6-1-1996

Controlling Experiment-wise Type I Error of Meta-analysis in the Solomon Four-group Design

Shlomo S. Sawilowsky
Wayne State University, shlomo@wayne.edu

Controlling Experiment-wise Type I Error of Meta-analysis in the Solomon Four-group Design

Shlomo S. Sawilowsky
shlomo@wayne.edu

Evaluation and Research
Wayne State University
First International Conference on Multiple Comparisons, June 23, 1996, Tel Aviv, Israel

Abstract. Stouffer=s Z, a meta-analytic technique, was proposed by W. Braver and Braver (1988) for analyzing data collected from a Solomon Four-group Design. Sawilowsky, Kelley, Blair, and Markman (1994) showed that this technique produces inflated Type I error rates. Recommendations are made to control the false positive inflation of their procedure.

According to Campbell and Stanley (1963), the Solomon Four-group Design (Solomon, 1949) has "higher prestige" than most other quantitative research designs, because it permits "explicit consideration of *external validity* factors" (p. 24). It contains both pretested and non-pretested treatment and control groups. Campbell and Stanley noted the lack of a single statistic for analyzing all of the data collected from this design.

To resolve this data analysis problem, W. Braver and Braver (1988) proposed the use meta-analysis. They provided a flowchart that documents a sequence of testing procedures. The first test (Test A) is a 2×2 ANOVA on the means of the four posttests= scores, which examines the interaction of Pretest - Treatment. If this test is significant, pretest sensitization is indicated, which is an undesirable situation where the outcome is confounded by the effect of the pretest (see, e.g., Willson & Putnam, 1982; Lana, 1959). W. Braver and Braver (1988) suggested a variety of follow-up tests (called Test B and Test C) to further investigate the nature of the pretest effects. If Test A is not significant, the sequence continues with Test D, which is a test of main effects for the treatment vs. control group.

Conditioned on these two preliminary tests being non-significant, the procedure continues with Test E, an ANCOVA on the posttest scores for the treatment and control group that was pretested, with the pretest scores used as the covariate. (W. Braver & Braver, 1988, suggested other tests, referred to as Test F and Test G, which could be substituted for the ANCOVA.) If the ANCOVA is not significant, the next step is to conduct an independent samples t test on the posttest scores for the non-pretested treatment and control groups, which they referred to as Test H. A significant result of Test E (ANCOVA) or Test H (t test) indicates a significant treatment outcome.

If neither of the final two classical tests (Test E and Test H) are significant, the sequence of testing concludes with Test I, Stouffer=s Z (Stouffer, et al., 1949) which is a meta-analytic procedure. W. Braver and Braver (1988) noted that nothing precludes the use of meta-analysis in

analyzing "several different tests of the same effect within but one study, as in the case of the Solomon four-group design" (p. 152).

In combining the two tests (Test E & Test F), Stouffer=s Z is

$$Z_{Meta} = \frac{Z_{p1} + Z_{P2}}{\sqrt{2}}$$

where Z_{p1} is the one-tailed p value associated with Test E (ANCOVA) and Z_{p2} is the one-tailed p value of Test H (t test). Stouffer=s Z is a nonparametric omnibus test of the null hypothesis that there is no treatment effect in any of the groups.

The current study

On the basis of a Monte Carlo study, Sawilowsky et al. (1994) found that the experiment-wise Type I error rate of W. Braver and Braver=s (1988) sequential procedure inflates to about 0.14, nearly triple the size of nominal α. (Other concerns with the procedure were discussed in Sawilowsky et al., 1990a; 1990b.) These results were based on population normality. The purpose of the current study is to a) present new Type I error results for nonnormal data, and more importantly, b) offer solutions to control the experiment-wise error rate associated with the use of Stouffer=s Z in the analysis of the Solomon Four-group Design.

The new Type I error study is necessary because, in an effort to control Type I errors, W. Braver and Braver (1995) subsequently revised their procedure to eliminate one of the preliminary tests (Test D). Solutions to control the Type I error rate inflations are necessary, because workers are now using this procedure. For example, Deanne, Spicer, and Leathem (1992) used W. Braver and Braver=s (1988) procedure in their study on the effects of videotaped preparatory information on client's state anxiety, expectations, and psychotherapy outcome.

New Type I error results

Results from Monte Carlo techniques reported below were obtained from a Fortran program written for a Pentium-based microcomputer using Microsoft Fortran Powerstation, accessing IMSL (1987) and RANGEN (1987) subroutines.

