Abstract Title Page

Title: On Cornfield's penalties for group randomization: When do degrees of freedom matter?

Author(s): Chris Rhoads, Northwestern University.

Abstract Body

Background/context:
Description of prior research, its intellectual context and its policy context.

Recent years have seen a great increase in interest in social experiments as a means for informing policy decisions about the efficacy of social programs (Greenberg and Shroder, 2008). Since many interventions are naturally delivered to intact social groups, this interest in social experimentation has led to a need to understand the statistical properties of experimental designs where the units randomly assigned to treatments (units of assignment) are composed of several smaller units (units of observation) which each individually provide data to the researcher. Experiments of this sort are commonly referred to as *cluster randomized experiments* (CREs). Cluster randomized experiments are particularly prevalent in education research where the natural hierarchical structure of the educational system (students nested within classrooms which are nested within schools, etc.) lends itself to experiments of this sort.

Cornfield (1978), in a short but widely cited paper, noted that two penalties are paid when a researcher opts to randomize intact groups to treatment rather than individuals. The first penalty is due to an inflated variance of the estimator of the effect of treatment. To illustrate, consider an educational intervention designed to impact student test scores in which $2m$ schools, each consisting of n students, will participate. If individual randomization is used to assign mn students to each of the treatment and control groups, then the sampling variance of the estimator of the mean impact of treatment is $\frac{2\sigma_T^2}{mn}$, where σ_T^2 denotes the variance in student test scores in the population of interest. We can decompose σ_T^2 as $\sigma_T^2 = \sigma_b^2 + \sigma_w^2$, where σ_b^2 represents the within treatment group variance between school mean test scores in the population and σ_w^2 represents the variance of individual test scores around the school mean.

On the other hand, consider the experiment that randomly assigns m intact schools to each of the treatment and control groups, so that, again mn students are in each experimental condition. Under this design the variance of the estimated treatment effect is $\frac{2\sigma_T^2(1+(n-1)\rho)}{mn}$, where $\rho = \frac{\sigma_b^2}{\sigma_T^2}$ denotes the *intra-class correlation coefficient*. The quantity $1 + (n-1)\rho$ is known as the *" design effect"*. The design effect quantifies the increase in variance due to randomization by cluster.

The second of Cornfield's penalties comes about because in the cluster randomized design there are fewer degrees of freedom available to measure the variance of the estimator of the average impact of the treatment. Referring to the example above, in the design that randomly assigns individuals to treatment the usual t test conducted has $2mn - 2$ degrees of freedom, whereas in the design that randomly assigns schools there are $2m - 2$ degrees of freedom.

In situations where cluster randomization cannot be avoided it is impossible to avoid a treatment effect estimator with increased variance. However, Blair and Higgins (1986) pointed out that if the ICC is known prior to conducting the experiment, Cornfield's second penalty can be avoided and a test of the treatment effect can be conducted with $2mn - 2$ rather than $2m - 2$ degrees of freedom. Konstantopoulos (forthcoming) has made the same point with regard to experimental designs with three levels of nesting. While the methods

proposed by Blair and Higgins (1986) and Konstantopoulos (forthcoming) are intriguing ways to boost power, they can only be used if the exact value of the ICC (or ICCs in the case of a three level design) is known prior to the experiment. In situations where schools form clusters and the schools participating in the experiment are a clearly defined sample from a population of schools, and when extensive administrative data is available for the population of schools, it may be reasonable to assume that the ICC is known with enough precision to use the methods mentioned above.

In many situations some estimator of the ICC will be available from data external to the main experiment. For example, a pilot study may provide information about the ICC. However, the pilot study may not be very large and so the available estimate of the ICC will be subject to considerable imprecision due to sampling variance. In a situation like this it is not reasonable to assume that the ICC is known exactly, however, it would still be useful to be able to use the prior information about the ICC in order to improve the power of the experiment.

