Psychological Bulletin
1988, Vol. 104, No. 1, 150–154

0033-2909/88/$00.75

Statistical Treatment of the Solomon Four-Group Design: A Meta-Analytic Approach

Mary C. Walton Braver and Sanford L. Braver
Arizona State University

One of the causes of the underuse of the Solomon four-group design may be that the complete details for the statistical analysis have not previously been presented. The primary issue previously unaddressed was how to combine an analysis of the effect of the treatment in the posttest-only groups with the same effect in the pre- and posttest groups (after an earlier phase of the analysis has shown no evidence of pretest sensitization.) A meta-analytic solution for this problem is proposed, and the entire analysis is presented, complete with flowchart and example. It is shown that the analysis has adequate statistical power even if the total N is not increased from that of a posttest-only design, removing the last of the serious objections to the Solomon design.

Almost 40 years ago, Solomon introduced a new form of experimental design typically referred to today as the Solomon four-group design (Solomon, 1949). Campbell and Stanley (1963) discussed this design as one of three one-treatment condition experimental designs, the other two being the pre- and posttest control group design and the posttest-only control group design (see Table 1). Each of these designs is adequate to assess the effect of the treatment and is immune from most threats to internal validity. The Solomon four-group design, however, adds the advantage of being the only one of the three able to assess the presence of pretest sensitization. Pretest sensitization means that "exposure to the pretest increases . . . the Ss' sensitivity to the experimental treatment, thus preventing generalization of results from the pretested sample to an unpretested population"[1] (Huck & Sandler, 1973, p. 54). Thus, the Solomon four-group design adds a higher degree of external validity in addition to its internal validity, and hence, according to Helmstadter (1970), is therefore "the most desirable of all the . . . basic experimental designs" (p. 110).

Despite its strength, however, the Solomon four-group design is underused. Four reasons may account for this underuse. First, the Solomon four-group design requires twice the number of groups used by the other two designs, a requirement frequently misunderstood to mean that the Solomon design requires twice the number of *subjects* as do the other designs. We will show later in this article that by cutting the size of each group in half, thus retaining the same total sample size as the other designs, one may still enjoy the benefits of the Solomon design. This strategy typically will result in quite adequate statistical power; indeed, the power will be greater than that of the posttest-only control group design.

Second, few investigators have research interest in pretest sensitization effects per se, so they object to using a design for which the strongest advantage is the ability to examine pretest sensitization. This objection misses the point, however. Pretest sensitization, like the experimenter effect, is a potential artifact that could limit the generalizability of effects in which the researcher does have interest, and its existence should be explored despite no direct interest in the phenomenon. Somewhat more to the point may be the researcher's belief that pretest sensitization artifacts probably do not exist in his or her research arena. This belief is supported by several literature reviews (Bracht & Glass, 1968; Lana, 1959; Lana, 1969; Rosnow, 1971; Willson & Putnam, 1982) showing that pretest sensitization artifacts rarely occur. Nevertheless, this artifact should be considered an effect that could potentially jeopardize the external validity of any research finding until and unless the Solomon design has been used to explicitly rule it out.

Third, as Oliver and Berger (1980) have argued, conclusions may become far more complicated using the Solomon design because of the number of comparisons it permits. This complexity may dissuade researchers from using the design. Although such a motivation is understandable, the complexity of a phenomenon or an analysis is certainly not a scientifically justifiable reason to fail to conduct it.

Fourth, and perhaps the most important reason why the Solomon design is underused, is the lack of certainty concerning the proper statistical treatment of this rather complicated design. Campbell and Stanley (1963) made some recommendations,

A preliminary version of this article was presented at the annual meeting of the Western Psychological Association, Seattle, Washington, May 1986.

Correspondence concerning this article should be addressed to Sanford L. Braver, Department of Psychology, Arizona State University, Tempe, Arizona 85287.

[1] In the omitted portion of the preceding quote, Huck and Sandler (1973) included the phrase "(or decreases)". For the purposes of the present article, however, we wish to limit the term and our analysis to the "increased" case, that is, the case where the treatment's effect in the pretested groups is both in the same direction and at least as large as its effect in the unpretested groups.

but their presentation was incomplete. Huck and Sandler (1973) modified Campbell and Stanley's analysis but still left some details unattended. The remainder of this article is devoted to the issue of the proper statistical treatment of the Solomon design. The analysis we recommend is in agreement with the work of Campbell and Stanley and Huck and Sandler but covers more contingencies and has more statistical power (i.e., ability to detect significance). For the sake of completeness, the entire analysis is presented here, even though the first stages were originally presented by the previous authors.

