The Limitations of Random Assignment: A Computational Simulation

Boster, F, Hao, Q, Reynolds, R

Abstract

In social research, estimates of the effect of one variable on another can be distorted by individual variation associated with the dependent variable. In experimental designs, random assignment of subjects to conditions helps address this challenge. However, random assignment does not always produce equivalent groups, and when distributions of an extraneous variable differ across conditions the apparent effect of a treatment may differ substantially from the true effect. On this basis, a simulation was conducted to examine (a) the extent that random assignment of subjects to conditions affects the distribution of pretest scores in a posttest-only, control group experiment, and (b) the extent to which these various distributions of pretest scores influence the observed effect of the experimental treatment on the dependent variable. Results indicate that random assignment often fails to create equivalent groups, and that these failures lead to substantial discrepancies between observed and actual effects of a treatment.

In their seminal monograph on experimental design Campbell and Stanley (1963) emphasized that the equivalence of subjects assigned to the various experimental and control groups that comprise an experiment poses a challenge to the validity of experiments. They termed this potential source of invalidity the *selection bias*. Subsequently, methods texts have reinforced this point (Singleton & Straits, 2010, pp. 233-234). Campbell and Stanley (1963) argue that subject differences in demographic characteristics, personality traits, and other dimensions of individual variation associated with the dependent variable can result in unwarranted conclusions being drawn to the extent that their distributions differ across conditions of the experiment. The random assignment of subjects to the various conditions of an experiment provides a means of addressing this challenge.[1]

Investigators have asserted that random assignment assures the comparability or near comparability of subjects in the various conditions of the experiment. Examination of canonical texts produces examples that reinforce this point of view. For example, Campbell and Stanley (1963) write: "Perhaps Fisher's most fundamental contribution has been the concept of achieving pre-experimental equation of groups through randomization" (p. 2). And, Fisher (1947) writes that randomization, ". . . relieves the experimenter from the anxiety of considering and estimating the magnitude of the innumerable causes by which the data may be disturbed" (p. 43). Unsurprisingly, some methods textbook authors have adopted the same point of view. Singleton and Straits provide an example when they write:

> Random assignment means that the procedure by which subjects are assigned (in this case, tossing a coin) ensures that each subject has an equal chance of being in either group. By virtue of random assignment, individual characteristics or experiences that might confound the results will be about evenly distributed between the two groups.

> Thus, the number of students who are bright or dull, motivated or unmotivated, fully nourished or hungry, in love or not in love, and so forth, should be about the same in each group" (p. 197).

A strong sense of "comparability" as used in the preceding paragraph requires equal means and standard deviations on the extraneous variable(s) in each condition of the experiment. A weakened version requires equal means only. Notice that equal means on an extraneous variable both in a control group and in an experimental group implies that the correlation between the induced variable and the extraneous variable equals zero. To the extent that randomization fails to yield this outcome, mean scores on the extraneous variable differ in the two conditions. And, when mean scores on the extraneous variable differ in the two conditions, the correlation between the experimentally induced variable and the extraneous variable departs from zero. For example, consider a one-factor, independent groups experiment in which the experimentally induced variable consists of an experimental condition and a control condition. Suppose that the mean score on an extraneous variable is 8.5 in the experimental condition, 6.5 in the control condition, the variance in both conditions equals 5, and that 20 *S*s participate in each condition. In this case $t\ (38) \approx 2.83$, and transforming t to r yields, $r \approx .42$. Notice that reversing the order of the means, i.e., 6.5 in the experimental condition and 8.5 in the control condition, results in $r \approx -.42$. Thus, "near comparability" refers to this correlation closely approximating zero.

Despite the assertions of Fisher, Campbell and Stanley, and others, some scholars question this assertion (Krause & Howard, 2003; Sidani, 2006). One reason for suspicion stems from considering the mathematical constraints on the possible values that correlations among three variables may assume. Consider, for example, a case in which an experimentally induced