Utilizing an estimate of the ICC that is subject to sampling error to improve power is the subject of two papers by Blitstein and colleagues: Blitstein, Hannan, Murray and Shadish (2005), and Blitstein, Murray, Hannan and Shadish (2005). These papers describe a method for utilizing external information about ρ when the external estimator of ρ is subject to sampling variance (so that ρ is not known exactly). They propose a method that multiplies the usual test statistics that are computed when observations are independent by a correction term that depends on ρ. Information about ρ is obtained from data existing prior to the experiment in question, and, where possible, this information is combined with information about ρ obtained from the experiment itself. The test statistic is then computed and compared to the critical value of a t distribution with df^* degrees of freedom, where df^* is a synthetic degrees of freedom estimate computed by a meta-analytic procedure given in Blitstein, Hannan, Murray and Shadish (2005).

However, the df^* method of Blitstein, et al. has major problems. The arguments presented in favor of the method are ad hoc in nature, and so there is no guarantee that the test procedures advocated will control type I error rates at the nominal level. Additionally, the properties of the procedure are such that the computed value of df^* may be greater than $2mn - 2$, which is the maximum possible degrees of freedom available when ρ is known exactly. Also, Blitstein, Hannan, Murray and Shadish (2005) propose three possible uses of the external estimate of ρ when utilized in conjunction with appropriate correction factors:

(1) To correct a test statistic that incorrectly ignored clustering.
(2) To correct a test statistic that incorrectly treated clusters as fixed effects.
(3) To improve the power of a test statistic that correctly accounted for clustering but that had low power.

As noted in Rhoads (2008), the maximum number of degrees of freedom available in these three cases is different, yet the method of Blitstein, Hannan, Murray and Shadish computes df^* in the same fashion in each case. Finally, a mistake in the Blitstein, Murray, Hannan and Shadish (2005) paper leads the authors to overstate the amount of power improvement that is possible due to prior knowledge of ρ. The authors consider a design scenario where, when using traditional methods, 17 clusters per treatment group are necessary to obtained the desired statistical power of 0.80. They argue that using the df^* method, it is possible to obtain the same power with only 6 clusters. In fact, Rhoads (2008) shows that, under the design situation considered, even if ρ is known exactly, at least 16 clusters per treatment

group are necessary to obtain the desired power.

Perrett (2006) considered the use of prior information about the ICC in the case of experiments with only one cluster per treatment group, corresponding to the case of $m = 1$ in the notation used above. In this case, it is impossible to construct a test that correctly accounts for clustering using only the data from the experiment so it is necessary to use external information about the ICC. Perrett advocates using prior information about ρ to place an upper bound on ρ. Since the test statistic used is a decreasing function of the ρ that is used, this approach ensures that the type I error rate is controlled at the nominal level, provided the upper bound really is larger than the true value of ρ.

Donner and Klar (2000) warned against the use of external estimates of ρ to attempt to improve power in cluster randomized trials. They state, "The potentially increased precision obtained from using an external estimate of ρ must be weighed against the accompanying risk of bias resulting from a wrongly assessed value of the parameter. Given the sensitivity of statistical inferences to the assigned value of ρ, we would discourage this practice unless very reliable and representative external data are available (p.116)."

Purpose / objective / research question / focus of study:
Description of what the research focused on and why

The current research builds on existing research in the following ways:

(1) Using prior information about the ICC is a tempting way to improve power. However, as evidenced by the disapproval that Donner and Klar (2000) show, the approach is controversial. Before considering the method, a researcher might wish to know just how much improvement in power is possible. If power improvement will be substantial, perhaps it is worth it to make the additional assumptions needed to justify the method. On the other hand, if the possible power improvement is slight, most researchers will conclude that it is preferable to utilize the standard test statistic. Previous work has made it difficult to determine exactly when substantial gains can be expected. The current work provides a bound on the possible power improvement due to the use of external information about the ICC.
(2) It is shown that the bounding approach recommended by Perrett (2006), while a viable approach in unreplicated experiments, will rarely improve the power of experiments where more than one cluster is randomized to treatment and control conditions.

