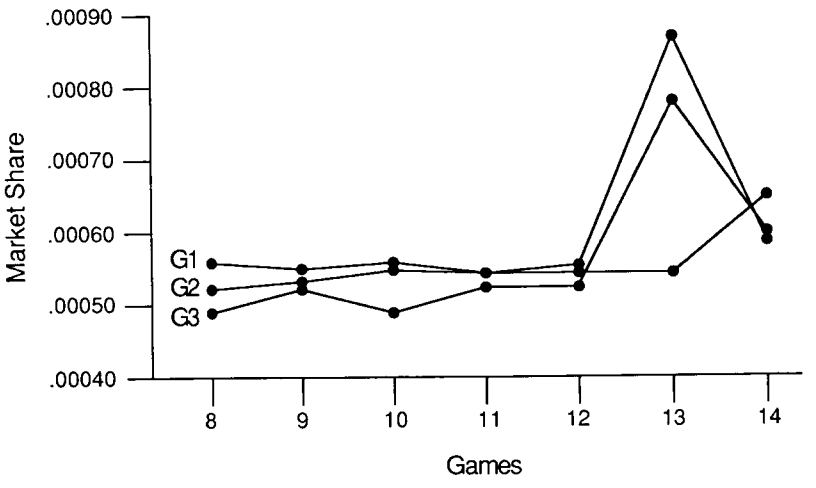

Figure 5: Mean Marketshare for Treatment Stores for Group 1, Group 2 and Group 3
Note: Group 1 (G1) began the program during game 13 and discontinued the program during game 14. Group 2 (G2) began the program during game 13 and continued during game 14. Group 3 (G3) began the program during game 14. Only the treatment stores are included.

turnover, increase in customer traffic). The high temporal reliability of the sales data and the use of the entire chain of stores provides further evidence against the threat of statistical regression.

Design 3 directly addressed the threats of selection by maturation and differential statistical regression through the use of an extended pretreatment and posttreatment series of observations. The Algina-Swaminathan (1979) procedure probed all forms of the selection by maturation threat including those not addressed by Design 1, resulting in the conclusion that a treatment effect exists. The comparability of the series for the control and treatment groups prior to the implementation of the program clearly rules out the threat of differential statistical regression. The declines in lottery ticket sales in both groups over the weeks prior to the beginning of the program clearly indicated that the control group was not at a ceiling level of lottery ticket sales.

Design 4 featured an extended preseries to address the threats of selection by maturation and differential statistical regression and replication of the program effect across different groups of stores and different games to address local history threats. This is the strongest design and provides strong rejection of virtually all alternative hypotheses for the treatment effect. Nearly all forms of each of the four threats to internal validity were ruled out prior to the implementation of Design 4; this final design provided substantiation of the treatment effect and provides an illustration of one of the strongest possible designs under these circumstances.

The use of multiple designs provides replication of the treatment effect over a series of different stores and times of implementation. This combination of replications makes it very implausible that an alternative causal process would covary with the treatment across so many different circumstances. That is, to be a viable alternative explanation, a causal process other than the treatment would need to covary with the treatment and exert its influence *in the same direction* despite the many different biases that operate in these different circumstances. It is extremely unlikely that these many different circumstances share such a systematic bias.

In summary, through the use of multiple designs that do not share a common bias we have been able to eliminate nearly all of the plausible threats to internal validity and can conclude with some certainty that the "Ask for the Sale" program does produce increases in sales. This increase may not extend over a long period of time given the decrease in the program's effectiveness observed in Design 4. The reasons for this decrease cannot be conclusively determined with the present data. However, a decay in the program's effectiveness might be expected as customers become habituated to the continual requests for sales. In addition, the program requires a certain degree of supervision to maintain, and the integrity of the program may decrease over time as managers tire of enforcing the program among their employees.

NOTES

1. The structural equation approach attempts to develop a causal model of selection process using multiple indicators of each construct (e.g., Magidson, 1977). If the model of

the selection process is correct, then treatment effects can be estimated using structural equation modeling techniques with a high degree of precision. The econometric modeling approach attempts to estimate treatment group assignment as a function of pretest scores (e.g., Barnow et al., 1980). Such models achieve identification by estimating probit models of group assignment or by assuming other restrictions on the group assignment model. Both approaches make strong statistical assumptions. For example, the econometric approach of Barnow et al. (1980) assumes bivariate normality of the error terms, correct specification of the selection model, appropriate selection of identifying restrictions, and no serious problems with multicollinearity in estimating the second-step regression equation. The effects of the assumptions on the estimate of the treatment effect need to be carefully probed to the extent possible before accepting the conclusions of these analyses. When the assumptions of these models are not met, serious misestimation of treatment effects can result (see Bentler and Woodward, 1978; Murnane et al., 1985, for examples of this problem). Typically most problematic is the assumption that an unknown selection process has been correctly modeled. However, when the assumptions of these models are met they can provide highly efficient and unbiased estimates of treatment effects.