variable, X, correlates .3 with a dependent measure, Z. Denote a relevant extraneous variable, Y, and suppose that Y correlates .7 with the dependent measure, Z. It follows that the XY correlation can range from -.47 to .99 (Glass & Collins, 1970; Stanley & Wang, 1969). Consider three scenarios. In the first randomization works perfectly so that $r_{XY} = 0$, in the second $r_{XY} = -.3$ and in the third $r_{XY} = .3$. In the first case, performing the multiple regression analysis, or its equivalent the analysis of covariance, to examine the effect of X on Z controlling for the extraneous variable Y shows that, $\beta_X = r_{XY} = .30$. In the second case $\beta_X = .56$. And, in the third case, $\beta_X = .10$. When random assignment works perfectly, the XZ correlation estimates well the impact of X on Z. But, when $r_{XY} < 0$, the XZ correlation, or the t-test of the difference in the Z means in the control and experimental conditions, provides a substantial underestimate of the causal impact of X on Z. In this case Y suppresses the XZ correlation because random assignment placed people with lower mean scores on the extraneous variable in the experimental condition. And, when $r_{XY} > 0$, the XZ correlation, or the t-test of the difference in the Z means in the control and experimental conditions, overestimated the causal impact of X on Z because random assignment placed people with higher mean scores on the extraneous variable in the experimental group. The substantial differences among these coefficients, and the likely differences in the conclusions that investigators would draw from them, hint at the possibility that random assignment may depart from the ideal suggested by Fisher and Campbell and Stanley.

Nevertheless, although it is possible that r_{XY} ranges from -.47 to .99, each of the values in this range do not have equal probability. With even modest numbers of subjects assigned to conditions, enumerating all possible combinations of the distribution of the extraneous variable

across even two conditions raises a formidable challenge. Simulations, however, can provide an excellent approximation.

The Posttest-Only, Control Group Design

The posttest-only, control group design provides an important venue in which to test the effectiveness of random assignment. For example, with a dependent measure such as attitude change investigators may rely on random assignment to equate subjects on the focal pretreatment attitude so that analyzing the impact of an induction on the posttreatment attitude measure approximates closely a measure of attitude change, the variable of theoretical interest (e.g., Yan, Dillard, & Shen, 2012). The belief that random assignment of subjects to conditions assures that subjects in the two conditions have equal, or near equal, pretreatment scores on the dependent measure forms a pivotal premise in the argument for the conclusion that the post-treatment attitude measure provides a reasonable proxy for a measure of attitude change.[2] And, Campbell and Stanley, for instance, endorsed this point of view when in their discussion of the posttest-only, control group design they wrote, "Nonetheless, the most adequate all-purpose assurance of lack of individual biases between groups is randomization. Within the limits of confidence stated by the tests of significance, randomization can suffice without the pretest (p.25).

Conceiving of the pretest in a pretest-posttest, control group design as a type of extraneous variable, the reasons provided previously for questioning the effectiveness of random assignment apply. To examine this matter, a simulation was conducted to examine (a) the extent that random assignment of subjects to conditions affects the distribution of pretest scores in a posttest-only, control group experiment, and (b) the extent to which these various distributions of pretest scores influence the observed effect of the experimental induction on the dependent variable. The following section describes the details of that simulation.

Method

The simulation consisted of four stages. First, a model generated the posttest scores. Second, from a defined population of subjects mathematical manipulations constrained the pivotal population parameters, setting the effect size for the treatment. Third, computer simulations drew random samples of varying sizes from this population. Last, analyses examined the obtained results. Discuss of each of these four steps follows.

Modeling the Posttest-Only, Control Group Design

In a posttest-only, control group experiment a distribution of pretest scores (Y) exists, but is unobserved. In the simulation this distribution was modeled by drawing N observations from a normally distributed population. Thus, each observation, i, has a pretest score, y_i.

Pretest scores were then assigned randomly to one the two values of the experimentally induced variable (X), i.e., either to the control group or to the experimental group. This random assignment was conducted under the constraint of equal numbers of observations in the two conditions.

In the experimental group, but not in the control group, the treatment was presumed to have some positive and homogeneous effect, denoted by **a**, on the posttest score (Z). Moreover, it was presumed that numerous uncontrolled forces would exert an impact, both on the posttest score in the control group and on the posttest score in the experimental group. These sum of these forces contribute to the residual (R), its effect characterized by a change of **r** points from the pretest to the posttest. In the simulation the distribution of R, in both the control group and the experimental group, was normal with a mean of zero and a standard deviation σ_R. Furthermore, R was constructed so that in the population it was uncorrelated with Y.