Findings / Results:
Description of main findings with specific details.

Let y_{ijk} denote the k^{th} observation from the j^{th} cluster which has been assigned to the i^{th} treatment ($i = 1, \ldots, 2; j = 1, \ldots, m; k = 1, \ldots, n$). Let the "$\cdot$" notation stand for averaging over the given subscript. When there is no prior information about ρ, the uniformly most powerful unbiased (UMPU) test of the null hypothesis of no treatment effect is given by:

$$t_C = \frac{\sqrt{\frac{mn}{2}}(y_{1\cdot\cdot} - y_{2\cdot\cdot})}{\sqrt{\frac{SSB}{2(m-1)}}}. \tag{1}$$

In the above, SSB denotes the sum of squares between clusters, given by

$$SSB = n\sum_{i=1}^{2}\sum_{j=1}^{m}(y_{ij.} - y_{i..})^2.$$

When ρ is exactly known prior to the experiment, then the UMPU test of the null hypothesis of no treatment effect is given by:

$$t_{GLS}(\rho) = \frac{\sqrt{\frac{mn}{2}}(y_{1..} - y_{2..})}{\sqrt{\left[SSB + SSW\frac{(1+(n-1)\rho)}{1-\rho}\right]/(2(mn-1)}}. \tag{2}$$

Here SSW is defined as :

$$SSW = \sum_{i=1}^{2}\sum_{j=1}^{m}\sum_{k=1}^{n}(y_{ijk} - y_{ij.})^2. \tag{3}$$

Under a given alternative hypothesis, $H_0 : \mu_1 - \mu_2 \neq 0$, both t_{GLS} and t_C are distributed as a non-central t random variable with the same non-centrality parameter

$$\lambda = \frac{(\mu_1 - \mu_2)(\sqrt{mn/2})}{\sigma_T}\sqrt{\frac{1}{1+(n-1)\rho}}, \tag{4}$$

but different degrees of freedom, $2mn-2$ and $2m-2$ respectively. Thus, the power improvement that can result from the use of prior information about the ICC is logically bounded by the difference between the power of the test based on t_C and the power of the test based on t_{GLS}.

Previous treatments have tended to present the power differences between the two tests as a function of m, n, ρ and $\delta = \frac{\mu_1-\mu_2}{\sigma_T}$ for a fixed type I error rate, α. Given the myriad possible combinations of these parameters, the prior treatments made it difficult to determine a simple rule indicating when using prior information about the ICC might result in substantial power improvements. The current paper instead notes that ultimately power differences depend on only two parameters, the degrees of freedom and the non-centrality parameter, λ. Furthermore, while the non-centrality parameter depends on m, n, ρ and δ, the degrees of freedom depend only on the cluster and individual level sample sizes, m and n. Hence, given values for m and n, the power improvement due to using the test based on t_{GLS} is a function of only one parameter, λ. The maximum of this function over the parameter space of λ was computed for values of m ranging from $2, \ldots, 10$, for $n = \infty$, and for $n = 10$. Various one-sided α levels are considered. One-sided tests are used to prevent results being influenced by probabilities of rejection in the lower tail of the non-central t distribution. Since only realtively large power values are of practical interest (say power above 0.50), for all practical purposes, results for one-sided level α tests may be taken to be equivalent to results for two-sided level 2α tests. The results are presented in tables 1 and 2 in the appendix.

Examining table 1, and focusing attention on the $\alpha = 0.025$ column (corresponding to the usual two-sided $\alpha = 0.05$ level test), we see that, even when the cluster size is very large, once there are at least 20 clusters in the experiment it is impossible to obtain power improvements greater than 0.05 by using external information about ρ. Unless there are less than 10 clusters in the experiment, the maximmum possible power improvement will be 0.10. Turning attention to table 2, we note that, even with a relatively modest cluster sample size of $n = 10$, there is very little difference between table 1 and table 2. Hence, for most cluster randomized experiments, the values given in table 1 are reasonable approximations to the

amount of possible power improvement.