As Campbell and Stanley (1963) pointed out, the initial phase of the analysis is to determine whether evidence of pretest sensitization exists, that is, whether X affects O only when a pretest measure also is administered. If this were the case O_2 would be higher than O_4, but O_5 would not be higher than O_6. The test for this is a 2×2 between-groups analysis of variance (ANOVA) on the four posttest scores, as indicated in Table 2. The factors are treatment (yes vs. no) and pretest (yes vs. no). Evidence demonstrating pretest sensitization is detected by the interaction (referred to henceforth as Test A). In addition, there should be a significant simple effect for treatment in the first row (pretest present—Test B)[2] but not in the second (pretest absent—Test C). If the preceding pattern is present, the analysis terminates with the conclusion that there is evidence of a treatment effect, but it occurs only for pretested groups; there is thus pretest sensitization (a result unlikely to be welcomed by the investigator).

Huck and Sandler (1973) modified Campbell and Stanley's (1963) analysis by noting that if the simple effect in the second row (Test C) is significant also (still assuming a significant interaction in Test A), there is evidence that pretest sensitization is present but that it merely *enhances* the effect of the treatment, which is detectable as well even in an unpretested sample.

If the interaction is not significant, however, we conclude there is no evidence of pretest sensitization. Is there a treatment effect, however?

One answer to that question, suggested by Campbell and Stanley (1963), is found by looking at the main effect for treatment in the above analysis (Test D). If significant, there is unqualified evidence of the treatment effect. This test clearly is not the most powerful available, however, and if not significant it should not be considered conclusive evidence against a treatment effect. This is because Test D disregards the pretest information available for Groups 1 and 2, data that typically increase power substantially.

Table 1
Three One-Treatment Condition Experimental Designs

Design	Group		Pretest	Treatment	Posttest
Solomon four-group	1	R	O_1	X	O_2
	2	R	O_3		O_4
	3	R		X	O_5
	4	R			O_6
Pre- and posttest control group	1	R	O_1	X	O_2
	2	R	O_3		O_4
Posttest-only control group	1	R		X	O_5
	2	R			O_6

Note. O = outcome measure; X = treatment; R = randomization.

Table 2
2 × 2 Analysis of Posttest Scores

	Treatment (X)	
Pretest	Yes	No
Yes	O_2	O_4
No	O_5	O_6

Note. O = outcome measure.

No one has previously suggested how to use these data effectively, however. One approach is to do a separate additional analysis on Groups 1 and 2. The test could be a two-group analysis of covariance (ANCOVA) on the posttest scores, covarying the pretest scores (Test E); a two-group (independent) t test on "gain" (i.e., post minus pre) scores (Test F); or the test of the interaction in a 2×2 ANOVA, with treatment (yes = Group 1 vs. no = Group 2) and time (pre vs. post) as the factors, the second a "repeated measures" factor (Test G). Huck and McLean (1975, p. 513) showed that the interaction in Test G (which is the appropriate test of differential treatment effects) is identical to the outcome of Test F. Test E is usually the preferable one of these three, however, primarily because of its greater power or ability to detect the treatment effect (Cuervorst & Stock, 1978; Huck & McLean, 1975; Humphreys, 1976; Scheifley & Schmidt, 1978).

Although Tests E, F, and G of X's effectiveness are often more powerful than Test D, none is the most powerful available because each omits the data of Groups 3 and 4. If, despite the lack of power, the test is significant, evidence of treatment effects unqualified by pretest sensitization is obtained, and no further testing is necessary. On the other hand, should significance not be found, testing should continue, resorting to the more powerful tests described subsequently.

The test that uses the data of Groups 3 and 4 is an independent t test on O_5 and O_6 (Test H). This is probably the least powerful test of all, however, because it omits all of the data from Groups 1 and 2. Again, if the test is significant despite its lack of power, evidence for unqualified treatment effects is obtained, and further testing may cease. As before, however, should significance not be found, testing continues.