Consequently, the post-test scores in the experiment group (X=1) resulted from the following expression,

$$z = y + a + r = y + aX + r, (X = 1, r \sim N(0, \sigma_R^2)) \quad (1)$$

and the post-test scores in the control group (X=0) were calculated as,

$$z = y + r = y + aX + r, (X = 0, r \sim N(0, \sigma_R^2)) \quad (2)$$

Alternatively, generalizing (1) and (2) yields,

$$Z = y + aX + r, (X = 0,1; r \sim N(0, \sigma_R^2)). \quad (3)$$

Constraining Population Parameters and Effect Sizes

The correlation between X and Z in the population (ρ_{XZ}) is a constant so that any variance in sample XZ correlations results from the post-test only design with random assignment. Pivotal population parameters were set as follows.

var(Y) = 1

var(X) = 0.25 (by definition of equal cell size)

$a = \sqrt{0.72}$

var(R) = 0.82

Mean population correlations between variables in the model were set as follows (see Appendix B for elaboration).

$\rho_{XZ} = .30$

$\rho_{YZ} = \sqrt{0.5}$

$\rho_{XY} = \rho_{XR} = \rho_{YR} = 0$

Computer Simulations

Ten thousand simulations, each representing a distinct sample randomly drawn from the population, were conducted for five sample sizes (N = 20, 40, 60, 80, and 100). For each

iteration, N pre-test scores (y) were randomly drawn from a standard normal distribution, and N residual scores (r) were randomly drawn from a normal distribution with mean zero and variance.82. Subsequently, each of a randomly selected N/2 experimental group posttest observations were calculated from equation 1 ($z = y + \sqrt{0.72} + r$), and the remaining N/2 control group posttest observations were calculated from equation 2 ($z = y + r$). All simulations were conducted using STATA. See Appendix C an example of the STATA code.

Sample Statistics Calculation

The sample correlations between the experimentally induced variable, X, and the pretest score, Y (r_{xy}), the correlation between X and posttest score, Z (r_{xz}), the standardized regression coefficient estimating the impact of X on Z controlling for Y (β_x) were calculated for each iteration. Null hypothesis statistical significance tests (NHSST) were also conducted both for r_{xz} and β_x in each study.

RESULTS

Appendix A presents all tables. Table 1 shows the distributional properties of the obtained correlations between the experimentally induced variable, X, and the pretest, Y, for each of the five sample sizes. If randomization worked perfectly, then mean pretest scores in the control group would equal those in the experimental group with the result that the r_{xy} correlation would equal zero. It rarely does, however. Rounding to two decimal places, at N = 20 only 1.71% of the iterations produced an r_{XZ} which equaled 0.00 when rounded to two decimals. The comparable figures for the other sample sizes included: 2.64% (N = 40), 3.06% (N = 60), 3.70% (N = 80), and 4.34% (N = 100).

As the data in Table 1 indicate random assignment produces an r_{XZ} of zero to two decimal points, *on average*, for each sample size. As expected from the standard error of the

correlation coefficient, the standard deviation of these distributions decrease with sample size, although for each N the standard deviation exceeds slightly the standard error, a result that may be attributable to random assignment. Moreover, given that a value of three (3) indicates a perfectly mesokurtic distribution, these distributions differ only trivially from normality.

The striking feature of these data, however, is the range in the obtained values. For N = 20 (10 observations per condition) the range is 1.442 on a scale with a possible range of 2.0. And, although that figure decreases with sample size, it remains an ample 0.762 when N = 100. In conjunction with the size of the standard deviations this outcome indicates that random assignment can produce numerous and substantial differences in pretest scores between the control group and the experimental group. In this simulation these differences can, and do, have important implications for the observed effect of the induced variable, X, on posttest scores, Z.

Table 1 also contains the distributional properties of the correlation between the experimentally induced variable, X, and the posttest measure, Z. Notably, it is this mean difference, and the accompanying effect size measure (r in this case), that investigators would observe in a posttest-only, control group design, and that they would employ to draw conclusions as to the effectiveness of the treatment. The mean correlation hovers closely around a value of 0.31, slightly exceeding its population value of 0.30, for each value of N. As expected, the standard deviation decreases with sample size. The distribution of these values is skewed negatively, perhaps reflecting the bias in the Pearson Product Moment Correlation Coefficient when it is positive, with skewness decreasing with increasing N. Kurtosis is minimal.

Once again, the range is the striking feature of these data. When N = 20, the range is 1.454, with the largest obtained coefficient being 0.864 and the smallest values being -0.589. And, when N = 100, the range remains an ample .671 with a maximum value of approximately

0.60 and a minimum value less than zero. In conjunction with the size of the standard deviations this outcome shows that in posttest-only, control group designs with random assignment substantially different estimates of the impact of an experimental induction can result. And, as will be demonstrated subsequently, these differences may have important effects on the substantive conclusion drawn from experiments.