Population normality is an underlying assumption of the parametric tests that are conducted prior to the nonparametric meta-analytic Stouffer=s Z. Table 1 contains the results of a Monte Carlo simulation with samples of size n = 30 per group drawn from a variety of theoretical distributions, pretest-posttest correlation of r = 0.00, nominal α = 0.05, for 10,000 repetitions.

As noted above, W. Braver and Braver (1995) retracted Test D from their proposed sequence of testing. The upper part of the table contains results with Test D retained in the sequence of testing. The mean (median) error rate was .1298 (.1320). In accordance with W. Braver and Braver=s retraction of this test, however, the lower portion of the table indicates the error rates with Test D excluded from the simulation. The mean (median) error rate was reduced to .1185 (.1219).

According to Micceri (1989), it is appropriate to augment Monte Carlo results based on theoretical distributions with studies on large real data sets. Table 2 was constructed in the same fashion as the lower part of Table 2 (i.e., with test D excluded), except that real data sets were sampled with replacement instead of theoretical populations. The eight distributions were identified by Micceri (1989) as being representative of educational and psychological data. The data sets are available in Sawilowsky, Blair, and Micceri (1990); see Sawilowsky and Blair (1992) for histograms and summary statistics. The results were similar with a mean (median) experiment-wise Type I error rate of .1239 (.1233) for the eight real data sets.

Table 1. Conditional and *Experiment-wise* Type I Error Rates for a Variety of Distributions, α = 0.050, n = 30, Pretest-Posttest r = 0.0, 10,000 Repetitions.

Conditional and *Experiment-wise* Type I Error Rates For Tests A, D, E, H, & I										
Test	N	U	E	CS	LN	G	W	C	vM	t
A	0.0514	0.0308	0.0491	0.0488	0.0636	0.1047	0.0463	0.0204	0.0483	0.0518
D	0.0463	0.0489	0.0442	0.0530	0.0386	0.0433	0.0456	0.0211	0.0465	0.0462
E	0.0205	0.0250	0.0193	0.0161	0.0169	0.0113	0.0192	0.0125	0.0185	0.0183
H	0.0184	0.0220	0.0183	0.0182	0.0149	0.0108	0.0198	0.0117	0.0196	0.0170
I	0.0002	0.0008	0.0001	0.0005	0.0009	0.0002	0.0002	0.0032	0.0000	0.0003
	0.1368	*0.1275*	*0.1310*	*0.1366*	*0.1349*	*0.1703*	*0.1311*	*0.0689*	*0.1329*	*0.1336*
Conditional and *Experiment-wise* Type I Error Rates For Tests A, D, E, H, & I										
A	0.0514	0.0308	0.0491	0.0488	0.0636	0.1047	0.0463	0.0204	0.0483	0.0518
E	0.0337	0.0394	0.0318	0.0297	0.0272	0.0241	0.0337	0.0175	0.0338	0.0334
H	0.0326	0.0380	0.0330	0.0307	0.0241	0.0224	0.0352	0.0155	0.0342	0.0290
I	0.0071	0.0077	0.0057	0.0080	0.0077	0.0071	0.0065	0.0072	0.0066	0.0079
	0.1248	*0.1159*	*0.1196*	*0.1172*	*0.1226*	*0.1583*	*0.1217*	*0.0606*	*0.1229*	*0.1221*

Notes: N = Normal, U = Uniform, E = Exponential, CS = Chi-Squared (df = 1), G = Gama (Shape = 3), W = Weibull (Shape = 2), C = Cauchy, vM = von Mises (Shape = 1), and t = t(df = 3).

The experiment-wise Type I error rate is not affected by the pretest-posttest correlation or by the sample size. Table 3 contains results of simulations under population normality (with Test D excluded) for nominal α = 0.05, pretest-posttest correlations of r = 0.00 - 0.95 (0.05), and for samples of size n = 3, 10, 20, and 30. The results indicate that the mean (median) experiment-wise Type I error rate for meta-analysis in the Solomon Four-group Design is .1222 (.1233) for the various pretest-posttest correlation examined and .1248 (.1274) based on sample sizes examined.

In summary, after having traversed the flowchart of sequential testing in order to conduct the meta-analysis in the Solomon Four-group Design (W. Braver and Braver, 1988; 1990a; 1990b; 1995), the most important question a researcher must ask is "What is the probability of getting the

wrong answer?" In terms of false positives, it is about .14, which is nearly triple the nominal α rate.