For maximal power, what is needed, but has heretofore not been suggested, is a test that somehow combines the test of X in Groups 1 and 2, with the test of X in Groups 3 and 4. Such a test may not have been suggested previously because it may have

[2] Actually, if the interaction is significant, it can easily be shown that Test B *must* be significant for the case we are assuming, namely, that the treatment's effect in the pretested groups is in the same direction and at least as large as its effect in the groups that were not pretested.

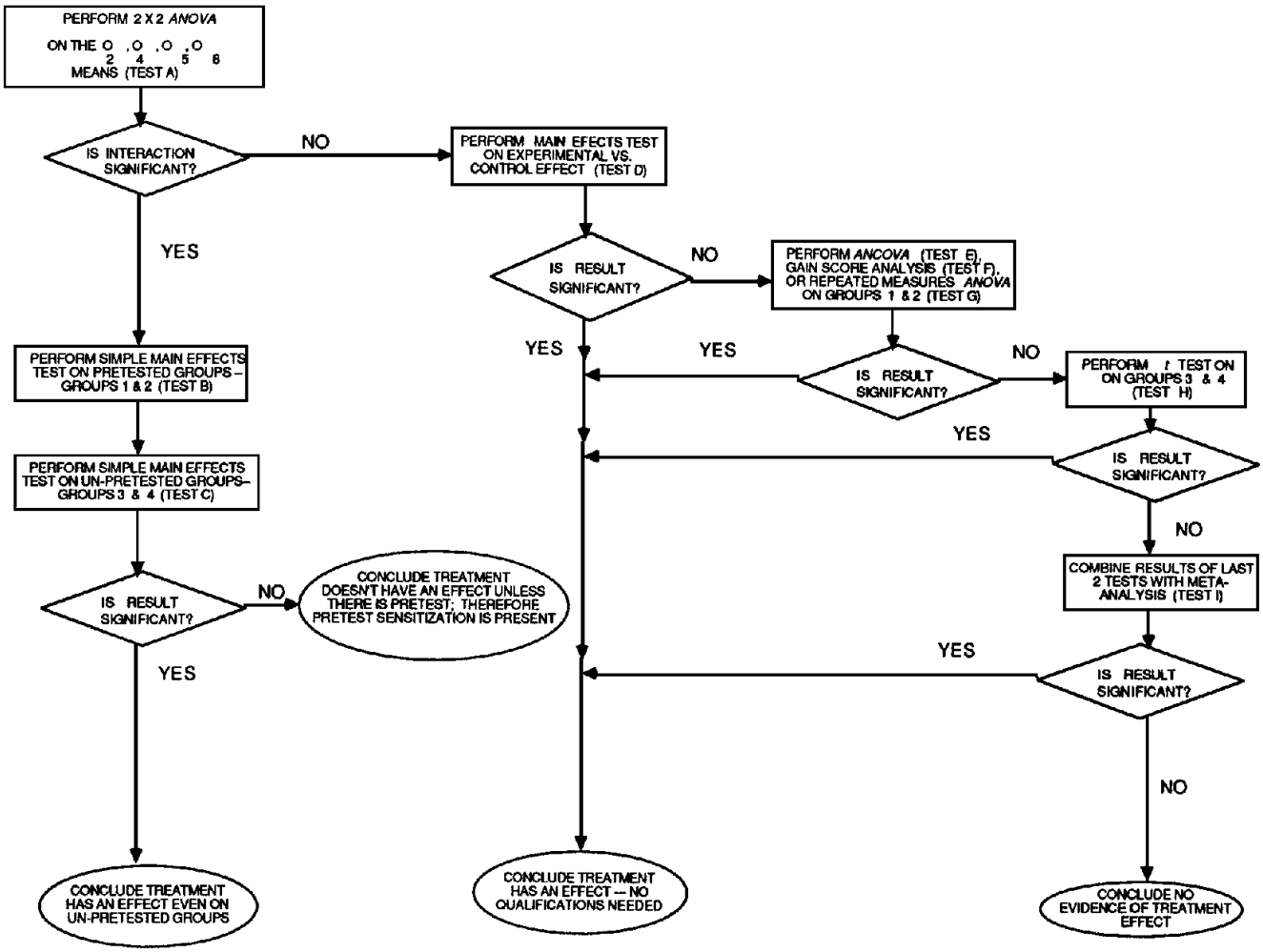

Figure 1. Flowchart of tests and conclusions. (O = outcome measure; ANOVA = analysis of variance; ANCOVA = analysis of covariance.)

been unclear how to combine the results of, say, an ANCOVA on one pair of groups within the experiment with a t test on a different pair of groups. The work on meta-analysis (Glass, 1978; Rosenthal, 1978; Smith & Glass, 1977), however, provides the ready answer.