Finally, Table 1 includes the distribution of the estimate of the impact of the experimentally induced variable (X) on the posttest measure (Z) controlling for any differences in the pretest variable (Y) between experimental groups (i.e., that would result in $r_{XY} \neq 0$). This statistic is denoted as β_X and refers to the standardized regression coefficient. The mean values of β_X exceed, but only slightly, the mean values of r_{XY} with the mean values of β_X falling into the 0.31 – 0.32 range. Standard deviations decrease with sample size, and notably are smaller than those associated with r_{XZ} at each value of N. The distribution of the β_X coefficients approximates closely the normal distribution with skewness and kurtosis being minimal.

Although the range of the β_X coefficients remains ample, it is less than the range of r_{XZ} for each value of N. Moreover, as N increases from 20 to 80 the difference between the ranges of β_X and r_{XZ} increases. At N = 100 this difference decreases, but remains substantial.

Table 2 expands on the observed differences in r_{XZ} and β_X. The first column presents the absolute value of the difference in these coefficients, and the remaining columns display the percentage of the iterations in which r_{XZ} and β_X differed by that amount or more for each value of N. So, for example, when N = 20, in 75.89% of the iterations the difference between these coefficients differed by more than .05 with either r_{XZ} exceeding β_X or the reverse.

Table 2 demonstrates that for any sample size, the percentage of iterations meeting or exceeding a given discrepancy decreases as the magnitude of the discrepancy increases. Thus,

small discrepancies between the two coefficients are more frequent than are large discrepancies. This table also demonstrates that as N increases r_{XZ} and β_X differ for fewer cases at any given discrepancy. Hence, however the difference is defined, larger samples produce fewer differences than do smaller samples.

It is important to note that even small discrepancies can result in substantially different substantive conclusions being drawn in individual experiments. Nevertheless, some of the values in Table 2 suggest that across a body of research the appearance of important substantive differences might emerge from being able to correct for or not correct for differences in pretest scores for those assigned randomly to the control group and to the experimental group. For example, at N = 60 β_X and r_{XY} differ by .10 or more in almost 30% of the iterations. Also, at N = 60 more than 11% of the iterations produced differences of .15 or more.

Table 3 examines the extent to which these two estimates of the impact of X on Z, r_{XZ} and β_X, differ from the estimated population parameter of .31. It is this value, both of r_{XZ} and β_X, that is obtained when randomization works perfectly, i.e., when $r_{XY} = 0$. From Table 3 it can be observed that for both coefficients the percentage of iterations that deviate from .30 for any N decrease as the discrepancy increases. It can also be observed that for any fixed level of discrepancy the percentage of coefficients deviating from .31 decrease as N increases.

The striking feature of these data, however, is that the percentage of cases that deviate for any given N and any given discrepancy are lower, often substantially so, for β_X than for r_{XZ}. So, for example, when N = 60, 19.95% of the XZ correlations differ from the population value by .15 or more, the corresponding percentage being 5.29% for β_X. Therefore, for this N and at this level of difference the latter is closer to the population value by a factor of 3.77. Or, when N = 100, these values are 2.25% and 0.10% respectively, so that they differ by a factor of 22.5. Thus,

Table 3 indicates that controlling for the pretest when assessing the impact of X on Z, as opposed to failing to doing so (as would necessarily be the case in the posttest-only, control group design), produces estimates nearer those of the population values, i.e., the estimate of the XZ effect when the control group and experimental group have equal pretest means on the dependent variable.

Finally, Table 4 presents the implications of non-equivalent experimental groups in terms of the null hypothesis statistical significance test (NHSST). The first two sections of the table show the percentage of the samples for which statistically significant effects were obtained, first for the XZ correlation and second for β_X. Because the simulation was created so that the impact of X on Z was non-zero, the fact that none of the entries in these sections of the table reaches 100% is a comment on the lack of statistical power when effects of this size are estimated from samples of these sizes. And, consistent with what is known of statistical power (e.g., Cohen, 1988), the entries in these sections of the table approach 100% as N increases. The most striking feature of these sections of the table, however, is that when $p \leq .05$, β_X is more likely to produce a statistically significant outcome than is r_{XZ}. For example, statistically significant results are produced for approximately the same percentage of β_X when N = 20 as for r_{XZ} when N is twice as large. And, only a slightly smaller percentage of β_X are statistically significant when N = 40 as for r_{XZ} when N = 100. Hence, these results indicate that β_X is a more powerful estimator of the XZ effect than is r_{XZ}.