Table 2. Conditional and *Experiment-wise* Type I Error Rates for a Variety of Real Data Sets From Micceri (1989) for Tests A, E, H, and I; α = 0.05, n = 30, Pretest-Posttest r = 0.0, 10,000 Repetitions.

	Data Set							
Test	1	2	3	4	5	6	7	8
A	0.0506	0.0500	0.0476	0.0513	0.0482	0.0509	0.0495	0.0503
E	0.0343	0.0349	0.0349	0.0317	0.0326	0.0339	0.0338	0.0349
H	0.034	0.0332	0.0294	0.0335	0.0289	0.0321	0.033	0.0323
I	0.0063	0.0068	0.0076	0.0059	0.0063	0.0071	0.0063	0.0070
	0.1252	*0.1249*	*0.1195*	*0.1224*	*0.1160*	*0.124*	*0.1226*	*0.1245*

Notes: Data set 1 = Smooth Symmetric, Achievement; 2 = Discreet Mass With Gap, Psychometric; 3 = Discreet Mass At Zero With Gap, Achievement; 4 = Extreme Asymmetry, Achievement; 5 = Extreme Asymmetry, Psychometric; 6 = Digit Preference, Psychometric; 7 = Extreme Bimodality, Psychometric; 8 = Multimodality and Lumpy, Achievement.

Table 3. Experiment-wise Type I Error Rates for Meta-analysis in the Solomon Four-group Design (Tests A, E, H, and I) with α = 0.05, 10,000 Repetitions.

r	Experiment-wise Type I Error Rate	$\bar{r}$		n	Experiment-wise Type I Error Rate
0.00	0.1248	0.50	0.1197	3	0.1324
0.05	0.1247	0.55	0.1225	10	0.1299
0.10	0.1242	0.60	0.1250	20	0.1122
0.15	0.1182	0.65	0.1211	30	0.1248
0.20	0.1246	0.70	0.1241		
0.25	0.1242	0.75	0.1261		
0.30	0.1251	0.80	0.1189		
0.35	0.1172	0.85	0.1261		
0.40	0.1215	0.90	0.1189		
0.45	0.1190	0.95	0.1178		

Notes: $\bar{r}$ = Pearson Product-Moment Coefficient of Correlation.

Proposed solutions

Three proposals are proffered, from the least to most desirable in my judgment, for resolving the data analysis problems inherent in the Solomon Four-group Design:

1. Given the inflation of Type I errors documented in Sawilowsky et al. (1994) and above, two methods of Bonferroni-type adjustments to control experiment-wise error are proposed. The values compiled in Table 4, obtained by Monte Carlo methods, indicate the required nominal α for Test A, E, H, and I to maintain robust results. The upper limit of robustness defined by Bradley=s (1968) conservative definition is empirical alpha α ` 1.1$\alpha$; with nominal $\alpha$ = 0.050 the upper limit is ` 0.055. The upper limit of the liberal definition is α ` 1.5 $\alpha$, or in this case ` 0.075.

The first method assumes nominal α should be evenly divided among the constituent tests. As indicated in Table 4 (following page), the conservative definition requires nominal α to be set at 0.020, while the liberal definition permits nominal α = 0.0275. The second method allocates 0.050 to Test A, because it is a preliminary test of assumptions. Nominal α for the remaining tests of treatment effects (Tests E, H, and I) are evenly divided in such a fashion so as to not exceed the upper limit of robustness. According to the conservative definition, the latter three tests require nominal α = 0.005, while the liberal definition of robustness permits nominal α = 0.020. (Although the Experiment-wise Type I error rate is controlled to an upper limit of robustness, note that the competing tests are nevertheless conducted at different α levels.)

2. Kennedy and Bush (1985, p. 310, 323) devoted nearly a chapter to the analysis of a 3-factor completely randomized posttest only ANOVA design, where one of the independent variables was whether a subject was pretested or was not pretested. They referred to this design as an extension to the Solomon Four-Group Design. It permits the examination of membership or non-membership in a pretested group as it interacts with the other independent variables, or in terms of a main effect. (Note, however, that the pretest scores are not used in the analysis.) The Experiment-wise Type I error rate of this 3-factor design, when conducted as discussed, is nominal α.