2. An instant lottery game involves the sales of lottery tickets over a period of 8 to 15 weeks. In the instant game, all of the information needed to play and win is printed on each ticket. Thus players find out immediately whether they have won and the amount they have won.

3. The use of marketshare figures introduces a negative correlation between datapoints, since they are constrained to sum to 1.00 across all outlets in the state. Given the very large number of lottery ticket outlets in the state, the magnitude of the bias resulting from nonindependence in the present data is trivial. In cases in which marketshare is divided among a small number of businesses this nonindependence would be of substantial magnitude, so that statistical adjustments for this problem would need to be made (see Kenny and Judd 1986).

4. Controversy exists over the most appropriate estimate of reliability with which to adjust for pretest differences between groups (e.g., Campbell and Boruch, 1975; Linn and Werts, 1973). Judd and Kenny (1981) and West (1985) recommend that two sets of Porter true-score analyses be performed, one using a lower-bound estimate and one using an upperbound estimate of the reliability.

5. Econometric approaches (Judge et al., 1982; Kmenta, 1986; Simonton, 1977) may also be taken to what is termed in that literature the cross-sectional time series design. In addition to the assumption of equal variance-covariance matrices in the two groups that must be made by the Algina-Swaminathan (1979) approach, econometric approaches make specific structural assumptions about the variance-covariance matrix. More specifically, the same autoregressive error structure, typically of order 1, is assumed to apply to all within-case disturbances, with the autoregressive parameter being assumed to be identical for all cases (Simonton, 1977: 499-500). When the assumptions of the econometric approach are met, it yields a powerful test of treatment effects. Unfortunately, for the short time series under consideration here, adequate tests of the structural assumptions of the econometric approach are not possible. Relatively little work has been done on the effects of violations of these structural assumptions, so that the extent and direction of bias in the estimation of the treatment effects when the assumptions are not met cannot at present be generally predicted.

REFERENCES

ALGINA, J. and H. SWAMINATHAN (1979) "Alternatives to Simonton's analyses of the interrupted and multiple-group time-series designs." Psych. Bull. 86: 919-926.

BARNOW, B. S., G. G. CAIN, and A. S. GOLDBERGER (1980) "Issues in the analysis of selectivity bias," in E. S. Stromsdorfer and G. Farkas (eds.) Evaluation Studies Review Annual, Vol. 5. Newbury Park, CA: Sage.

BENTLER, P. M. and J. A. WOODWARD (1978) "Head Start reevaluation: Positive effects are not yet demonstrable." Evaluation Rev. 2: 493-510.

BOOMSMA, A. (1982). "The robustness of LISREL against small sample sizes in factor analysis," in K. G. Jöreskog and H. Wold (eds.), Systems under Indirect Observation, Part 1. Amsterdam: North Holland.

BRYK, A. S. and H. I. WEISBERG (1977) "Use of the nonequivalent control group design when subjects are growing." Psych. Bull. 85: 950-962.

CAMPBELL, D. T. and R. F. BORUCH (1975) "Making the case for randomized assignment to treatments by considering the alternatives: Six ways in which quasi-experimental evaluations in compensatory education tend to underestimate effects," in C. A. Bennett and A. A. Lumsdaine (eds.) Evaluation and Experiment: Some Critical Issues in Assessing Social Programs. New York: Academic Press.

CAMPBELL, D. T., R. F. BORUCH, and A. E. ERLEBACHER (1970) "How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful," in J. Hellmuth (ed.) Compensatory Education: A National Debate, Vol. 3, Disadvantaged Child. New York: Brunner/Mazel.

COOK, T. D. (1983) "Quasi-experimentation: its ontology, epistemology, and methodology," in G. Morgan (ed.) Beyond Method: Strategies for Social Research. Newbury Park, CA: Sage.

COOK, T. D. (1985) "Post-positivist critical multiplism," in L. Shotland and M. M. Mark (eds.) Social Science and Social Policy. Newbury Park, CA: Sage.

COOK, T. D. and D. T. CAMPBELL (1979) Quasi-experimentation: Design and Analysis Issues for Field Settings. Boston: Houghton-Mifflin.

DWYER, J. H. (1984) "The excluded variable problem in nonrandomized control group designs." Evaluation Rev. 8: 559-572.