The last two sections of Table 4 present the percentage of cases for which r_{XZ} is statistically significant (r_{XZ} only) and β_X is not, and the percentage of cases for which β_X is statistically significant (β_X only) and r_{XZ} is not. Consistent with what is known about statistical power, these figures decrease as N increases. More striking, however, is that it is rarely the case

that r_{XZ} is statistically significant when β_X is not; whereas, it is more substantially more common for β_X to be statistically significant when r_{XZ} is not.

DISCUSSION

Summary

These results lead to at least five important conclusions. First, at least with $N \leq 100$, in a posttest-only, control group design random assignment of subjects to conditions is unlikely to produce two groups of equivalent subjects, defined as having equal means on the pretest (had one been administered). But, second, random assignment improves as sample sizes increase. Although even when $N = 100$ a small percentage of cases manifest equivalence, as defined by r_{XY} equaling zero to two decimals, the decreasing standard deviation in the r_{XY} distribution that accompanies increasing N indicates that as N increases the r_{XY} correlation is likely to be substantially closer to zero. Notably, this point was familiar to Campbell and Stanley when they wrote, “Thus, the assurance of equality is greater for large number of random assignments than small” (Campbell & Stanley, 1963, p. 15).

Third, the effectiveness of random assignment in producing equivalent groups has an impact on the substantive conclusions that would be drawn from standard analyses of the data. For example, for a disturbingly large proportion of cases either controlling for pretest differences (β_X) or failing to control for pretest differences (r_{XZ}) produces different substantive conclusions.

Fourth, analyses which control for pretest differences (β_X) reflect the magnitude of the population effect more accurately than those that do not control for pretest differences (r_{XZ}). Moreover, and fifth, controlling for pretest differences provides a more powerful test of the null hypothesis of no mean difference between the control group and experimental group means, or equivalently, no association between them.

Implications

The primary methodological implication of these results is to obtain a pretest measure of the dependent variable when possible (but see Note 2). Certainly occasions arise when a pretest of the dependent measure is impossible to obtain. Consider, for example, how a pretest measurement might be constructed in an experiment investigating the effectiveness of the legitimization of paltry favors compliance gaining technique (Andrews, Carpenter, Shaw, & Boster, 2008). In this example, *what* one would measure would be unclear. Additionally, if the sample varied little on the pretest, it would be reasonable to avoid it. In such a case the probability that random assignment could produce substantial pretest difference would be extremely low. For the vast majority of communication phenomena of interest, however, a sample of subjects with homogeneous scores on the dependent measure prior to a treatment would be unusual in the extreme (but see the classic Festinger & Carlsmith, 1959). Finally, if a pretest was believed to be reactive, it might be prudent to avoid it. On the other hand, two other strategies might be pursued instead. One could, for instance, separate the pretest from the experimental induction temporally to reduce any feared reactivity. Alternatively, a Solomon-Four Group Design might be employed as a means of estimating the impact, if any, of the pretest.

A second implication is the importance of sample size. It is well known that increased sample size increases statistical power and improves the precision of point estimates. Nevertheless, the data generated in this simulation show that it also increases the probability that random assignment will result in near equivalence of pretest scores in the control group and the experimental group, thus removing another source that might lead investigators to reach erroneous conclusions from their data analyses. These simulation data do not allow a

recommendation for any minimum sample size. Certainly, larger N is preferable up to the point that it might begin limiting research by investigators unable or unwilling to commit the resources necessary to reach some recommended number of subjects. Equally certain is that as N increases there are diminishing returns in the amount by which sampling error is reduced.

A third and final implication pertains to meta-analysis and replication. Schmidt and Hunter (2015, p. 41) detail an extended set of methodological artifacts, e.g., sampling error, measurement error, restriction in range, etc., that produce variance in a distribution of effect sizes compiled from studies examining a focal bivariate relationship. Schmidt and Hunter's 75% Rule asserts that when 75% - 99% of the variance in a distribution of effect sizes can be explained by a limited set of artifacts (e.g., sampling error, measurement error, and restriction in range only), the investigator need not search for moderator variables to explain the remaining variance. Instead, they argue that it is likely attributable to other, unexamined artifacts. As the simulation data indicate the difference in pretest scores between conditions that results from random assignment in a posttest-only, control group design produces another artifact that may explain a substantial percentage of the unexplained variance in a distribution of effect sizes.