When the value of ρ is not known perfectly in advance of the experiment, placing an upper bound on the true value of ρ, say ρ_b, and using the test statistic $t_{GLS}(\rho_b)$ seems like a reasonable way to proceed. Perrett (2006) demonstrated how this approach can be used in experiments with only one cluster per treatment arm. Unfortunately, when there is more than one cluster per treatment arm, the situations where this approach will give a more powerful test than the test based on t_C are quite rare. When a value for ρ_b that is other than the true value of ρ is plugged into the statistic T_{GLS}, the test statistic no longer has a t distribution, but its distribution is well approximated by a constant k times a t distribution with h degrees of freedom. The values of k and h are given by

$$k = \frac{\sqrt{2(mn-1)(1+(n-1)\rho)(1-\rho_b)}}{\sqrt{(1+\rho_b(n-1))(1-\rho)(2mn-2m)+(2m-2)(1+\rho(n-1)(1-\rho_b)}} \tag{5}$$

and

$$h = \frac{[(1+\rho_b(n-1))(1-\rho)(2mn-2m)+(2m-2)(1+\rho(n-1))(1-\rho_b)]^2}{(1+\rho_b(n-1))^2(1-\rho)^2(2mn-2m)+(2m-2)(1+\rho(n-1))^2(1-\rho_b)^2}. \tag{6}$$

The above result may be used to compute the power improvement of the test based on $T_{GLS}(\rho_b)$ relative to the usual test. Figure 1 considers the case of $m = 4$, $n = 25$ and $\delta = 0.5$. Power comparisons are presented in terms of the ratio of the true value of ρ to the bound. In the situation considered, if ρ is 0.05, the bound used can be no larger than about $1.7(0.05) = 0.085$ in order for the bounded test to improve power. If ρ is 0.25, the bound used can be no larger than about $1.2(0.25) = 0.30$ in order for the bounded test to improve power. The example considered involves an relatively small experiment with $2m = 8$ clusters randomized. As the value of m increases, it becomes even harder for the bounded test to improve power (in the sense that the bound must be even closer to the true value of ρ).

Conclusions:

Description of conclusions and recommendations based on findings and overall study.

Cornfield (1978) noted that two penalties are paid when clusters of individuals, rather than individuals themselves, are randomly assigned to treatments. Nothing can be done about the first penalty, the inflation of the variance of the estimated effect of treatment. However, the second penalty, a decrease in the degrees of freedom available to estimate the variance of the estimated treatment effect can be avoided if the intraclass correlation coefficient is known prior to the study. Since the ICC is almost never known with complete precision, improving the power of a CRE by treating the ICC as known prior to the study is somewhat controversial.

Given this controversy, it is useful to know in what situations substantial power improvements can be expected by the use of this method. The current study shows that, if usual type I error rates are used, the power improvement will never be greater than 0.05 when 20 or more clusters participate in the experiment. The power improvement will never be greater than 0.10 when 10 or more clusters particiapte in the experiment.

In situations where substantial power improvement is possible, but prior data on the ICC is subject to considerable sampling error, there is still no consensus on how this prior information might be used to improve power. One idea is to utilize an upper bound on the value of ρ. The present study shows that, unless the bound on ρ is quite close to the true value,

this strategy will usually fail to improve power relative to the usual test.

While not considered in this paper, a different method for utilizing prior information about ρ is presented in Rhoads (2008). This method is guaranteed to maintain type I error rates at the nominal level.

Appendices

Not included in page count.

Appendix A. References

References

[1] Blair, R.C. and Higgins, J.J. (1986). Comment on "Statistical power with group mean as the unit of analysis". *Journal of Educational Statistics*, **11**, 161-9.