HECKMAN, J. J. (1979) "Sample selection bias as specification error." Econometrica 47: 153-161.

HIGGINBOTHAM, H. N., S. G. WEST, and D. R. FORSYTH (forthcoming) Psychotherapy and Behavior Change: Social, Cultural, and Methodological Perspectives. New York: Pergamon.

HOUTS, A. C., T. D. COOK, and W. R. SHADISH (1986) "The person-situation debate: a critical multiplist perspective." J. of Personality 54: 52-105.

HUITEMA, B. E. (1980) The analysis of covariance and alternatives. New York: John Wiley.

JUDD C. M. and D. A. KENNY (1981) Estimating the Effects of Social Interventions. New York: Cambridge Univ. Press.

JUDGE, G. G., R. C. HILL, W. E. GRIFFITHS, H. LUETKEPOHL, and T-C LEE (1982) Introduction to the Theory and Practice of Econometrics. New York: McGraw-Hill.

KENNY, D. A. (1975) "A quasi-experimental approach to assessing treatment effects in the nonequivalent control group design." Psych. Bull. 82: 345-362.
KENNY, D. A. (1979) Correlation and Causality. New York: Wiley-Interscience.
KENNY, D. A. and C. M. JUDD (1986) "The consequences of violating the independence assumption in analysis of variance." Psych. Bull. 99: 422-431.
KMENTA, J. (1986) Elements of Econometrics (2nd ed.) New York: Macmillan.
LINN, R. L. and C. E. WERTS (1973) "Errors of inference due to errors of movement." Educational and Psych. Measurement 33: 531-545.
LIPSEY, M. W., D. S. CORDRAY, and D. E. BERGER (1981) "Evaluation of a juvenile diversion program: using multiple lines of evidence." Evaluation Rev. 5: 283-306.
MAGIDSON, J. (1977) "Toward a causal model approach for adjusting for preexisting differences in the nonequivalent control group situation: a general alternative to ANCOVA." Evaluation Rev. 1: 399-420.
MAGIDSON, J. and D. SÖRBOM (1980) "Adjusting for confounding factors in quasi-experiments: another reanalysis of the Westinghouse Head Start Evaluation." Presented at the American Statistical Association meetings, Houston, TX, August.
MURNANE, R. J., S. NEWSTEAD, and R. J. OLSON (1985) "Comparing public and private schools: the puzzling role of selectivity bias." J. of Business and Econ. Statistics 3: 23-35.
MUTHÉN, B. and K. G. JÖRESKOG (1983) "Selectivity problems in quasi-experimental studies." Evaluation Rev. 7: 139-174.
REICHARDT, C. S. (1979) "The statistical analysis of data from nonequivalent group designs," in T. D. Cook and D. T. Campbell (eds.) Quasi-Experimentation: Design and Analysis Issues for Field Settings. Boston: Houghton Mifflin, 1979.
SHADISH, W. R., Jr., COOK, T. D., and HOUTS, A. C. (1986) "Quasi-experimentation in a critical multiplist mode," in W.M.K. Trochim (ed.) Advances in Quasi-Experimental Design and Analysis. San Fransisco: Jossey-Bass.
SIMONTON, D. K. (1977) "Cross-sectional time-series experiments: some suggested analyses." Psych. Bull. 84: 489-502.
SÖRBOM, D. (1978) "An alternative to the methodology for analysis of covariance." Psychometrika 43: 381-396.
SÖRBOM, D. (1981) "Structural equation models with structured means," in K. G. Jöreskog and H. Wold (eds.) Systems under Indirect Observation: Part 1. Causalty, Structure, Prediction. Amsterdam: North-Holland.
TUKEY, J. W. (1986) "Sunset salvo." Amer. Statistician 40: 72-76.
WEST, S. G. (1985) "Beyond the laboratory experiment: experimental and quasi-experimental designs for interventions in naturalistic settings," in P. Karoly (ed.) Measurement Strategies in Health Psychology. New York: John Wiley.

Kim D. Reynolds is currently Postdoctoral Research Associate at the Center for Research in Disease Prevention, Stanford University School of Medicine. His primary research interests focus on the design and evaluation of theory-based prevention programs in the areas of cancer, heart disease, and childhood psychopathology.

Stephen G. West is Professor and Director of the Graduate Training Program in Social Psychology at Arizona State University, currently on sabbatical leave at UCLA. He is

coeditor of Evaluation Studies Review Annual *(Volume 4) and editor of a special issue of the Journal of Personality on methodological developments in personality research. His primary research interests are in research methodology and applied social psychology.*