Recently, many psychologists have a lack of confidence in the replicability of certain psychological experiments (Pashler & Wagenmakers, 2012). As the present data indicate, both for r_{XZ} and β_X, extreme and misleading effect sizes may be observed as a result of the limitations of random assignment. The possibility must be considered that failures to replicate certain experiments could arise, in whole or in part, from the manner in which subjects are assigned to conditions. This point is especially pertinent to small N experiments.

Limitations

This simulation has a number of limitations that require mention. Initially, there are limitations in the scope of the project. For example, only a pretest-posttest, control group design

with independent groups was simulated. There was only one experimental factor, and it was limited to two values. Only one population effect size, r = .30, was examined; only one population autocorrelation value, r = .70, was examined; and only five sample sizes were employed. Although these values are, in our experience, not unusual, and perhaps even typical, the question of the extent of the generality of these results remains to be determined.

A second set of limitations arises from simplified features of the simulation. For example, the data set was constructed so that the population pretest score was uncorrelated with the residual, R. Moreover, it was assumed that there would be no non-additivity between the experimental induction, X, and the pretest score, Y. In actual experimental data, these features may be violated. Again, the question of the generalizability of these results with these restrictions relaxed must be addressed.

Future research

Future simulations are planned that address these limitations. Additionally, this line of inquiry will be expanded to include another type of extraneous variable(s); namely, those that would be thought of as potential covariates because they are associated substantially with the dependent variable. When mean differences on these variables in a control group and an experimental group result from random assignment, results similar to those found in this simulation arise. Furthermore, multiple covariates may exist, complicating the pattern of results. The topic of randomization and its limitations deserves more attention. It is important for researchers to better understand the meaning of observations that result from designs that rely on randomization, particularly when key individual differences remain unobserved.

Notes

[1] This claim applies to random assignment to conditions, not to probability or non-probability sampling of the subjects participating in the experiment.

[2] With a dependent variable unmeasurable prior to the treatment, when the sample does not vary on the dependent variable prior to the treatment, or when a pretest measure of the dependent variable might prove reactive, investigators may, for good reason, avoid obtaining a pretreatment measure.

References

Andrews, K.R., Carpenter, C.J., Shaw, A.S., & Boster, F.J. (2008). The legitimization of paltry favors effect: A review and meta-analysis. *Communication Reports*, *21*, 59-69.

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

Cohen, J. (1988). *Statistical power analyses for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Festinger, L., & Carlsmith, J.M. (1959). Cognitive consequences of forced compliance. *Journal of Abnormal and Social Psychology*, *58*, 203-210.

Fisher, R.A. (1947). *The design of experiments, 4th ed.* Edinburgh: Oliver & Boyd.

Glass, G.V., & Collins, J.R. (1970). Geometric proof of the restrictions on the possible values of r_{xy} when r_{xz} and r_{yz} and fixed. *Educational and Psychological Measurement*, *30*, 37-39.

Krause, M.S., & Howard, K.I. (2003). What randomization does and does not do. *Journal of Clinical Psychology*, *59*, 751-766.

Pashler, H., & Wagenmakers, E.J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence. *Perspectives on Psychological Science*, *7*, 528-530.

Schmidt, F.L., & Hunter, J.E. (2015). *Methods of meta-analysis: Correcting error and bias in research findings*. Los Angeles, CA: Sage.

Sidani, S. (2006). Random assignment: A systematic view. In R.R. Bootzin & P.E. McKnight (Eds.), *Strengthening research methodology: Psychological measurement and evaluation* (pp., 125-141). Washington, D.C.: American Psychological Association.

Singleton, R.A., Jr., & Straits, B.C. (2010). *Approaches to social research* (5th Ed.). New York, New York: Oxford.

Stanley, J.C., & Wang, M.D. (1969). Restrictions on the possible values of r_{12}, given r_{13} and r_{23}. *Educational and Psychological Measurement*, *29*, 579-581.

Yan, C., Dillard, J.P., & Shen, F. (2012). Emotion, motivation, and the persuasive effects of message framing. *Journal of Communication*, *62*, 682-700.