[2] Blitstein, J.L., Hannan, P.J., Murray, D.M. and Shadish, W.R. (2005a). Increasing the degrees of freedom in existing group randomized trials through the use of external estimates of the intraclass correlation: The df* approach. *Evaluation Review*, **29**, 241-67.

[3] Blitstein, J.L., Murray, D.M.,Hannan, P.J. and Shadish, W.R. (2005b). Increasing the degrees of freedom in future group randomized trials through the use of external estimates of the intraclass correlation: The df* approach. *Evaluation Review*, **29**, 268-86.

[4] Cornfield, J. (1978). Randomization by Group: a formal analysis. *American Journal of Epidemiology*, **108**, 100-2.

[5] Donner, A. and Klar, N. (2000). *Design and Analysis of Cluster Randomization Trials in Health Research*. New York: Oxford University Press.

[6] Greenberg, D. and Shroder, M. (2008). *The Digest of Social Experiments, Third Edition*. Washington, D.C.: Urban Institute Press.

[7] Konstantopoulos, S (forthcoming). Constructing a More Powerful Test in Three-Level Cluster Randomized Designs. *Journal of Research on Educational Effectiveness*.

[8] Perrett, J. (2002). A Method for Analyzing Unreplicated Experiments Using Information on the Intraclass Correlation Coefficient. *Journal of Modern applied Statistical Methods*, **5(2)**, 432-442.

[9] Rhoads, C.H. (2008). *Utilizing External Information about the Covariance Structure in Experiments with Clustering*. Unpublished Doctoral Dissertation. Northwestern University, Evanston, IL.

m=	$\alpha =$	**0.005**	**0.01**	**0.025**	**0.05**	**0.075**	**0.1**
2	*max improve*	0.783	0.667	0.469	0.305	0.214	0.156
	GLS at max	0.968	0.944	0.897	0.852	0.826	0.810
3	*max improve*	0.486	0.376	0.236	0.145	0.100	0.072
	GLS at max	0.872	0.838	0.795	0.770	0.761	0.758
4	*max improve*	0.330	0.249	0.153	0.093	0.064	0.047
	GLS at max	0.805	0.779	0.752	0.740	0.738	0.740
5	*max improve*	0.246	0.184	0.113	0.068	0.047	0.034
	GLS at max	0.766	0.747	0.729	0.725	0.728	0.732
6	*max improve*	0.195	0.146	0.089	0.054	0.037	0.027
	GLS at max	0.740	0.727	0.716	0.716	0.721	0.727
7	*max improve*	0.161	0.120	0.073	0.045	0.031	0.022
	GLS at max	0.723	0.713	0.707	0.710	0.716	0.724
8	*max improve*	0.137	0.102	0.062	0.038	0.026	0.019
	GLS at max	0.710	0.703	0.701	0.706	0.713	0.721
9	*max improve*	0.120	0.089	0.054	0.033	0.023	0.017
	GLS at max	0.701	0.696	0.696	0.703	0.711	0.720
10	*max improve*	0.106	0.079	0.048	0.029	0.020	0.015
	GLS at max	0.693	0.690	0.692	0.701	0.710	0.718

TABLE 1. Max possible power improvement and GLS power at max when $n = \infty$, various α