3. Wilson and Putnam (1982) found that when pretest effects in psychology, or other disciplines within the social and behavioral sciences, are present they are small and not likely to interact with treatments. W. Braver and Braver (1988) also cited five literature reviews that indicated pretest sensitization "rarely occur(s)" (p. 150). If theory suggests and empirical results show *no* pre-test sensitization, given the prospects of Type I error inflations with meta-analysis, a prudent recommendation would be to minimize concerns regarding pretest sensitization and avoid the Solomon Four-group Design until a more robust and powerful data analysis procedure is found. Using a randomized posttest-only treatment vs. control group design and analyzing the results with the independent samples t test with nominal α = 0.05, or using a randomized pretest-posttest treatment vs. control group design and analyzing the results with ANCOVA with nominal α = 0.050, is preferable to conducting the same tests at nominal α = 0.005 which is necessary to control the Experiment-wise Type I error inflation in order to conduct a meta-analysis.

Table 4. Nominal α Levels Required to Maintain Robust Experiment-wise Type I Error Rates for Stouffer=s Z (Meta-analysis) in the Solomon Four-Group Design for Nominal Alpha = 0.050 Based on Bradley=s (1968) Conservative and Liberal Definition of Robustness, n = 30 , r = 0.0, 10,000 Repetitions.

Test	Conservative Definition of Robustness ($\alpha < 1.1\alpha$)		Liberal Definition of Robustness ($\alpha < 1.5\alpha$)	
	Nominal α	Conditional & *Experimental-Wise Type I Error*	Nominal α	Conditional & *Experiment-wise Type I Error*
		Method 1		
A	0.0200	0.0221	0.0275	0.0283
E	0.0200	0.0137	0.0275	0.0200
H	0.0200	0.0136	0.0275	0.0194
I	0.0200	0.0042	0.0275	0.0045
		0.0536		*0.0722*
		Method 2		
A	0.0500	0.0514	0.0500	0.0514
E	0.0050	0.0018	0.0200	0.0105
H	0.0050	0.0017	0.0200	0.0103
I	0.0050	0.0020	0.0200	0.0041
		0.0569		*0.0763*

Notes: Method 1 allocates nominal α evenly. Method 2 allocates 0.050 to Test A, a preliminary test of assumptions, and evenly divides the remaining alpha so as to not exceed the upper limit of robustness. The *Experiment-wise Type I error rate*) is within sampling error of the theoretical upper limit of robustness.

References

Bradley, J. V. (1968). *Distribution-free statistical tests*. Englewood Cliffs, NJ: Prentice Hall.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Boston: Houghton Mifflin.

IMSL (1987, April). *International Mathematical and Statistical Libraries. IMSL STAT/Library Version 1.0*. Houston, TX.

Deane, F. P., Spicer, J., & Leathem, J. (1992). Effects on videotaped preparatory information on expectations, anxiety, and psychotherapy outcome. *Journal of Consulting and Clinical Psychology*, *60*, 980-984.

Kennedy, J. J., & Bush, A. J. (1985). *An introduction to the design and analysis of*

experiments in behavioral research. Lanham: University Press of America.

Lana, R. E. (1959). Pretest-treatment interaction studies in attitudinal studies. *Psychological Bulletin*, *4*, 293-300.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*, 156-166.

RANGEN. (1987, May, Version 1.0). Boca Raton: IBM.

Sawilowsky, S., S., & Blair, R. C. (1992). A more realistic look at the robustness and Type II error properties of the t test to departures form population normality. *Psychological Bulletin*, *111*, 352-360.

Sawilowsky, S., Blair, R. C., & Micceri, T. (1990). A PC FORTRAN subroutine library of psychology and education data sets. *Psychometrika*, *55*, ,729.

Sawilowsky, S., Kelley, D. L., Blair, R. C., & Markman, B. S. (1994). Meta-analysis and the Solomon four-group design. *Journal of Experimental Education*, *62*(4), 361-376.

Sawilowsky, S., S., & Markman, B. S. (1990a). Another look at the power of meta-analysis in the Solomon four-group design. *Perceptual and Motor Skills*, *70*, 177-178.

Sawilowsky, S. S., & Markman, B. S. (1990b). Rejoinder to Braver and Braver. *Perceptual and Motor Skills*, *71*, 424-426.

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, NY: Princeton University Press.

W. Braver, M. C., & Braver, S. L. (1988). Statistical treatment of the Solomon four-group design: A meta-analytic approach. *Psychological Bulletin*, *104*, 150-154.

W. Braver, M. C., & Braver, S. L. (1995). Meta-analysis for Solomon-Four Group Designs Redeemed: A reply to Sawilowsky, Kelley, Blair, and Markman (1994). Manuscript submitted to *Journal of Experimental Education*.

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

Acknowledgments

This research was supported in part by a 1995 - 1996 Wayne State University Career Development Chair Award.