APPENDIX A

Table 1

Distribution Properties—10,000 Simulations

r_{XY}	$N = 20$	$N = 40$	$N = 60$	$N = 80$	$N = 100$
Mean	-0.001	0.001	0.001	-0.001	0.001
S.D.	0.228	0.159	0.129	0.114	0.100
Skewness	0.024	-0.010	-0.049	-0.001	0.007
Kurtosis	2.763	2.918	2.863	2.934	2.941
Max.	0.728	0.566	0.462	0.499	0.413
Min.	-0.713	-0.600	-0.454	-0.460	-0.348
r_{XZ}	$N = 20$	$N = 40$	$N = 60$	$N = 80$	$N = 100$
Mean	0.314	0.315	0.314	0.312	0.313
S.D.	0.203	0.141	0.116	0.100	0.088
Skewness	-0.374	-0.281	-0.245	-0.178	-0.169
Kurtosis	3.067	2.998	2.963	3.012	2.982
Max.	0.864	0.744	0.667	0.716	0.597
Min.	-0.589	-0.222	-0.175	-0.092	-0.073
β_X	$N = 20$	$N = 40$	$N = 60$	$N = 80$	$N = 100$
Mean	0.324	0.318	0.316	0.315	0.314
S.D.	0.142	0.096	0.077	0.067	0.060
Skewness	-0.090	-0.026	-0.031	-0.039	-0.027
Kurtosis	3.157	2.980	3.019	3.059	2.951
Max.	0.813	0.664	0.603	0.611	0.537
Min.	-0.524	-0.124	0.000	0.054	0.082

Note: $N = 20$ implies 10 per condition (x = 0, x = 1).

Table 2

Correlation / Beta Discrepancy

Discrepancy	**Percentage of 10,000 Simulations**				
|r - β|	***N* = 20**	***N* = 40**	***N* = 60**	***N* = 80**	***N* = 100**
0.05	75.89	66.82	60.40	54.67	49.69
0.10	54.40	38.97	29.33	23.57	17.25
0.15	36.74	20.21	11.07	7.38	4.08
0.20	22.99	8.50	3.47	1.68	0.63
0.25	13.81	3.37	0.96	0.30	0.08
0.30	7.78	1.31	0.32	0.05	0.00
0.35	4.17	0.47	0.05	0.01	0.00
0.40	2.03	0.13	0.01	0.01	0.00
0.45	1.06	0.06	0.00	0.00	0.00
0.50	0.48	0.02	0.00	0.00	0.00

Note: The percentage displayed for each discrepancy refers to the proportion of samples where the discrepancy was *at least* the displayed value.

Table 3

Discrepancy Between Obtained Coefficient and Estimated Population Parameter

Discrepancy	**Percentage of 10,000 Simulations**				
r_{XZ}	***N* = 20**	***N* = 40**	***N* = 60**	***N* = 80**	***N* = 100**
0.05	80.79	73.02	66.26	61.67	57.52
0.10	63.16	48.35	39.25	31.63	25.88
0.15	46.63	28.98	19.95	13.40	8.70
0.20	32.81	15.60	8.14	4.60	2.25
0.25	22.12	7.54	2.78	1.36	0.46
0.30	13.74	3.18	0.87	0.27	0.04
0.35	7.86	1.24	0.26	0.04	0.01
0.40	4.43	0.53	0.05	0.02	0.00
0.45	2.47	0.17	0.02	0.00	0.00
0.50	1.37	0.05	0.00	0.00	0.00
β_X	***N* = 20**	***N* = 40**	***N* = 60**	***N* = 80**	***N* = 100**
0.05	71.66	60.84	51.81	44.93	40.66
0.10	47.44	29.92	19.83	13.81	9.09
0.15	29.06	12.01	5.29	2.62	1.10
0.20	16.02	3.87	1.05	0.30	0.10
0.25	7.79	0.82	0.12	0.02	0.00
0.30	3.49	0.17	0.01	0.00	0.00
0.35	1.54	0.04	0.00	0.00	0.00
0.40	0.48	0.01	0.00	0.00	0.00
0.45	0.12	0.00	0.00	0.00	0.00
0.50	0.05	0.00	0.00	0.00	0.00

Note: The percentage displayed for each discrepancy refers to the proportion of samples where the discrepancy was *at least* the displayed value.