Meta-analysis demonstrates how the results from disparate, independent tests of the same hypotheses may be statistically combined even when the significance tests arise through different statistical techniques (Rosenthal, 1978). Although meta-analytic techniques are typically applied to the results of many *different* studies, nothing prevents their application to several different tests of the same effect within but one study, as in the case of the Solomon four-group design. A large number of methods that combine test results into one overall test are available and each has advantages and disadvantages. It has been shown that none of these methods is uniformly most powerful, so that under varying conditions, different techniques may be preferred. Rosenthal (1978) provided an excellent review of these methods, and Hedges and Olkin (1985) reviewed the literature concerning the differential power of the techniques. Although several of these would be appropriate in the present circumstance, we describe and recommend Stouffer's z method (Mosteller & Bush, 1954; Stouffer, Suchman, DeVinney, Star, & Williams, 1949) simply because it is the most straightforward. In this approach, the p level from each statistical test is converted to a normal deviate (z) value, and then the resulting zs are combined into a single z_{meta} by the following formula:

$$z_{\text{meta}} = \sum_i z_{pi}/\sqrt{k} \qquad (1)$$

where z_{pi} is the z value corresponding to the one-tailed p value of the ith statistical test and k is the number of such tests. In the present instance, $k = 2$; thus the test reduces to:

$$z_{\text{meta}} = (z_{p1} + z_{p2})/\sqrt{2} \qquad (2)$$

where z_{p1} is the z value corresponding to the one-tailed p value

Table 3
Hypothetical Data for a Solomon Four-Group Design

Group	Pretest		Correlation between pre- and posttest	Posttest	
	M	Variance		*M*	Variance
1	10.5	19.3		12.4	22.0
			.58		
2	10.7	17.2		10.2	16.5
			.62		
3				12.5	19.0
4				10.3	22.5

Note. $N = 14$ per group.

of Test E, F, or G, and z_{p2} is the z value corresponding to the one-tailed p value of Test H. One then refers to a z table for the significance of z_{meta}. We refer to this test as Test I.

Test I allows full use of all data and thus is the most powerful single test of the treatment effect available. A flowchart summarizing the recommended sequence[3] of testing and decision is presented in Figure 1.

An example will serve to make the discussion more concrete. Consider the hypothetical data in Table 3. Note that inspection of the means suggests a treatment effect of X unqualified by pretest sensitization, because the posttest means of Groups 1 and 3, the groups receiving X, are similar to one another and higher than all of the other means. As specified by the flowchart, first we perform a 2 × 2 ANOVA on the posttest scores. The results are displayed in Table 4. There is no interaction whatever (Test A), so we proceed to an examination of the main effect of treatment. This main effect (Test D) is substantial but not quite significant ($p = .0714$) by conventional standards, however. Thus we next need to test the treatment effect in Groups 1 and 2, the two pretested groups. In Table 5 are presented the results of the ANCOVA (Test E), the preferred test of Tests E, F, and G. The ANCOVA is not significant at conventional levels, $p = .0993$. Because of this lack of significance, we proceed to an independent-samples t test (Test H) performed on the scores of Groups 3 and 4, the posttest-only groups. This result, too, does not achieve conventional levels of significance, $t(26) = 1.28$, $p = .2127$. Finally, the meta-analysis (Test I) is performed, which combines the t-test result with the ANCOVA result. The meta-analytic result is significant, $z_{meta} = (1.65 + 1.25)/\sqrt{2} = 2.05$, $p = .040$.

These hypothetical data show the superior power of the meta-analytic technique, because none of the customary analyses reached significance but the meta-analysis did. Thus, when the data conformed completely to the pattern predicted by a treatment effect uncontaminated by pretest sensitization, only the meta-analysis reached significance in detecting that fact.