m=	$\alpha =$	**0.005**	**0.01**	**0.025**	**0.05**	**0.075**	**0.1**
2	*max improve*	0.768	0.651	0.454	0.293	0.206	0.150
	GLS at max	0.965	0.941	0.893	0.849	0.824	0.809
3	*max improve*	0.463	0.356	0.223	0.136	0.094	0.068
	GLS at max	0.865	0.831	0.790	0.767	0.759	0.757
4	*max improve*	0.310	0.233	0.143	0.087	0.060	0.043
	GLS at max	0.798	0.773	0.748	0.738	0.737	0.740
5	*max improve*	0.229	0.171	0.104	0.063	0.044	0.032
	GLS at max	0.759	0.742	0.727	0.723	0.726	0.731
6	*max improve*	0.180	0.134	0.082	0.050	0.034	0.025
	GLS at max	0.734	0.722	0.713	0.715	0.720	0.726
7	*max improve*	0.149	0.110	0.067	0.041	0.028	0.021
	GLS at max	0.718	0.709	0.705	0.709	0.716	0.723
8	*max improve*	0.126	0.094	0.057	0.035	0.024	0.017
	GLS at max	0.706	0.700	0.699	0.705	0.713	0.721
9	*max improve*	0.110	0.081	0.050	0.030	0.021	0.015
	GLS at max	0.697	0.693	0.694	0.702	0.711	0.719
10	*max improve*	0.097	0.072	0.044	0.027	0.018	0.013
	GLS at max	0.690	0.687	0.690	0.700	0.709	0.718

TABLE 2. Max possible power improvement and GLS power at max when $n = 10$, various α

Appendix B. Tables and Figures

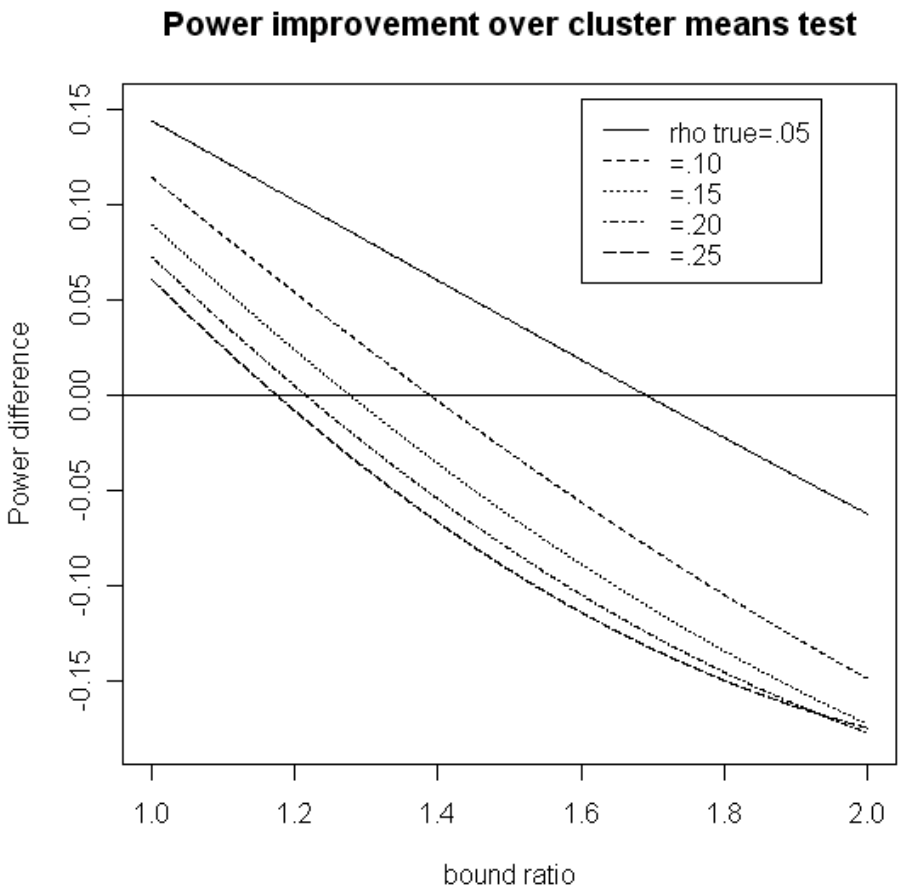

FIGURE 1. Power improvement of test using upper bound on ρ as function of ratio of bound to true ρ, $\delta = 0.5$, $m = 4$, $n = 25$