Table 4

Obtained Coefficient Significance Level

Statistic	**Percentage Significant from 10,000 Simulations**				
r_{XZ}	***N* = 20**	***N* = 40**	***N* = 60**	***N* = 80**	***N* = 100**
$p < .05$	39.13	59.22	74.28	84.41	91.47
$p < .01$	24.20	39.62	55.03	67.81	78.71
$p < .001$	13.52	21.29	32.45	44.12	56.15
β_X	***N* = 20**	***N* = 40**	***N* = 60**	***N* = 80**	***N* = 100**
$p < .05$	61.63	89.36	97.51	99.42	99.91
$p < .01$	39.31	73.64	91.37	97.53	99.48
$p < .001$	18.86	48.17	74.65	89.54	96.38
r_{XZ} only	***N* = 20**	***N* = 40**	***N* = 60**	***N* = 80**	***N* = 100**
$p < .05$	5.72	1.86	0.41	0.11	0.02
$p < .01$	6.88	2.99	0.99	0.27	0.09
$p < .001$	6.29	3.55	1.64	0.79	0.24
B_X only	***N* = 20**	***N* = 40**	***N* = 60**	***N* = 80**	***N* = 100**
$p < .05$	27.90	32.00	23.64	15.12	8.46
$p < .01$	21.99	37.01	37.33	29.99	20.86
$p < .001$	11.63	30.43	43.84	46.21	40.47

Note: In this table, "r_{XZ} only" refers to samples where only r_{XZ} was significant.

APPENDIX B

On the population level the effect size of the treatment (ρ_{XZ}) is determined by the value of **a**, the variance of **Y** and the variance of **R**. By choosing the right combination of these three parameters, we will be able to control for the population correlation of treatment and post-test scores. In the current study the variance of Y is arbitrarily set to be 1, ρ_{XZ} is arbitrarily set to be 0.3 and the auto-regression ρ_{yz} is set to be $\sqrt{0.5}$.

$$\rho_{XZ}=\frac{COV(X,Z)}{\sqrt{var(X)var(Z)}}=\frac{COV(X,Y+aX+R)}{\sqrt{var(X)var(Z)}}=\frac{COV(X,Y)+aCOV(X,X)+COV(X,R)}{\sqrt{var(X)var(Z)}}; \quad \text{(equation 4)}$$

suppose on the population level, with infinitely many data, COV(X,Y) and COV(X,R) are zero. Then

$$\rho_{XZ}=\frac{avar(X)}{\sqrt{var(X)var(Z)}}=\frac{a\sqrt{var(X)}}{\sqrt{var(Z)}}=0.3 \quad \text{(equation 5)}$$

Similarly:

$$\rho_{yz}=\frac{COV(Y,Z)}{\sqrt{var(Y)var(Z)}}=\frac{COV(Y,Y+aX+R)}{\sqrt{var(Y)var(Z)}}=\frac{COV(Y,Y)+aCOV(Y,X)+COV(Y,R)}{\sqrt{var(Y)var(Z)}}; \quad \text{(Equation 6)}$$

suppose on the population level, with infinitely many data, COV(X,Y) and COV(Y,R) are zero. Then

$$\rho_{yz}=\frac{COV(Y,Y)+aCOV(Y,X)+COV(Y,R)}{\sqrt{var(Y)var(Z)}}=\frac{var(Y)}{\sqrt{var(Y)var(Z)}}=\frac{\sqrt{var(Y)}}{\sqrt{var(Z)}}=\sqrt{0.5}; \quad \text{(equation 7)}$$

because X has an equal number of 1s and 0s, variance of X is 0.25.

Based on equation 5 and equation 7, substituting var(Y) for 1 and var(X) for 0.25, we get var(Z)=2, and $a=\sqrt{0.72}$.

According to equation 3,

$$var(Z)=var(Y)+a^2var(X)+var(R)+2COV(X,Y)+2COV(X,R)+2COV(Y,R); \quad \text{(equation 8)}$$

suppose on the population level, with infinitely many data, COV(X,Y), COV(X,R) and COV(Y,R) are zero. Then

$$var(Z)= var(Y)+a^2var(X)+var(R)=1+0.72*0.25+var(R)=2$$

so that var(R)=0.82 on the population level.

To sum up, when the population parameters are set as the following values:

$$\begin{cases} var(Y)=1 \\ var(X)=0.25 \\ a=\sqrt{0.72} \\ var(R)=0.82 \end{cases}$$

and when post-test score (Z) is calculated according to equation 1,

the correlations between variables in the model on the population level will be set values:

$$\begin{cases} \rho_{XZ}=0.3 \\ \rho_{YZ}=\sqrt{0.5} \\ \rho_{XY}=\rho_{XR}=\rho_{YR}=0 \end{cases}$$

APPENDIX C

Example STATA Code—Generating Values of X and Y

```
foreach var of newlist Y1-Y`i' {
gen `var' = rnormal(0,1)
 }

foreach var of newlist X1-X`i' {
gen R = runiform()
sort R
gen `var' = 0
capture replace `var' = 1 if _n>_N/2
capture replace R = runiform()
sort R
capture drop R
}
```