Another demonstration of the power of the technique may also be seen in this example. As mentioned earlier, many researchers are reluctant to use the Solomon design because of an erroneous belief that it requires twice the number of subjects as the simpler designs. In the present example, had only the two posttest groups been used, concentrating the total N into these groups (resulting in a doubled N of 28 per group), an informative result would have been obtained. Assuming the same means and standard deviations for Groups 3 and 4, the test of the treatment effect would *not* be significant, $t(54) = 1.807$, $p > .07$. Clearly, even when splitting the N per group in half, there is no loss of power for the Solomon design analyzed meta-analytically as compared to the use of the posttest-only design (in fact, there generally is a slight gain in power.)

We contend, then, that there is only one defensible objection to regularly employing Solomon designs: the greater difficulty in running subjects both with and without pretests. This objection is more than compensated for by the Solomon design's advantage in providing information concerning the external validity of the treatment effect.

Table 4
Results of Analysis of Variance on Posttest Scores

Source	*MS*	*df*	*F*	*p*
Pretest vs. Not (P)	.14	1	.007	.9337
Treatment vs. Not (T)	67.76	1	3.388	.0714
P × T	0.0	1	0.0	.999
Error	20.00	52		

Table 5
Analysis of Covariance on Groups 1 and 2

Source	*MS*	*df*	*F*	*p*
Treatment vs. Not	37.74	1	2.93	.0993
Error	12.87	25		

[3] In sequential tests, it is difficult to specify a priori the experimentwise Type I error rate over the entire sequence. Thus, it is perhaps most appropriate to consider the significance level at each step of the sequence as a conditional probability, conditional both on having reached that step and on the truth of the null hypothesis at that step.

References

Bracht, G. H., & Glass, G. V. (1968). The external validity of experiments. *American Educational Research Journal, 5,* 437–474.

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

Cuervorst, R. W., & Stock, W. A. (1978). Comments on the analysis of covariance with repeated measures designs. *Multivariate Behavioral Research, 13,* 509–513.

Glass, G. V. (1978). Integrating findings: The meta-analysis of research. *Review of Research in Education, 5,* 351–379.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* New York: Academic Press.

Helmstadter, G. C. (1970). *Research concepts in human behavior: Education, psychology, sociology.* New York: Appleton-Century-Crofts.

Huck, S., & McLean, R. (1975). Using a repeated measures ANOVA to analyze the data from a pretest-posttest design: A potentially confusing task. *Psychological Bulletin, 82,* 511–518.

Huck, S., & Sandler, H. M. (1973). A note on the Solomon 4-group design: Appropriate statistical analyses. *Journal of Experimental Education, 42,* 54–55.

Humphreys, L. (1976). Analysis of data from pre- and posttest designs: A comment. *Psychological Reports, 38,* 639–642.

Lana, R. E. (1959). Pretest treatment interaction effects in attitudinal studies. *Psychological Bulletin, 56,* 293–300.

Lana, R. E. (1969). Pretest sensitization. In R. Rosenthal & R. L. Rosnow (Eds.), *Artifact in behavioral research.* New York: Academic Press.

Mosteller, F. M., & Bush, R. R. (1954). Selected quantitative techniques. In G. Lindzey (Ed.), *Handbook of social psychology: Volume I. Theory and method.* Cambridge, MA: Addison-Wesley.

Oliver, R. L., & Berger, P. K. (1980). Advisability of pretest designs in psychological research. *Perceptual and Motor Skills, 51,* 463–471.

Rosenthal, R. (1978). Combining results of independent studies. *Psychological Bulletin, 85,* 185–193.

Rosnow, R. L. (1971). Experimental artifact. In L. C. Deighton (Ed.), *The encyclopedia of education* (Vol. 3). New York: Macmillan and Free Press.

Scheifley, V., & Schmidt, W. H. (1978). Analysis of repeated measures data: A simulation study. *Multivariate Behavioral Research, 13,* 347–362.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist, 12,* 752–760.

Solomon, R. L. (1949). An extension of control group design. *Psychological Bulletin, 46,* 137–150.

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., & Williams, R. M., Jr. (1949). *The American soldier: Adjustment during Army life* (Vol. 1). Princeton, NJ: Princeton University Press.

Willson, V. L., & Putnam, R. R. (1982). A meta-analysis of pretest sensitization effects in experimental design. *American Educational Research Journal, 19,* 249–258.

Received February 2, 1987
Revision received August 19, 1987
Accepted October 27, 1987 ■